Autenticação Comportamental de Motoristas em Redes Veiculares

Paulo H. L. Rettore[1], André B. Campolina[1], Artur Souza[1], Guilherme Maia[1], Leandro A. Villas[2], Antonio A. F. Loureiro[1]

[1]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte – MG – Brasil

[2]Instituto de Computação, Universidade de Campinas (UNICAMP)
Campinas, SP – Brasil

{rettore,andre.campolina,arturluis,jgmm,loureiro}@dcc.ufmg.br,

leandro@ic.unicamp.br

Abstract. *The community has discussed the potential of processing and wireless communication of vehicles in a transportation system. In this sense, VANETs aim to exploit the communication and sensing capabilities of vehicles to feed data into applications and services. VANETs also contribute to the emergence of ADAS and ITS, which seek to provide services to users as safety and less tiring trips. Many of these systems need to authenticate their users, but they do so in a way that an attacking driver can use them. This work explores the driver identification as an extra authentication factor to local services and vehicular networks. Then, a virtual sensor was developed to determine the driver identity, with precision above 98%, using embedded sensor data. This sensor was also used to identify driver suspects. Besides, based on the suspect's identification, we discuss the impacts of these drivers in the data dissemination in a vehicular network.*

Resumo. *Muito se discute sobre o potencial de processamento e comunicação sem fio de veículos em um sistema de transporte. Nesse sentido, as redes veiculares (VANETs) têm como objetivo explorar a capacidade de comunicação e sensoriamento de veículos para alimentar com dados aplicações e serviços. VANETs também contribuem para o surgimento de Sistemas de Assistência ao Motorista (ADAS) e Sistemas de Transporte Inteligente (ITS), que buscam fornecer serviços aos usuários como viagens mais seguras e menos cansativas. Muitos desses sistemas necessitam autenticar seus usuários, porém o fazem de maneira que um motorista invasor possa utilizá-los. Este trabalho explora a identificação de motoristas como fator extra de autenticação em serviços locais e de uma rede veicular. Para isso, foi desenvolvido um sensor virtual para determinar a identidade de motoristas, com precisão acima de 98%, usando dados de sensores embarcados. Esse sensor também é utilizado na identificação de motoristas suspeitos. Além disso, diante da identificação de suspeitos, são discutidos os impactos desses motoristas na disseminação de dados em uma rede veicular.*

1. Introdução

Os sistemas de controle de veículos modernos dependem dos dados de sensores para que a estabilidade e experiência de condução seja mais segura e confortável ao motorista. Esses dados estão disponíveis através da interface universal de Diagnóstico On-Board – On-Board Diagnostic (OBD) –, introduzida para fins de regulamentação e manutenção, mas que permite também obter valores de diferentes tipos de variáveis do veículo.

Os dados dos sensores veiculares por si só não apresentam informações valiosas para os motoristas, uma vez que a maioria desses dados é usada pela Unidade de Controle do Motor – Engine Control Unit (ECU) – para melhor ajustá-lo. Por exemplo, oxigênio e sensores de pressão de combustível, dentre outros, não representam significado claro para um motorista inexperiente. Além disso, os sensores que indicam informações úteis para o condutor apresentam seus valores no próprio painel do veículo (por exemplo, rotações por minuto do motor e velocidade atual). Um dos desafios que surge com o acesso a esses dados está em apresentar informações úteis, bem como fornecer serviços aos condutores e a uma rede veicular, com base nas leituras dos sensores de seus veículos.

Nesse sentido, as redes veiculares – Vehicular *Ad-hoc* Networks (VANETs) – exploram a capacidade de comunicação e sensoriamento dos veículos com o objetivo de fornecer dados às aplicações e serviços que executarão nesse ambiente. Isso contribui para o surgimento de Sistemas de Assistência ao Motorista – Advanced Driver Assistant Systems (ADAS) – e Sistemas de Transporte Inteligente – Intelligent Transportation System (ITS) –, os quais fornecem vários serviços, dentre eles mecanismos de segurança às pessoas no transito, e de conforto aos motoristas e passageiros como acesso às redes sociais, *streams* de vídeo e rotas. Muitos desses sistemas necessitam autenticar seus usuários, para um conteúdo direcionado, porém o fazem de maneira que um motorista invasor possa utilizá-los.

Assim, este trabalho apresenta um sensor virtual para autenticar motoristas de um veículo baseado nos seus comportamentos de condução. Esse sensor, por sua vez, é utilizado na diferenciação entre motoristas legítimos e suspeitos. Essa identificação é tratada como um fator extra na autenticação do motorista e tem os objetivos de (i) permitir uso serviços locais (*intra-vehicle*) e (ii) serviços de rede (*extra-vehicle*) de forma segura e autenticada. O sensor virtual usa dados de sensores embarcados para identificar a pessoa que está dirigindo o veículo, dado um conjunto de dados rotulados previamente. Baseado nessa identificação, o sensor virtual possibilita que serviços locais e de rede sejam habilitados para o motorista legítimo ou desabilitados, caso contrário.

O restante deste trabalho está organizado da seguinte forma. A Seção 2 apresenta os trabalhos relacionados. A Seção 3 discute o segundo fator de autenticação de motoristas e as preocupações com a privacidade e segurança dos dados. A Seção 4 descreve o processo de coleta e as características dos dados adquiridos nos veículos de teste. A Seção 5 descreve o estágio de preparação desses dados e redução de variáveis. A Seção 6 apresenta o sensor de identificação dos motoristas legítimo e suspeito, bem como os resultados das avaliações. A Seção 7 apresenta os resultados da simulação da disseminação de motoristas suspeitos na rede veicular e uma discussão do impacto de veículos suspeitos nessa rede. Finalmente, a Seção 8 apresenta as conclusões e trabalhos futuros.

2. Trabalhos Relacionados

Na literatura, existem estudos dedicados a reconhecer o comportamento e identidade do motorista. Analisar dados de condução é tema de interesse devido ao aumento das questões relacionadas à segurança de veículos em um sistema de transporte. Nesse sentido, vários trabalhos focam no reconhecimento de estilo de direção [Bergasa et al. 2014, Hallac et al. 2016, Johnson and Trivedi 2011, Martínez et al. 2016]. Alguns desses trabalhos identificam quem é o motorista e outros caracterizam seu comportamento, como agressivo ou normal, por exemplo. [Zhang et al. 2016] desenvolveram um modelo de identificação de motorista, usando sensores disponíveis no *smartphone* e no veículo, usando a interface OBD. Eles avaliaram três veículos em dois ambientes diferentes, controlado e normal. Considerando apenas os sensores veiculares, o modelo de classificação obteve uma precisão de 30,36% no ambiente controlado com 14 condutores e 85,83% no ambiente normal com dois condutores por veículo.

[Carmona et al. 2015] propuseram uma nova ferramenta para analisar o comportamento do motorista, detectando o comportamento agressivo em tempo real. [Aoude et al. 2011] desenvolveram algoritmos para estimar o comportamento do motorista nas interseções rodoviárias. Eles introduziram duas classes de algoritmos que podem classificar os condutores como compatíveis ou violadores.

Outros estudos buscam fortalecer a autenticação de motoristas em seus veículos. Nessa linha, destacam-se trabalhos que propõem mecanismos para autenticar os condutores utilizando características biométricas. [Yuan and Tang 2011], por exemplo, propuseram um mecanismo de autenticação baseado em características da palma da mão e distribuição de veias do motorista. Similarmente, [Silva et al. 2012] propuseram um mecanismo de autenticação baseado em leituras do eletrocardiograma do motorista. As leituras das mãos do motorista são feitas por sensores colocados no volante do veículo.

[Burton et al. 2016] utilizaram um simulador para obter padrões de direção de motoristas como a pressão aplicada aos pedais, distância média viajada e o desvio médio do volante. De posse desses dados, os autores utilizaram *Support Vector Machines* (SVM) para tentar identificar e autenticar os motoristas. [Salemi 2015] propôs autenticar motoristas considerando dados obtidos por meio da interface OBD. Foram extraídas sete características de direção, como frenagem e variações na velocidade, e aplicada SVM para identificar e autenticar os motoristas. O resultado mostrou uma precisão de até 94% na identificação de motoristas.

Nosso trabalho difere das propostas anteriores de identificação e autenticação por considerar os seguintes aspectos: apenas dados extraídos do próprio veículo são usados (diferente de [Burton et al. 2016]) e por considerar o comportamento do motorista em vez de dados biométricos estáticos ([Yuan and Tang 2011, Silva et al. 2012]). Além disso, o nosso trabalho difere do trabalho de Salemi [Salemi 2015] na metodologia adotada para identificar os motoristas e, consequentemente, obteve precisão maior (acima de 98%). Além disso, combinamos a autenticação de motoristas para fornecer serviços personalizados de assistência aos condutores legítimos, e serviços em uma rede veicular.

3. Fator Extra de Autenticação do Motorista

Nesta seção iremos discutir e propor uma abordagem para autenticar motoristas, baseado em seus hábitos de condução do veículo, com o objetivo de identificá-los. Para fazer isso,

é necessário conhecer o motorista a partir de um conjunto de dados de condutores conhecidos de um determinado veículo compartilhado, ou seja, um veículo conduzido por diferentes motoristas. Uma vez que os conjuntos de dados de motoristas individuais são rotulados, a identificação do motorista pode ser traduzida em um problema de classificação. O desenvolvimento de uma metodologia que possibilite um fator extra de autenticação de motoristas, permite fornecer serviços intra-veículo e inter-veículo. No serviço intra-veículo é possível a personalização de Sistemas de Assistência ao Motorista (*Advanced Driver Assistance Systemse – ADAS*), como entretenimento, ergonomia, serviços de rotas e serviços que auxiliam na eficiência de combustível. Já no serviço inter-veículo, uma rede veicular pode permitir ou não a troca de mensagens, entretenimento e sugestões personalizadas de rotas, caso o veículo seja considerado suspeito/ilegítimo. Entendemos que o sistema atual de autenticação é frágil o suficiente para permitir o acesso, de qualquer indivíduo, às informações pessoais dos motoristas principais ao romper a única etapa de autenticação (chave). Estas informações estão vinculadas aos sistema embarcados que armazenam as preferencias do usuário, tais como rotas e pontos de interesses, lista de contatos e acesso à residencia e locais de trabalho.

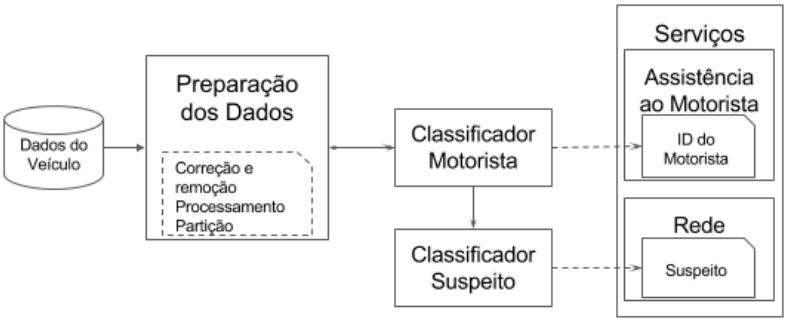


Figura 1. Fluxo de Identificação de Motoristas e Suspeitos

O procedimento de identificação do condutor é dividido em seis etapas. A partir dos dados coletados, a primeira etapa consiste em preparar os dados corrigindo e eliminando variáveis que contenham valores faltantes ou que não são influenciados pelo comportamento do motorista. Em seguida (segunda etapa), usamos a Análise de Componentes Principais – *Principal Component Analysis (PCA)* – para reduzir o espaço de análise, mantendo os dados com maior variabilidade. Na terceira etapa, particionamos os dados em uma base de treinamento e uma base de teste, considerando um particionamento aleatório e um particionamento que considere características de inicio e fim de cada viagem. A quarta etapa consiste em classificar os motoristas usando o algoritmo *Extremely Randomized Tree (Extra-Trees)*. No fim desta etapa é possível identificar o condutor e fornecer dados para a próxima etapa que visa verificar se o motorista é suspeito ou não.

A quinta etapa desconsidera a identidade real do motorista e se dedica a verificar se o motorista é autêntico ou suspeito. Finalmente, no sexto passo, realizamos uma análise exploratória para determinar quais tratamentos podem ser feitos nos dados para melhorar a precisão do classificador, tais como, sem tratamento, dados normalizados e janelas de observação entre 30 e 180 segundos com a média móvel dos dados. Além disso, é verificada a importância das variáveis, usando os recursos do algoritmo de florestas aleatórias [Pedregosa et al. 2011], mantendo as variáveis que mais contribuem para a precisão da predição. A Figura 1 apresenta o fluxo de identificação de motoristas e suspeitos proposto.

Vale ressaltar que esses passos descrevem a metodologia que sustenta a proposta desse trabalho. Em outras palavras, a principal contribuição está no uso de sensores do próprio veículo para determinar e diferenciar o comportamento dos motoristas, possibilitando habilitar/desabilitar serviços em um contexto local e de rede, com a autenticação e a identificação de suspeitos, diferentemente do trabalho de Salemi, por exemplo. Entendemos que os serviços locais de ADAS carregam informações pessoais dos motoristas principais, não sendo desejável o acesso à indivíduos suspeitos. Em serviços de redes, não é conveniente que as mensagens sejam entregues para motoristas suspeitos, por exemplo.

3.1. Privacidade e Segurança dos Dados Veiculares

Atualmente, o principal mecanismo de autenticação do motorista com o veículo é a chave de ignição. Nesse mecanismo, a chave atua como um *token* de autenticação e qualquer usuário com posse desse *token* é considerado autêntico. Esse mecanismo é ineficiente para autenticar o motorista, uma vez que toda segurança é baseada no *token* que pode ser roubado juntamente com o veículo. Ou seja, um adversário/suspeito ao roubar o veículo pode utilizar a chave de ignição para se autenticar no veículo normalmente e acessar todas as aplicações e serviços existentes. Por exemplo, o suspeito torna-se capaz de acessar dados sigilosos do motorista autenticado, como suas preferências de rotas e mensagens trocadas ou até mesmo usar o próprio veículo para lançar ataques à rede, debilitando aplicações como sistemas de roteamento (disseminando mensagens falsas) ou sistemas de proteção de outros motoristas (omitindo mensagens de segurança).

Um dos objetivos deste trabalho é aumentar a segurança do sistema de autenticação, utilizando o comportamento do motorista como um segundo fator de autenticação. Esta solução autenticação depende de algo inerente à pessoa, e que o adversário não consegue se apropriar. Porém, por depender do comportamento do motorista, a solução se torna reativa, identificando um adversário apenas após este contornar o sistema primário de autenticação. Nesse ponto, não é mais viável simplesmente bloquear todo o acesso do motorista atual ao veículo, devido ao risco de causar acidentes ou prejudicar todo um sistema de transporte e sua eficiência. Por outro lado, a partir da identificação de um motorista suspeito, feita pela proposta apresentada neste trabalho, pode-se tomar uma série de medidas tanto intra-veículo (por exemplo, definir uma velocidade máxima de condução) como inter-veículo (por exemplo, usar uma interface de comunicação sem fio específica instalada no veículo para notificar uma empresa de segurança, autoridade policial ou mesmo uma pessoa do ocorrido e da localização corrente).

De qualquer forma, para contornar esse problema, a alternativa mais viável é permitir ao veículo bloquear parcialmente o acesso do intruso ao ADAS. Nessa linha, o veículo bloqueia todas as aplicações que não são fundamentais para o funcionamento correto do veículo e da rede. Isto é, todas as aplicações de entretenimento e conforto, assim como aquelas que possuem conteúdo sensível, são bloqueadas. Novamente, aplicações de proteção aos motoristas e mensagens de posicionamento não podem ser bloqueadas devido ao risco de colocar outros participantes da rede em perigo.

Para complementar essa alternativa, propomos também que o veículo periodicamente alerte os outros veículos que o condutor atual é um suspeito. Ao receberem esse alerta, os veículos na mesma região do suspeito propagam o alerta aos seus vizinhos mais próximos, até um certo limite de distância ou até que essa mensagem chegue a um veículo

oficial que pode tomar providências específicas. Além disso, os veículos (indiretamente os motoristas) podem tomar ações adicionais quando recebem o alerta. Um motorista cauteloso pode, por exemplo, optar por evitar rotas que contenham veículos suspeitos, mudar de rota se estiverem próximos ao suspeito ou até mesmo contribuir para sua captura ao aumentar o fluxo de transito nas regiões onde encontram-se suspeitos.

Por fim, argumentamos que a solução proposta incorre em um impacto mínimo na privacidade dos motoristas. Isso porque em momento algum o veículo divulga, na rede veicular, a identidade do motorista que está ao volante. A divulgação dessa identidade é limitada apenas à informação de motorista suspeito ou não. Não é factível a um agente externo descobrir a identidade do condutor baseando-se apenas nessa informação. Se os motoristas legítimos de veículos seguirem rotinas bem definidas, um agente externo pode dizer se esse veículo segue um comportamento previsível. Porém, até mesmo essa informação é de pouca utilidade para um adversário.

4. Coleta de Dados

Hoje em dia, os veículos modernos têm sistemas embarcados sofisticados que objetivam melhorar a segurança da condução, o desempenho, o conforto e o consumo de combustível. Para alcançar esses objetivos, os fabricantes têm investido tanto na quantidade quanto na qualidade dos sensores [Fleming 2001]. Atualmente, um veículo coleta informações de centenas de sensores que estão conectados à Unidade de Controle da Máquina – *ECU* – através de uma rede interna de sensores com fio [Qu et al. 2010] e os dados de saída são acessíveis por meio de uma interface *OBD*.

A interface OBD-II foi introduzida para padronizar o conector físico, os protocolos e o formato das mensagens com as quais eles lidam. O sistema é geralmente empregado para monitorar e regular as emissões de gás e está presente em todos os carros produzidos na Europa e nos Estados Unidos desde 1996 e, no Brasil, desde 2010. A interface OBD também auxilia os serviços de manutenção, ao rastrear a origem de problemas mecânicos [Lin et al. 2009]. Ao possibilitar o armazenamento dos códigos de falha do motor, essas informações fornecem aos mecânicos um histórico de problemas do veículo e possíveis fontes associadas. O processo de coleta utiliza a interface OBD-II como meio de acesso aos dados do veículo, transferindo-os via conexão *Bluetooth* para um *smartphone* com o sistema operacional Android, onde são processados e registrados através de um aplicativo.

Os dados coletados dos sensores veiculares estão disponíveis através dos PIDs da interface OBD. A Tabela 1 apresenta uma amostra dos dados coletados do veículo, *smartphone* e sensores virtuais. Existem também outras centenas de sensores que podem ser acessados através dos PIDs, alguns dos quais são definidos pelo padrão OBD e outros pelos fabricantes dos veículos. Contudo, este trabalho tem por objetivo responder a seguinte questão: *Os dados de sensores de um veículo são capazes de identificar o motorista, baseando-se em seu comportamento?* Desse modo, nos concentraremos nos dados coletados do veículo e também nos dados dos sensores virtuais projetados a partir dos sensores físicos existentes. Um sensor virtual é basicamente um sensor que gera dados mais sofisticados a partir de um algoritmo que recebe dados de sensores físicos. Por exemplo, a interface OBD pode não oferecer a um usuário comum a marcha corrente do veículo e, assim, é necessário executar um algoritmo que recebe como entrada dados de sensores

físicos (velocidade e rotação do motor) para inferir a marcha em determinado instante.

Para responder à questão tratada neste trabalho, foi feito um estudo de caso baseado em dados de sensores de dois veículos compartilhados entre quatorze motoristas. A Tabela 2 apresenta a configuração do processo de coleta de dados. Um dos aspectos importantes desse processo diz respeito aos tipos de viagens registradas por ambos os veículos: todos os quatro motoristas que compartilharam o Veículo #2 foram convidados a percorrer duas rotas diferentes, enquanto os dez motoristas do Veículo #1 o utilizaram para vários fins em suas rotinas diárias.

Tabela 1. Dados da ECU, *smartphone* e sensores virtuais

Dados Coletados				
Smartphone		*Veículo*		*Sensor Virtual*
Data/Hora	Distância da Viagem	Torque*	Rotação Por Minuto*	Aceleração*
GPS	Nível de Combustível Restate	Fluxo de Combustível*	Velocidade*	Tempo de Reação
Velocidade (GPS)	Temperatura do Ambiente	Temperatura do Motor*	Média de CO_2*	Força de Atrito do Ar
GPS HDOP	Custo do km Inst (R$)	Voltagemda Bateria*	CO_2 Instantâneo*	Velocidade/RPM*
Bússola	Custo da Viagem (R$)	KPL Instantâneo*	Posição do Pedal*	Marcha*
Giroscópio	Barômetro	Temperatura de Entrada do Ar*	Média de KPL	
Altitude		Média de KPL da Viagem	Nível de Combustível	

Tabela 2. Configuração da coleta

	Veículo 1	Veículo 2
Motor	1.0 16v	1.6 16v
RPM Max	7000	7000
Transmissão	5	5
Potência	76	122
Peso	1025 kg	1000 kg
Fabricante	Renault	Hyundai
Modelo	Sandero	HB20
Viagens	36	8
Tempo de Viagem	28 horas	3 horas
Tipo de Viagem	Natural	Controlada
Tempo Médio da Viagem	180 min	45 min
Motoristas	10	4
Gênero	6 M - 4 F	2 M - 2 F
Idade	25-61	20-53

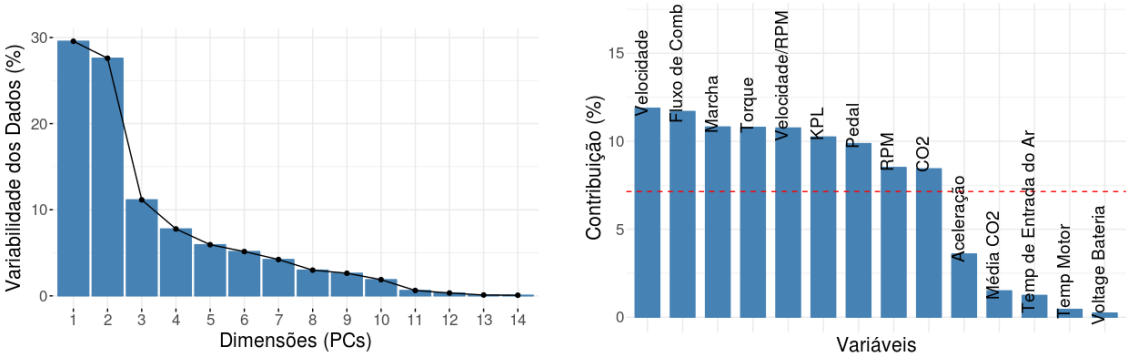
5. Preparação dos Dados

Em nossas análises, consideramos a premissa de que apenas os dados de sensores (reais ou virtuais) do veículo são suficientes para prover informações valiosas sobre a identidade do motorista e seu comportamento.

Baseado nessa premissa, foram removidos os dados coletados do *smartphone* e os demais que apresentam valores inválidos ou que não refletem o comportamento do motorista como, por exemplo, a força de atrito do ar e nível de combustível. Sendo assim, foram preservadas 14 variáveis das 40 coletadas. A Tabela 1 destaca as variáveis selecionadas (*) para a próxima etapa da preparação dos dados. Nesta etapa, também foram gerados dados adicionais com base nos dados de sensores do veículo, com o objetivo de descrever melhor o comportamento do veículo e do motorista. O sensor de marcha foi desenvolvido em [Rettore et al. 2017], e o processo de tratamento dos dados foi guiado em [Rettore et al. 2016]. Este trabalho nos levou a eliminar e tratar problemas de dados como *outliers*, conflitos, incompletude, ambiguidade, correlação e disparidade.

Mesmo que o estágio de preparação reduza e trate os dados, o número de variáveis usadas para identificar os motoristas pode levar a tempos de processamento inviáveis, dependendo do contexto aplicado, enfatizando a necessidade de reduzir ainda mais a quantidade de dados utilizados no processo de classificação. Usamos um procedimento estatístico chamado *Principal Component Analysis (PCA)* para extrair um conjunto de variáveis relevantes. Esse processo extrai as características com maior variação de um conjunto de dados multivariáveis, representando-os como um conjunto de novas características chamadas Componentes Principais (PCs). Esses PCs representam as direções ao longo das quais a variabilidade nos dados é máxima.

A Figura 2(a) mostra a porcentagem de variância em 14 PCs, número de recursos avaliados. O primeiro componente principal tem a maior variação possível. Em outras palavras, o primeiro PC contém a maior variabilidade nos dados. E cada componente sucessor contém a maior variação possível menor do que o antecessor. Os vetores resultantes são um conjunto ordenado não correlacionado. Depois disso, considerando os



(a) Componentes principais ordenados pela porcentagem de variabilidade

(b) Relevância dos dados considerando os primeiros dois componentes principais

Figura 2. Variáveis mais representativas do conjunto de dados

dois primeiros PCs, podemos explicar quase 60% da variância do conjunto de dados. A Figura 2(b) ilustra a variabilidade dos dodos explicada entre os dois primeiros componentes principais (também chamados de dimensões). Como podemos ver, cada variável possui uma variância explicada por mais de 60% (linha tracejada), ou seja, nove das quatorze variáveis representam a maior variabilidade de dados. Desse modo, esses dois PCs podem contribuir para melhor determinar o comportamento do motorista e sua identidade.

6. Identificação de Motoristas e Suspeitos

Um dos desafios na resolução de problemas de aprendizado de máquina é determinar o algoritmo certo para resolver uma dada questão. Isso porque o algoritmo adequado depende do conjunto de dados, dos resultados esperados, das restrições de tempo, do tamanho dos dados, da qualidade e natureza dos dados. Considerando estes fatores, foi conduzido um estudo com o objetivo de encontrar mecanismos que possam servir de guia para resolver as questões mencionadas. Feito isso, podemos usar ferramentas para selecionar um algoritmo de aprendizagem e seus hiper-parâmetros automaticamente, utilizando-se uma abordagem exploratória de algoritmos de hiper-parâmetros. Desse modo, dentre as ferramentas AutoML –*Auto Machine Learning* – mais conhecidas, foi escolhida a mais recente, TPOT [Olson et al. 2016], para explorar os algoritmos e configurações de hiper-parâmetros possíveis que mais se adeque aos dados de sensores veiculares.

Antes de submeter os dados à ferramenta TPOT, é necessário particioná-los. Para isso, foram criadas duas abordagens de particionamento, sendo: (1) Viagem: foram considerados todos os dados rotulados disponíveis de viagens e motoristas, dividindo-os em subconjuntos de treinamento (70%) e teste (30%). Esse particionamento considera o inicio de todas as viagens como base no treinamento e o fim como a base de teste. Também permite capturar um conjunto mais abrangente de comportamentos de cada motorista entre suas viagens. Esse conjunto de dados pode identificar o motorista em diferentes interações com ambiente (trajetos) e veículo; (2) Aleatório: o particionamento foi conduzido de forma aleatória, com o objetivo de eliminar o viés que pode ser introduzido no

particionamento por viagens. Posteriormente, todos os dados de treinamento e teste dos motoristas foram agrupados, resultando em uma base de treinamento final e uma base de teste final, respectivamente.

Depois de executar o TPOT, considerando as partições criadas, o algoritmo *Extremely Randomized Trees or Extra-Tree (ET)* [Geurts et al. 2006] foi selecionado por apresentar melhores resultados com os dados brutos. Esse algoritmo é usado para realizar a classificação ou regressão e exige que todos os preditores sejam numéricos e valores ausentes não são permitidos. Por exemplo, falhas na comunicação entre os sensores e o dispositivo de armazenamento e falhas de leitura inerentes aos sensores, gerando valores faltantes ou com erros. O algoritmo Extra-Trees cria um conjunto de árvores de decisão sem poda, usando uma estratégia *top-down*. As principais diferenças entre os outros métodos baseados em árvores estão relacionadas à forma de dividir os nós: ET escolhe aleatoriamente os pontos de corte e usa toda a amostra de aprendizagem para aumentar as árvores.

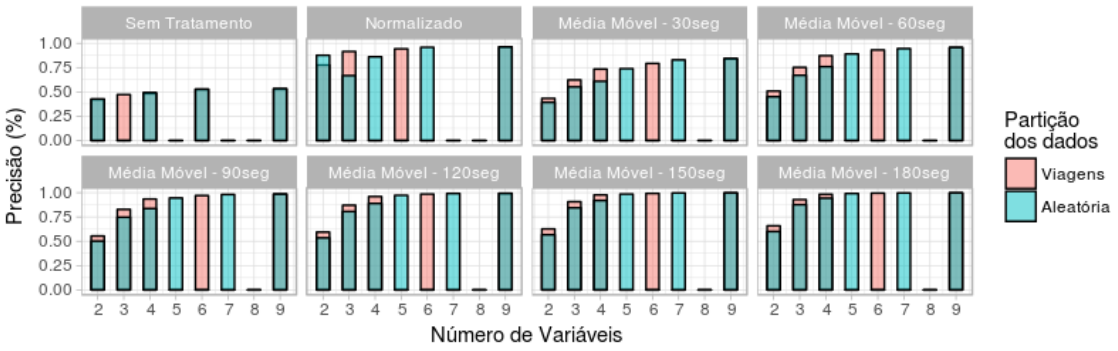
O algoritmo Extra-Tree foi avaliado em termos da precisão e número de variáveis consideradas pelo classificador. Assim, primeiro foi realizada a classificação usando dados brutos (sem tratamento), mas os resultados não foram suficientes para atender o objetivo em direção a serviços de assistência ao motorista e rede personalizados. Desse modo, a avaliação considerou inicialmente nove variáveis e, a cada iteração, a redução foi aplicada com base na importância da variável para o classificador. Esse procedimento é realizado por meio da métrica de importância da característica/variável (*feature importance*) incluída nos pacotes padrões do algoritmo de floresta aleatória. Uma maneira de calculá-lo pode ser a contabilização do número de vezes que um conjunto de dados passa por um nó cuja decisão é baseada em uma determinada característica. Se essa característica aparecer com frequência, então, mais importante é a contribuição dessa variável para a função de predição.

Também foi investigado o uso de janelas temporais de observação, semelhante a [Aoude et al. 2011, Carmona et al. 2015, Zhang et al. 2016], onde é processado o conjunto de dados criando um novo subconjunto que é calculado usando média móvel. Desta forma, é possível explorar diferentes tamanhos de médias móveis e suas consequências para o classificador. Foram avaliados os dados brutos, normalizados e médias móveis com tamanhos de 30, 60, 90, 120, 150 e 180 segundos de observação e de nove a duas variáveis. Além disso, duas métricas de particionamento dos dados foram utilizadas (por viagem, aleatória), com o objetivo de comprovar a validade das abordagens. O número de variáveis foi escolhido considerando sua importância, acima de 85%, para a função de predição. Por esse motivo, as características são destacadas de forma diferente entre os veículos, tipo de processamento e particionamento dos dados, tornando esse processo uma abordagem personalizada para identificar motoristas e suspeitos.

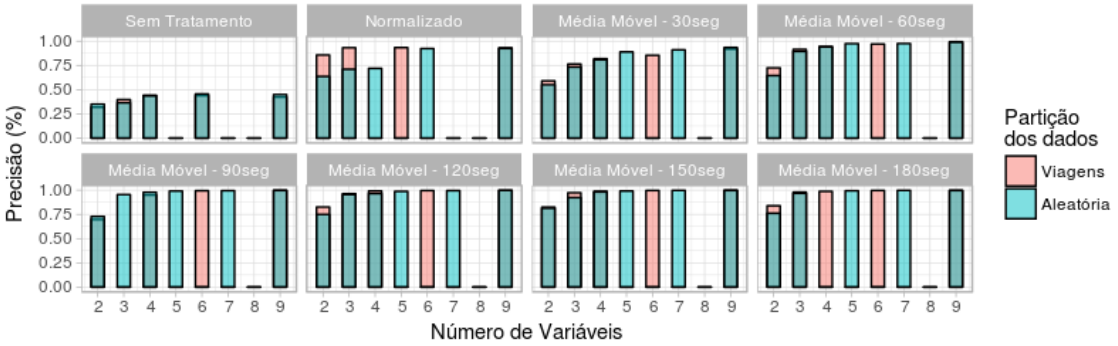
6.1. Avaliação da Identificação de Motoristas

Considerando a métrica *Viagem* de particionamento de dados, ao avaliar a classificação com os dados brutos (sem tratamento), foi observado que a precisão do classificador alcançou 54% com nove variáveis caindo para 43% quando apenas duas foram consideradas, e 42% com nove variáveis caindo para 39% com duas, para os Veículo #1 (Figura 3(a)) e #2 (Figura 3(b)), respectivamente. A normalização dos dados representou o

primeiro tratamento realizado nos dados, com o objetivo de avaliar o comportamento do classificador e determinar qual o ponto de corte ideal para cada veículo. Nessa avaliação, o Veículo #1 com nove variáveis apresentou 96% caindo para 77% com duas, enquanto o Veículo #2 apresentou 93% caindo para 85% com nove e duas variáveis.



(a) Veículo 1



(b) Veículo 2

Figura 3. Precisão vs. número de variáveis usando diferentes tratamentos nos dados

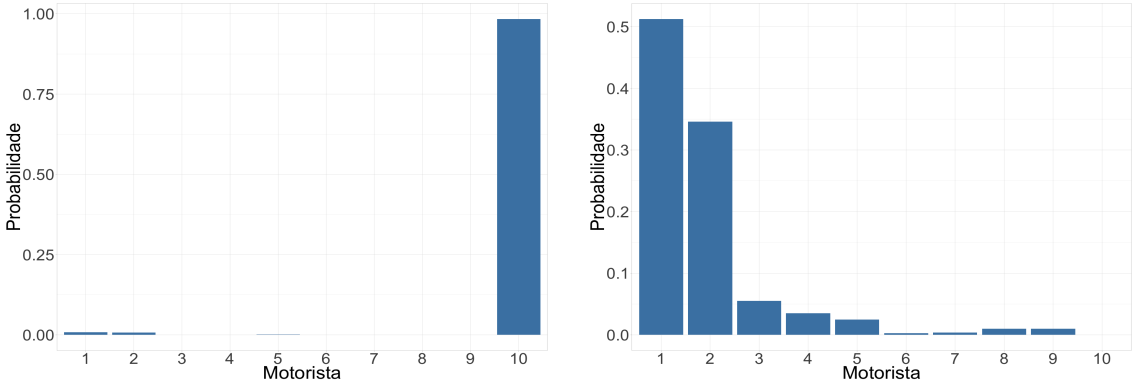
O segundo tratamento nos dados, fez uso de médias móveis com janelas entre 30 e 180 segundos. Esse processamento possibilitou o aumento da precisão do classificador e a redução das características/variáveis avaliadas. Isso ocorre porque o dado instantâneo do sensor faz da decisão uma tarefa difícil e confusa. Dessa forma, aplicando uma média móvel de 30 segundos no Veículo #1, a precisão foi superior a 83% com nove características, 79% com seis, 73% com quatro, 62% com três e 43% com duas características, mesmo resultado considerando os dados brutos e duas variáveis. Aumentando o tamanho da janela para 60 segundos de observação, foi observada uma melhora na precisão chegando a 95% com nove variáveis e 50% com duas. Esse comportamento de melhora se repete à medida que a janela aumenta, o que possibilitou identificar um limiar de parada para que a precisão da classificação seja superior a 98%. Foi considerado então, o tamanho da janela da média móvel de 120 segundos, resultando em 99% com nove variáveis, acima de 98% com seis e chegando a 60% com duas. Esse cenário se repete com o Veículo #2, contudo, é possível manter uma precisão acima de 99% com apenas quatro variáveis.

Essa investigação mostrou o custo-benefício entre precisão e número de características. A partir dessa análise exploratória, foram escolhidas as melhores relações para cada veículo, sendo seis e quatro características com uma média móvel de 120 segundos, para o Veículo #1 e #2, respectivamente. Isso levou a uma precisão acima de 98% para o Veículo #1 e 99% para Veículo #2. Ao considerar ambos os veículos, a precisão do classificador alcançou precisão superior a 98%. A diferença na precisão da identificação,

e também nos aspectos de desempenho (tempo de execução e memória – não discutidos neste texto), entre os veículos está relacionada às diferentes rotas utilizadas e à quantidade de dados coletadas, enquanto o Veículo #2 foi usado em rotas controladas com oito viagens e quatro motoristas, o Veículo #1 foi usado de forma normal com 26 viagens e 10 motoristas. Além disso, o Veículo #2 permite uma maior variação em sua condução, por ter motorização superior ao Veículo #1, resultando em uma melhor distinção entre os motoristas.

Os resultados obtidos possibilitaram definir a melhor configuração do método de classificação de motorista para cada veículo, permitindo o desenvolvimento de serviço personalizados de assistência ao motorista como entretenimento, ergonomia, serviços de rotas e serviços que auxiliam na eficiência de combustível. Além disso, essa configuração serve de entrada para a etapa de identificação de suspeitos (motoristas ilegítimos/intrusos), visando auxiliar os serviços em uma rede veicular, como permitir ou não a troca de mensagens entre os nós/veículos, entretenimento e sugestões personalizadas de rotas. Foi mantida como configuração o particionamento dos dados por *Viagem*, considerando sua leve melhora nos resultados, uma janela com média móvel de 120 segundos, seis variáveis para o Veículo #1 e quatro variáveis para o Veículo #2.

6.2. Avaliação da Identificação de Suspeitos



(a) Legítimo (b) Intruso

Figura 4. Resultados do classificador quando trata o motorista 10 como legítimo e intruso

A simulação da presença de suspeitos entre os motoristas parte da premissa de que não se conhece a maneira como esse motorista dirige. Essa condição faz com que um suspeito dirija, sob a óptica da identificação de motoristas conhecidos, de maneira semelhante a vários desses condutores. Para simular a presença de suspeitos entre os motoristas dos veículos, cada um dos motoristas conhecidos foi tratado como um suspeito e seus dados foram removidos dos dados de treinamento.

Usando os dados produzidos pelo classificador, avaliado na Seção 6.1, treinado tanto com o conjunto completo, quanto com conjuntos faltando dados de cada um dos motoristas, foi possível identificar motoristas suspeitos. Avaliando o comportamento desse primeiro classificador, é possível perceber que existem diferenças na precisão e na distribuição dos seus resultados quando submetidos a dados de motoristas conhecidos e intrusos. A Figura 4 mostra as distribuições de probabilidade em dois casos: quando o motorista 10 é identificado em uma viagem e quando esse mesmo motorista é tratado como um intruso no veículo de testes.

Apesar de existirem diferenças visíveis entre ambas distribuições, elas nem sempre podem ser notadas, ou metodologicamente diferenciadas. Assim, um novo classificador é responsável por diferenciar distribuições de probabilidades dos valores produzidos pelo classificador de motoristas quando o motorista é autêntico ou intruso. Para diferenciar essas situações, o classificador usa os dados produzidos pela identificação de motoristas conhecidos. Ao invés de usar somente o último resultado do primeiro classificador, a identificação de suspeitos usa a distribuição de probabilidade dos resultados até o momento. Treinando um segundo classificador com as distribuições de motoristas suspeitos e legítimos, foi possível alcançar precisões acima de 99%. Uma característica importante desse processo é que, assim como a identificação de motoristas conhecidos, separar suspeitos de motoristas legítimos é uma tarefa que não depende de dados vindos da rede, o que possibilita o processamento no próprio elemento computacional.

7. Dados Suspeitos em Redes Veiculares

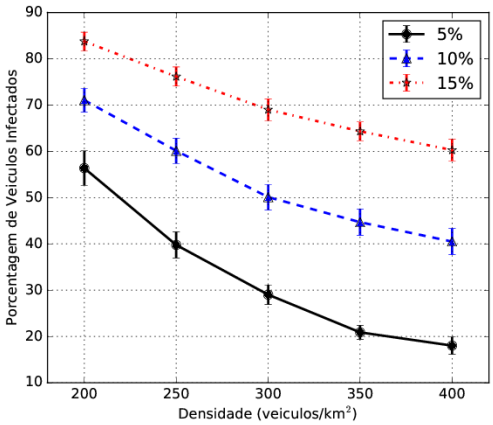


Figura 5. Propagação de veículos infectados na rede veicular

Com o objetivo de analisar o impacto que veículos suspeitos podem ter nos serviços de comunicação em redes veiculares, esta seção apresenta um estudo considerando uma disseminação de dados em um centro urbano. Neste cenário, um veículo fonte, o qual não é um veículo suspeito, dissemina 100 pacotes de dados para todos os veículos contidos em um Manhattan Grid com 10 ruas de sentido duplo na horizontal e vertical em uma área de 1 km^2. O protocolo de disseminação utilizado é o de inundação tradicional (Flooding), em que um veículo, ao receber um pacote de dados não duplicado, o repassa para seus vizinhos. Durante a simulação, variou-se a densidade de veículos no centro urbano (200, 250, 300, 350 e 400 veículos/km^2) e a quantidade de veículos suspeitos (5, 10 e 15%). Este cenário foi implementado utilizando-se o framework de simulação OMNeT++ 4.2.2, o simulador de redes veiculares Veins 2.1 e o simulador de mobilidade SUMO 0.17.0. Definiu-se a potência de transmissão em 0,98 mW, resultando em um raio de comunicação de aproximadamente 200 m. Replicações foram realizadas como forma de obter um intervalo de confiança de 95%.

A Figura 5 mostra a propagação de veículos infectados durante o processo de disseminação. Um veículo é dito infectado se ele recebeu dados diretamente de um veículo suspeito ou se os dados recebidos passaram por um veículo suspeito em algum momento durante o processo de disseminação. Conforme pode-se observar, em cenários com densidades mais baixas, a presença de uma pequena porcentagem de veículos suspeitos (5%), pode resultar em uma quantidade de infecções de mais de 50%. Conforme a

densidade aumenta, a quantidade de infecções diminui. No entanto, dependendo da quantidade de veículos suspeitos na rede, a quantidade de veículos infectados pode ultrapassar 40%. Este resultado mostra que a presença de veículos suspeitos pode comprometer a qualidade de serviços em redes veiculares. Por exemplo, veículos suspeitos podem alterar os dados que estão sendo disseminados na rede. Esse resultado mostra a importância em desenvolver mecanismos de veículos suspeitos em uma rede veicular.

8. Conclusões e Trabalhos Futuros

A capacidade de comunicação e sensoriamento dos veículos possibilita o desenvolvimento de uma variedade de aplicações e serviços, como serviços para gerenciar e oferecer maior segurança às pessoas no transito, além de serviços de conforto para motoristas e passageiros. Muitos desses sistemas necessitam/deveriam autenticar seus usuários, para que o conteúdo seja direcionado, porém o fazem de maneira que um motorista invasor possa utilizá-los.

Neste trabalho, propusemos um sensor virtual para determinar localmente quem está dirigindo o veículo em determinado momento. Exploramos a identificação de motoristas como fator extra de autenticação em serviços de assistência ao motorista e serviços de redes veiculares. O algoritmo de classificação demonstrou ser simples e eficiente, mantendo sua precisão acima de 98%. Além disso, discutimos a importância da abordagem no contexto de VANETs, simulando um cenário onde o motorista suspeito é identificado na rede e avaliamos seu impacto na disseminação de dados, já que esse suspeito pode modificar a informação, comprometendo a rede.

Observamos que os algoritmos de classificação mais comuns encontrados na literatura com o objetivo de caracterizar o comportamento do motorista são *Support Vector Machines (SVM)*, *Hidden Markov Models (HMM)* e *Fuzzy Inference Systems (FIS)*. Em contraste, o algoritmo de classificação utilizado, Extra-Trees, demonstrou ser simples e eficiente, mantendo sua precisão acima de 98%, considerando seis características analisadas do Veículo #1 e quatro características do Veículo #2 com uma janela de 120 segundos de média móvel. Este classificador foi utilizado no reconhecimento de motoristas legítimos e suspeitos. Notamos que o identificador de motoristas se comporta de maneira diferente quando submetido aos dados de um motorista legítimo e um motorista suspeito, e esse comportamento se reflete em distribuições de probabilidade visualmente diferentes. O resultado do classificador treinado para distinguir entre os dois tipos de distribuições atingiu precisão acima de 99%.

Como extensão deste trabalho, pretendemos expandir a autenticação comportamental de motoristas, embarcando o sistema no veículo. Investigar do custo computacional da autenticação, levando em consideração as características utilizadas. Além disso, pretendemos avaliar soluções para contornar a presença de suspeitos em VANETs.

Referências

Aoude, G. S., Desaraju, V. R., Stephens, L. H., and How, J. P. (2011). Behavior classification algorithms at intersections and validation using naturalistic data. *IEEE Intelligent Vehicles Symposium, Proceedings*, (Iv):601–606.

Bergasa, L. M., Almeria, D., Almazan, J., Yebes, J. J., and Arroyo, R. (2014). DriveSafe: An app for alerting inattentive drivers and scoring driving behaviors. *IEEE Intelligent Vehicles Symposium, Proceedings*, (Iv):240–245.

Burton, A., Parikh, T., Mascarenhas, S., Zhang, J., Voris, J., Artan, N. S., and Li, W. (2016). Driver identification and authentication with active behavior modeling. In *12th International Conference on Network and Service Management (CNSM)*.

Carmona, J., García, F., Martín, D., Escalera, A., and Armingol, J. (2015). Data Fusion for Driver Behaviour Analysis. *Sensors*, 15(10):25968–25991.

Fleming, W. J. (2001). Overview of Automotive Sensors. *IEEE Sensors Journal*, 1(4):296–308.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63(1):3–42.

Hallac, D., Sharang, A., Stahlmann, R., Lamprecht, A., Huber, M., Roehder, M., Leskovec, J., et al. (2016). Driver identification using automobile sensor data from a single turn. In *Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on*, pages 953–958. IEEE.

Johnson, D. A. and Trivedi, M. M. (2011). Driving style recognition using a smartphone as a sensor platform. In *2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pages 1609–1615. IEEE.

Lin, J., Chen, S., Shih, Y., and Chen, S.-h. (2009). A study on remote on-line diagnostic system for vehicles by integrating the technology of OBD, GPS, and 3G. *World Academy of Science, Engineering and Technology*, 32(8):435–441.

Martínez, M., Echanobe, J., and del Campo, I. (2016). Driver identification and impostor detection based on driving behavior signals. In *Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on*, pages 372–378. IEEE.

Olson, R. S., Bartley, N., Urbanowicz, R. J., and Moore, J. H. (2016). Evaluation of a tree-based pipeline optimization tool for automating data science. In *Proceedings of the Genetic and Evolutionary Computation Conference 2016*, GECCO '16, pages 485–492, New York, NY, USA. ACM.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Qu, F., Wang, F. Y., and Yang, L. (2010). Intelligent transportation spaces: Vehicles, traffic, communications, and beyond. *IEEE Communications Magazine*, 48(11):136–142.

Rettore, P. H., André, B. P. S., Campolina, Villas, L. A., and A.F. Loureiro, A. (2016). Towards intra-vehicular sensor data fusion. In *Advanced perception, Machine learning and Data sets (AMD'16) as part of the 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC 2016)*, Rio de Janeiro.

Rettore, P. H. L., Campolina, A. B., Villas, L. A., and Loureiro, A. A. F. (2017). A method of eco-driving based on intra-vehicular sensor data. In *2017 IEEE Symposium on Computers and Communications (ISCC)*, pages 1122–1127, Heraklion, Greece. IEEE.

Salemi, M. (2015). *Authenticating drivers based on driving behavior*. Rutgers The State University of New Jersey-New Brunswick.

Silva, H., Lourenço, A., and Fred, A. (2012). In-vehicle driver recognition based on hand ecg signals. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*.

Yuan, W. and Tang, Y. (2011). The driver authentication device based on the characteristics of palmprint and palm vein. In *International Conference on Hand-Based Biometrics*, pages 1–5.

Zhang, C., Patel, M., Buthpitiya, S., Lyons, K., Harrison, B., and Abowd, G. D. (2016). Driver Classification Based on Driving Behaviors. *Proceedings of the 21st International Conference on Intelligent User Interfaces - IUI '16*, pages 80–84.