SISTEMA DE IDENTIFICAÇÃO DE CONDUTORES BASEADO EM MÉTODOS DE EXTRAÇÃO DE CARACTERÍSTICAS ESTATÍSTICAS E TÉCNICAS DE REDUÇÃO DE DIMENSIONALIDADE

Andrey Gustavo de Souza*, Wilian Soares Lacerda*, Danton Diego Ferreira*, Gustavo Lobato Campos†

*Departamento de Engenharia
Universidade Federal de Lavras
Lavras, Minas Gerais, Brasil

†Departamento de Engenharia
Instituto Federal de Minas Gerais
Formiga, Minas Gerais, Brasil

Emails: andreygsouza@hotmail.com, lacerda@dcc.ufla.br, danton@deg.ufla.br, gustavo.lobato@ifmg.edu.br

Abstract— The increasing number of stolen vehicles, specialy in Brazil, shows a demand for systems that is to avoid this type of occasion. Aware of this, this work seeks the development of a driver identification system by means of data from sensors embedded in the vehicle, along with the application of data processing and machine learning techniques. The system makes use of statistical feature extraction, in order to maximize the representativeness of the data and also dimensionality reduction techniques, like Fisher Discriminant Analysis (FDA), Principal Component Analysis (PCA), Incremental Principal Component Analysis (IPCA) and Independent Component Analysis (ICA), in order to optimize the performance of the classifier. The system was efficient in carrying out the proposed task with a correct driver identification rate of over 99%.

Keywords— Driver identification, driver behavior modeling, data processing, machine learning.

Resumo— O crescente número de roubos e furtos de veículos no Brasil, mostra a demanda iminente por sistemas que possam evitar este tipo de ocasionalidade. Ciente disto, este trabalho busca o desenvolvimento de um sistema identificação de condutores por meio de dados oriundos de sensores embarcados no próprio veículo, juntamente com a aplicação de técnicas de processamento de dados e aprendizado de máquina. O sistema faz uso de métodos de extração de características estatísticas, a fim de se maximizar a representatividade dos dados e também de técnicas de redução de dimensionalidade, como Análise Discriminante de Fisher (FDA), Análise de Componentes Principais (PCA), Análise de Componentes Principais Incrementais (IPCA) e Análise de Componentes Independentes (ICA), com o objetivo de otimizar o desempenho dos classificadores. O sistema se mostrou eficiente na execução da tarefa proposta, com uma taxa de correta identificação do condutor superior à 99%.

Palavras-chave— Identificação de condutores, modelagem de comportamento do condutor, processamento de dados, aprendizado de máquina.

1 Introdução

A cada ano, o número de veículos roubados ou furtados tem crescido consideravelmente. Somente no Brasil, em 2015, 509.978 veículos foram roubados (562,4 por 100 mil veículos), de acordo com o Fórum Brasileiro de Segurança Pública (2017). Uma série histórica é apresentada na Figura 1 para ilustrar o crescimento no número de ocasionalidades desta natureza. Estes números contribuem para a discussão de novas aplicações que possam reduzir estes números. Tais aplicações trariam benefícios para os proprietários de veículos, companhias de seguro e forças de segurança pública, que diretamente sofrem com este problema. Portanto, novas soluções tecnológicas que evitem esse tipo de situação são promissoras, especialmente quando conseguem impedir rapidamente a ação criminosa. Neste contexto, surge o conceito de Sistemas de Identificação de Condutores.

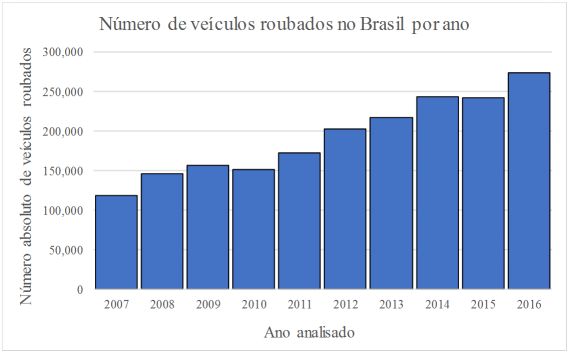


Figura 1: Veículos Roubados no Brasil por ano. Fonte: Adaptado do Fórum Brasileiro de Segurança Pública (2017).

Sistemas de Identificação de Condutores podem ser definidos como técnicas utilizadas para autenticar o usuário do veículo e possivelmente evitar que quaisquer condutores não autorizados a operar o veículo o utilizem. O sistema preferencialmente deve cumprir sua tarefa somente com o uso de recursos mínimos, sem contato direto com o condutor (tal como autenticação biométrica). Isto pode ser implementado de diversas formas, sendo

uma destas através do uso de dados de direção extraídos do barramento CAN (CAN-bus) e sensores externos a este (Sensores inerciais, GPS, entre outros). Estes dados carregam em si certos padrões que podem ser utilizados para identificar um indivíduo específico. Sendo assim, este trabalho pretende apresentar o desenvolvimento de um modelo computacional de identificação de condutores baseado em técnicas de processamento de dados e inteligência computacional.

2 Extração de Características

Amplamente utilizada em aprendizado de máquina, reconhecimento de padrões e processamento de imagem, a extração de características começa por um conjunto inicial de dados medidos e retorna valores derivados (características) com a intenção de serem informativas e não redundantes, melhorando a capacidade de aprendizagem e generalização do classificador e, em alguns casos, levando a uma melhor interpretação dos dados (Guyon and Elisseeff, 2006). Neste trabalho foi empregado como técnica de extração de características as Estatísticas de Ordem Superior.

O termo Estatística de Ordem Superior (EOS) refere-se à funções que usam a terceira potência ou superiores de uma certa amostra que fazem uso de termos constantes, lineares e quadráticos (potências zero, um e dois). EOS pode ser definido em termos de momentos e cumulantes. Enquanto momentos são próprios para descrever sinais determinísticos, cumulantes são adequados para a análise de sinais probabilísticos (Guedes et al., 2016). Considere x como um processo aleatório, real e discreto com média zero. Dado um caso unidimensional ($d = 1$), os momentos de uma variável aleatória X podem ser definidos como (Naves et al., 2016):

$$\begin{aligned} m_1 &= \langle x \rangle \\ m_2 &= \langle x^2 \rangle \\ &\vdots \\ m_n &= \langle x^n \rangle \end{aligned} \tag{1}$$

Os cumulantes podem ser escritos em forma de momentos:

$$\begin{aligned} c_1 &= m_1 \\ c_2 &= m_2 - {m_1}^2 = \sigma^2 \\ c_3 &= m_3 - 3m_1m_2 + 2{m_1}^3 \\ c_4 &= m_4 - 3{m_2}^2 - 4m_1m_3 \\ &\quad +12{m_1}^2m_2 - 6{m_1}^4 \end{aligned} \tag{2}$$

Os cumulantes c_1, c_2, c_3 e c_4 são a média, a variância, a assimetria e a curtose, respectivamente. Dentre os quais somente a média não é uma EOS.

Como descrito em Mendel (1991), as EOS podem levar a resultados mais representativos quando aplicadas como ferramenta de extração de características em processos não lineares e não gaussianos. De acordo com Naves et al. (2016), a maior vantagem do uso dos cumulantes como extrator de características em problemas de classificação é sua propriedade de imunidade à ruido gaussiano.

3 Redução de Dimensionalidade

Um método útil para melhorar a eficiência da tarefa de classificação em termos de custo computacional é a redução de dimensionalidade. Estas ferramentas podem ser definidas como o processo de redução do número de variáveis de entrada pela obtenção de um conjunto de variáveis principais. Em outros termos, estas técnicas transformam um conjunto de dados de amplo espaço dimensional em um espaço de menor dimensão (Wang and Chang, 2006).

A principal técnica para redução de dimensionalidade é a Análise de Componentes Principais. Outra técnica que pode ser explorada para esta tarefa é a Análise de Componentes Independentes, embora esta não seja convencionalmente usada para redução de dimensionalidade, mas para a separação de sinais sobrepostos. Contudo, existem exemplos na literatura que aplicam Análise de Componentes Independentes para redução de dimensionalidade (Wang and Chang, 2006) (Cao et al., 2003).

3.1 Análise de Componentes Principais e PCA Incremental

Análise de Componentes Principais (PCA) é um procedimento matemático que usa uma transformação ortogonal para para converter um conjunto de variáveis possivelmente correlatadas em um conjunto de variáveis não correlacionadas chamadas de componentes principais (Wold et al., 1987). Sejam $x_t (t = 1, \ldots, l$ e $\sum_{t=1}^{l} x_t = 0)$ um conjunto de vetores de entrada de m dimensões $x_t = (x_t(1), x_t(2), \ldots, x_t(m))^T$, então PCA transforma linearmente x_t em um novo vetor s_t pela seguinte expressão (Cao et al., 2003):

$$s_t = U^T x_t, \tag{3}$$

onde U é uma matriz ortogonal de $m \times m$ dimensões no qual a i-ésima coluna u_i é o i-ésimo autovetor da matriz de covariância C. Portanto, PCA primeiro soluciona o problema dos autovalores:

$$\lambda_i u_i = C u_i, \qquad i = 1, \ldots, m, \tag{4}$$

onde λ_i é o i-ésimo autovalor de C e u_i é seu o autovetor relativo. Então, os componentes de s_t, baseado no obtido em u_i, são calculados como a transformação ortogonal de x_t por:

$$s_t(i) = {u_i}^T x_t, \qquad i = 1, \ldots, m. \tag{5}$$

Estes novos componente $s_t(i)$ são os componentes principais. Tal qual a equação 5 demonstra, o número de componentes principais de s_t pode ser reduzido pelo uso de somente alguns dos primeiros autovetores ordenados na ordem descendente dos autovalores. Assim, pode-se concluir que PCA tem a característica de reduzir a dimensão de um conjunto de dados (Cao et al., 2003).

O PCA é uma ferramenta útil para problemas de redução de dimensionalidade, porém possui certas limitações quando utilizados em *datasets* maiores. Isto se dá porque o processamento do PCA é feito em lotes, que faz com que todos os dados a serem processados devam ser armazenados na memória do dispositivo utilizado. Isto pode ser um problema para aplicações em *hardware* embarcado que tem limitações quanto a capacidade de memória (Pedregosa et al., 2011).A técnica de PCA Incremental (IPCA) contorna este problema, utilizando uma forma diferente de processamento que permite cálculos parciais que praticamente obtém os mesmos resultados do PCA, porém realizando o processamento em mini lotes (Weng et al., 2003).

A versão incremental do algoritmo funciona da seguinte maneira: assume-se que já de posse do conjunto de autovetores $U = [u_j], j = 1, \ldots, p$ do vetor de entrada $x_i, i = 1, \ldots, n$. os autovalores correspondentes são $\lambda = diag(\Lambda)$ e a média dos valores é $\overline{x}$. A construção dos incrementos requer a atualização destes autovalores e autovetores levando em conta uma nova entrada x_{n+1} (Artac et al., 2002). Este procedimento resulta em um processamento mais eficiente em termos de utilização de memória (Balsubramani et al., 2013).

3.2 Análise Componentes Independentes

Análise Componentes Independentes (ICA) (Lee, 1998) é uma técnica que originalmente foi desenvolvida para separação cega de fontes, que recupera sinais mutualmente independentes, mas com fontes desconhecidas a partir de suas misturas lineares sem saber os coeficientes de mistura.

ICA considera que os dados são linearmente combinados por um conjunto de fontes independentes e estes sinais podem ser separados de acordo com sua independência estatística (Wang and Chang, 2006). Seja x_t a mistura linear e s_t denota o sinal original, então o objetivo do ICA é estimar s_t por:

$$s_t = U x_t, \tag{6}$$

onde U é a matriz $m \times m$ de separação de misturas. Os componentes retornados por s_t são tão independentes estatisticamente quanto possível.

Existem um amplo número de algoritmos que foram desenvolvidos para executar o ICA. Neste trabalho, considerou-se o FastICA de ponto fixo, proposto por Hyvarinen (1999). Este algoritmo é considerado um dos melhores métodos já desenvolvidos e também o mais aplicado. FastICA faz uso de informações mútuas como critério para estimar s_t, ao passo que também é uma medida natural de independência entre variáveis aleatórias. A maximização da negentropia (medida do grau de organização do sistema) corresponde à minimização das informações mútuas entre os componentes. Entretanto, esta negentropia não pode ser feita diretamente uma vez que as densidades de probabilidade dos componentes são desconhecidos. A explanação completa do funcionamento do algoritmo FastICA é encontrada em (Koldovsky et al., 2006).

As principais duas diferenças entre o PCA e o ICA são, primeiramente, os componentes retornados pelo ICA são estatisticamente independentes, não simplesmente descorrelacionadas tal qual ocorre nos gerados pelo PCA. A segunda distinção é que a matriz de separação de misturas do ICA não é ortogonal tal qual a do PCA (Cao et al., 2003).

3.3 Análise de Discriminante de Fisher

Análise de Discriminante de Fisher (FDA) é uma técnica de redução de dimensionalidade, otimizado em termos da maximização da separação entre classes. Dado um conjunto de dados de $n \times m$ dimensões representado pela matriz X com vetor coluna x_i, a matriz de dispersão total é dada por (Chiang et al., 2004):

$$S_t = \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^T, \tag{7}$$

onde μ é o vetor de médias totais dos elementos correspondentes às colunas de X. Considerando X_j como o conjuntos de vetores x_i que pertencem à uma determinada classe j, ma matriz de dispersão interna para a classe j é

$$S_t = \sum_{x_i \in X_j}^{n} (x_i - \mu_j)(x_i - \mu_j)^T, \tag{8}$$

onde μ_j é o vetor de médias para a classe j. Considerando c o número de classes dos dados, então

$$S_w = \sum_{i=1}^{c} S_i, \tag{9}$$

é a matriz de dispersão entre classes, onde n_j é o número de observações na classe j.

O primeiro vetor FDA w_1 pode ser determinado como:

$$\underset{w_1}{\text{argmax}} \frac{{w_1}^T S_b w_1}{{w_1}^T S_w w_1} \tag{10}$$

O segundo vetor FDA é calculado de modo a maximizar a dispersão entre classes enquanto minimiza a dispersão entre classes entre todos os eixos perpendiculares com o primeiro vetor FDA em seguida com os vetores FDA restantes. Pode ser provado matematicamente que os vetores FDA são iguais aos autovetores w_k do problema de autovalores generalizados

$$S_b w_k = \lambda_j S_w w_k, \tag{11}$$

onde os autovalores λ_k indicam o grau geral de separabilidade entre as classes pela projeção de todas as classes em w_k. Com os vetores FDA calculados, as observações são então classificadas de forma a reduzir o espaço dos vetores FDA por meio de uma análise discriminante (Sugiyama, 2007).

4 Materiais e Métodos

Nesta seção são apresentados os insumos utilizados para a implementação do sistema, bem como os procedimentos experimentais que foram adotados no desenvolvimento do mesmo.

4.1 O Dataset

Neste trabalho, foi utilizado o *dataset* disponibilizado por Kwak et al. (2016), juntamente com o *Hacking and Countermeasure Research Lab*[1]. Este *dataset* consiste em dados extraídos do barramento CAN do veículo, coletados por por meio da interface *On Board Diagnostics 2* (OBD-II) do veiculo e do *scanner* CarbigsP. O veículo usado na coleta dos dados (Kia Soul) possui diversos sensores e atuadores que fornecem um montante representativo de dados que podem ser usados para detectar certos padrões inerentes ao comportamento do condutor e até mesmo sua identidade. Estes dados foram amostrados em uma frequência de 1Hz com 51 sensores medidos.

Na coleta de dados, dez indivíduos conduziram o veículo em quatro rotas distintas em Seul, Coreia do Sul. Os autores consideraram três tipos de vias: urbano, rodovia e estacionamento. Cada condutor foi submetido à duas sessões de direção que passa por todos os tipos de vias, totalizando 23 km.

A grande vantagem desde conjunto de dados em comparação à outros utilizados em abordagens similares é que este *dataset* é mais recente, colhido a partir de julho de 2015, e mais completo visto que veículos mais novos possuem mais sensores embarcados. Outra vantagem relevante do mesmo é a preocupação que os autores tiveram em relação à confiabilidade dos dados, uma vez que os experimentos foram conduzidos em situações similares, como o fator temporal (entre 8 da manhã e 11 da noite em dias de semana) e duas viagens diferentes para cada condutor, aumentando a extração de traços característicos de direção de cada condutor, e portanto, proporcionando uma classificação de condutores mais eficiente.

O conjunto de dados fornecidos tem um total de 94.401 amostras, com amostragem de 1Hz, com tamanho total de 16,7Mb.

4.2 Seleção de variáveis

O conjunto de dados é dividido em dez condutores numerados entre 0 e 9. Porém o objetivo do trabalho é classificar os condutores entre não autorizado e autorizado. Para tal, foram selecionados aleatoriamente três condutores que foram definidos como autorizados (classe 1), como membros de uma mesma família ou empresa que compartilham o mesmo veículo, e restante dos condutores foram definidos como não autorizados (classe 0).

A implementação do sistemas de testes passa inicialmente pela seleção de quais variáveis disponíveis no conjunto de dados colhidos no barramento CAN serão utilizadas na tarefa de identificação dos condutores. Neste estágio, foi utilizado a técnica FDA, implementado para Python por Li et al. (2016) especialmente para seleção de características para classificadores. Foram selecionadas as oito variáveis de entrada com maior relevância, as quais são descritas na Tabela 1.

Tabela 1: Variáveis de entrada selecionadas.

Variável	Descrição
1	Valor do pedal de acelerador
2	Ângulo de esterçamento do volante
3	Velocidade do Veículo
4	Velocidade do Motor
5	Abertura da Válvula reguladora
6	Aceleração Longitudinal
7	Aceleração Lateral
8	Pressão do ar de entrada

4.3 Extração de Características

De posse destas variáveis, passa-se agora ao estágio de extração de características de forma a minimizar os efeitos de flutuação dos valores e caracterizar a distribuição de características efetivamente. Além dos valores de cumulantes em atraso zero (variância (segunda ordem), assimetria (terceira ordem) e curtose (quarta ordem)), foram utilizados os valores de mediana e média, totalizando assim cinco características estatísticas por variável.

Estas estatísticas foram obtidas por meio de certo período de amostragem determinadas por meio de um janelamento temporal. De forma a também determinar qual o intervalo de tempo de janelamento é mais adequado para a tarefa de identificação dos condutores, foram testadas janelas de 15, 30, 60, 90 e 120 segundos.

[1] http://ocslab.hksecurity.net

4.4 Redução da Dimensionalidade

Pode-se observar que no estágio de extração de características estatística, de cada variável de entrada são derivados cinco elementos, totalizando quarenta variáveis de entrada. Esta quantidade de valores dificulta o desempenho do classificador em termos de eficiência computacional e até comprometer a capacidade de generalização do mesmo. Sendo assim, mais uma vez foi aplicado o Discriminante de Fisher de forma a selecionar destas quarenta varáveis derivadas as dez mais significantes para a tarefa de classificação.

Ainda com a redução obtida no passo anterior, o número de variáveis de entrada pode dificultar o desempenho do classificador. Para contornar esta situação, foram adotados métodos de redução de dimensionalidade de forma a reduzir o espaço de entrada de dez variáveis para somente quatro, sem que ocorram perdas de características significativas. As técnicas testadas são as de Análise de Componentes Principais (PCA), Análise de Componentes Principais Incrementais (IPCA) e Análise de Componentes Independentes (ICA), todos considerando quatro componentes. Portanto o espaço de dez variáveis oriundos da seleção por meio do Discriminante de Fisher foi reduzido a um espaço de quatro componentes na entrada do classificador.

4.5 Classificação

Para a tarefa de identificação dos condutores foram considerados três algoritmos de classificação distintos. Os algoritmos foram implementados em Python utilizando a biblioteca *scikit-learn - Machine Learning in Python* (Pedregosa et al., 2011). Os classificadores foram configurados de acordo com os seguintes parâmetros:

- K – *Nearest Neighbor* (KNN): número de vizinhos: 3; pesos uniformes.
- *Random Forest* (RF): 50 estimadores; profundidade máxima: 10; amostras mínimas por folha: 10.
- Redes Neurais Artificiais (MLP): Número de camadas escondidas = 2, número de neurônios por camada = 10, máximo de épocas de treinamento = 500.

4.6 Configuração dos experimentos

Os experimentos foram conduzidos de forma a testar todas as combinações de técnicas possíveis de forma a se conseguir o maior número de acertos na identificação dos condutores. Estas combinações envolvem as cinco janelas temporais (15, 30, 60, 90 e 120 segundos), três técnicas de redução de dimensionalidade (PCA, IPCA e ICA) e três classificadores (KNN, RF e MLP). A combinação destes elementos totalizam quarenta e cinco cenários diferentes. Cada classificador é treinado e testado dez vezes e cada uma destas é efetuada a separação entre dados de treino (80%) e teste (20%). Em cada um dos testes é feita uma permutação aleatória dos dados, de forma a não comprometer a capacidade de generalização do classificador. Além disso, em cada uma das iterações são selecionados aleatoriamente diferentes condutores que serão definidos como autorizados, de modo a testar diferentes combinações.

O objetivo final dos testes é maximizar a taxa de acertos entre os condutores autorizados (classe 1) e não autorizados (classe 0). O fluxo geral dos experimentos é descrito na Figura 2.

4.7 Avaliação dos Modelos

Para se mensurar o desempenho dos classificadores, foram considerados algumas métricas adequadas para este tipo de problema. A primeira, e mais importante, é a medida de acurácia, que pode ser expressa por:

$$A_{cc}(R) = P(H|B) = \frac{P(HB)}{P(B)} = \frac{f_{hb}}{f_b} \qquad (12)$$

onde f_{hb} é a quantidade de classificações corretas e f_b é a quantidade total de classificações. A segunda métrica utilizada é o F1-*Score*, que é também uma medida de desempenho do modelo, que considera a precisão e a revocação dos testes. F1-*score* é obtido pela seguinte expressão:

$$F_1 = 2\frac{precision.recall}{precision + recall} \qquad (13)$$

O última medida colhida é a análise de concordância kappa (*Cohen's kappa score*), que varia entre 0 e 1, os quais podem ser interpretados de acordo com a Tabela 2. Uma explanação mais completa destas métricas pode ser encontrado em (Banerjee et al., 1999).

Tabela 2: Interpretação dos valores de Kappa

Kappa	Interpretação
<0	Sem concordância
0-0.19	Concordância baixa
0.20-0.39	Concordância razoável
0.40-0.59	Concordância moderada
0.60-0.79	Concordância substancial
0.80-1.00	Concordância quase perfeita

5 Resultados Experimentais

Um dos objetivos do presente trabalho é testar a aptidão de diferentes técnicas de processamento e classificação de dados para tarefa de classificação de condutores. Os gráficos da Figura 3 mostram o desempenho em termos da acurácia dos testes juntamente com a dispersão dos resultados dos classificadores, divididos em cada umas das técnicas de redução de dimensionalidade.

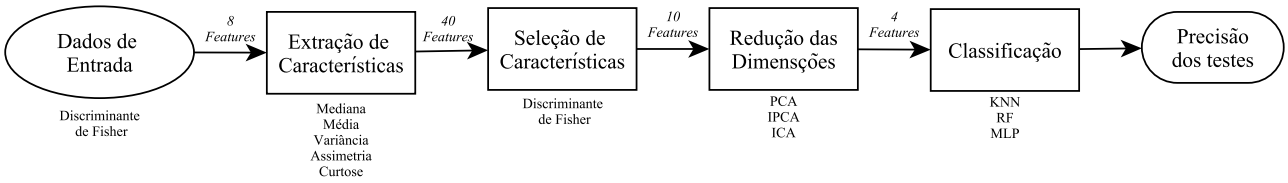
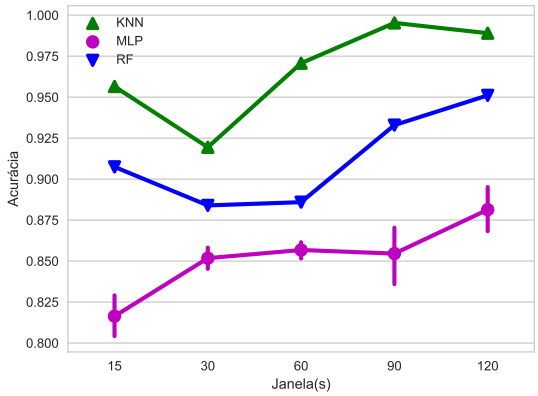
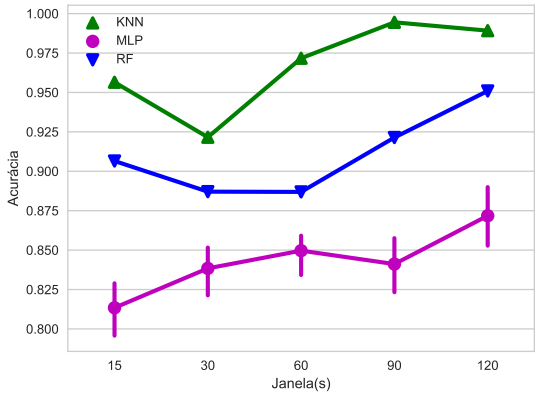


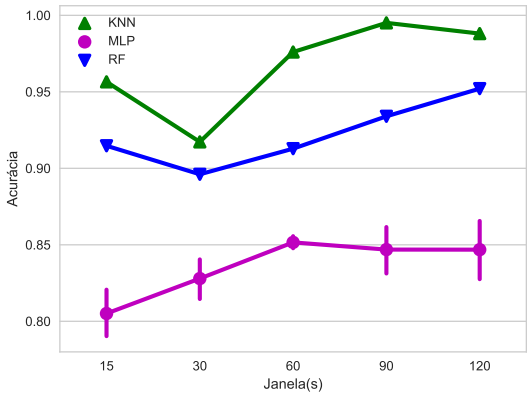
Figura 2: Fluxo de desenvolvimento e testes do sistema proposto.



(a) PCA



(b) IPCA



(c) ICA

Figura 3: Acurácia média de cada método de redução de dimensionalidade.

Em todos os cenários, o classificador KNN obteve desempenho superior em comparação aos outros classificadores, enquanto o MLP teve um desempenho consideravelmente inferior, além de apresentar valores de desvio superiores. Considerando a janela temporal ideal para a identificação de condutores, conclui-se que a janela de 90 segundos maximiza o desempenho do classificador KNN, porém nos casos das técnicas MLP e RF, quanto maior a janela temporal, melhor a taxa de acertos dos mesmos. A janela temporal de 30 segundos apresentou desempenho inferior no KNN e RF, enquanto que no caso do MLP, os valores se estabilizaram a partir da janela de 30 segundo, com acréscimo razoável na janela de 120 segundo.

Este desempenho superior do KNN pode ser explicado pelo uso das técnicas de redução de dimensionalidade (PCA, ICA) que buscam maximizar a separação entre classes e esta técnica se baseia em distâncias entre uma amostra desconhecida e o restante dos dados. Já o desempenho da RF foi satisfatório visto que somente foram utilizadas quatro variáveis de entrada e, no caso da MLP, deve-se realizar um estudo de maneiras de otimizar o desempenho desta, ou mesmo o uso de topologias alternativas, uma vez que seu desempenho inferior frente à estas técnicas não é convencional.

Pode-se observar que a janela de 90 segundos tem o melhor desempenho no que diz respeito aos acertos quanto à autenticidade do condutor. É valido ressaltar que, em aplicações reais futuras, deve-se determinar um valor de janela ótimo que alie a rapidez e acurácia na identificação do condutor, visto que uma janela de 90 segundos pode ser muito quanto se trata de uma ação criminosa, porém, uma classificação errônea é uma situação que deve ser evitada para a confiabilidade do sistema.

Com relação à eficiência dos métodos de redução de dimensionalidade, nenhumas das técnicas se mostrou superior em questão de desempenho em relação às outras, no qual o objetivo de diminuir o espaço amostral de entrada do classificador foi alcançado sem que se houvesse comprometimento na classificação. No caso do melhor classificador, o KNN, as três técnicas obtiveram acurácias semelhantes nas mesmas condições de janela temporal. No entanto, nos classificadores RF e MLP não houve uma técnica de redução com

desempenho superior em todas as circunstâncias.

Como o classificador KNN teve um desempenho consideravelmente superior aos demais, optou-se por efetuar uma análise de desempenho mais aprofundada do mesmo, considerando, além da acurácia, as métricas de análise de concordância *Cohen's kappa Score* e *F1-Score*. As Tabelas 3, 4 e 5 apresentam os valores médios destas métricas, onde o valor do desvio padrão foi insignificante e portanto foi desconsiderado. Pode-se observar que em todos os casos, a janela de 90 segundos resulta em um resultado melhor chegando à 99.5% de acertos com redução PCA e ICA e 99.4% na redução IPCA. A matriz de confusão de um dos experimentos está contida na Tabela 6, onde é possível que a taxa de falsos positivos e negativos foi baixa, o que mostra a confiabilidade do modelo, mesmo em situações de desbalanceamento de classes.

Tabela 3: Desempenho do Classificador KNN e Redução de Dimensionalidade PCA em diferentes métricas

PCA			
Janela (s)	KNN		
	Acurácia	F1-Score	Cohen Kappa
15	0.956	0.927	0.896
30	0.919	0.838	0.785
60	0.971	0.951	0.930
90	0.995	0.992	0.988
120	0.989	0.981	0.974

Tabela 4: Desempenho do Classificador KNN e Redução de Dimensionalidade IPCA em diferentes métricas

IPCA			
Janela (s)	KNN		
	Acurácia	F1-Score	Cohen Kappa
15	0.956	0.927	0.896
30	0.922	0.842	0.7905
60	0.972	0.952	0.932
90	0.994	0.990	0.986
120	0.989	0.981	0.974

Tabela 5: Desempenho do Classificador KNN e Redução de Dimensionalidade ICA em diferentes métricas

ICA			
Janela (s)	KNN		
	Acurácia	F1-Score	Cohen Kappa
15	0.957	0.927	0.896
30	0.917	0.834	0.778
60	0.976	0.960	0.943
90	0.995	0.991	0.988
120	0.988	0.980	0.971

Tabela 6: Matriz de Confusão para o classificador KNN, redução ICA e janela de 120 segundos.

Real	**Classificado**	
	Não Autorizado	Autorizado
Não Autorizado	14113	89
Autorizado	135	4183

6 Conclusões

O presente trabalho teve como objetivo apresentar um estudo sobre técnicas de extração de características, redução de dimensionalidade e classificação para o problema de identificação de condutores, por meio de dados de direção oriundos de sensores embarcados no veículo e colhidos por meio do barramento CAN do mesmo. O sistema utilizando o classificador KNN, com uma janela temporal de 90 segundos para extração de características estatísticas se mostrou eficiente para a execução da tarefa proposta, com acurácia superior à 99%, independentemente da técnica de redução de dimensionalidade empregada.

O desempenho superior do classificador KNN pode ser explicado pela maximização da separação de classes proporcionada pelo métodos de redução de dimensionalidade, uma vez que este classificador trabalha com medidas de distância. Porém, este classificador não é adequado para conjunto de dados extensos e de alta dimensionalidade, uma vez que em cada predição ele realiza um número elevado de cálculos. Já o modelo RF obteve um desempenho razoável, visto que o número de entradas foi limitada a quatro e esta técnica é adequada para dados de alta dimensionalidade. Portanto, espera-se avaliar esta técnica com diferentes dimensões de entrada e determinar o número otimizado de variáveis de entradas. No caso da MLP, será efetuada uma análise mais detalhada com o objetivo de maximizar seu desempenho, como variações do método de treinamento e da estrutura, como número de neurônios e cadas escondidas, bem como a dimensão das variáveis de entrada.

Além destas análises, para trabalhos futuros, espera-se avaliar o impacto do número de condutores autorizados no desempenho do sistema, além empregar dados de sensores de outras fontes, como unidades de medidas inerciais. Também espera-se o desenvolvimento de um *software* embarcado que possa desempenhar a identificação de condutores em situações reais, de modo a comprovar a viabilidade do sistema.

Agradecimentos

Os autores gostariam de agradecer ao Dr. Byung Il Kwak e o *Hacking and Countermeasure Research Lab* por fornecerem o *dataset* utilizado neste trabalho. Os autores agradecem também o apoio financeiro da CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior e da

UFLA - Universidade Federal de Lavras, para a publicação deste artigo.

Referências

Artac, M., Jogan, M. and Leonardis, A. (2002). Incremental pca for on-line visual learning and recognition, *Object recognition supported by user interaction for service robots*, Vol. 3, pp. 781–784 vol.3.

Balsubramani, A., Dasgupta, S. and Freund, Y. (2013). The fast convergence of incremental pca, *in* C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger (eds), *Advances in Neural Information Processing Systems 26*, Curran Associates, Inc., pp. 3174–3182.

Banerjee, M., Capozzoli, M., McSweeney, L. and Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures, *Canadian journal of statistics* **27**(1): 3–23.

Cao, L., Chua, K. S., Chong, W., Lee, H. and Gu, Q. (2003). A comparison of pca, kpca and ica for dimensionality reduction in support vector machine, *Neurocomputing* **55**(1): 321–336.

Chiang, L. H., Kotanchek, M. E. and Kordon, A. K. (2004). Fault diagnosis based on fisher discriminant analysis and support vector machines, *Computers & chemical engineering* **28**(8): 1389–1401.

Fórum Brasileiro de Segurança Pública (2017). 11º anuário brasileiro de segurança pública.

Guedes, J. D., Ferreira, D. D. and Barbosa, B. H. (2016). A non-intrusive approach to classify electrical appliances based on higher-order statistics and genetic algorithm: a smart grid perspective, *Electric Power Systems Research* **140**: 65–69.

Guyon, I. and Elisseeff, A. (2006). An introduction to feature extraction, *Feature extraction*, Springer, pp. 1–25.

Hyvarinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis, *IEEE transactions on Neural Networks* **10**(3): 626–634.

Koldovsky, Z., Tichavsky, P. and Oja, E. (2006). Efficient variant of algorithm fastica for independent component analysis attaining the cramér-rao lower bound, *IEEE Transactions on neural networks* **17**(5): 1265–1277.

Kwak, B. I., Woo, J. and Kim, H. K. (2016). Know your master: Driver profiling-based anti-theft method, *PST 2016*.

Lee, T.-W. (1998). Independent component analysis, *Independent Component Analysis*, Springer, pp. 27–66.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J. and Liu, H. (2016). Feature selection: A data perspective, *arXiv preprint arXiv:1601.07996* .

Mendel, J. M. (1991). Tutorial on higher-order statistics (spectra) in signal processing and system theory: Theoretical results and some applications, *Proceedings of the IEEE* **79**(3): 278–305.

Naves, R., Barbosa, B. H. and Ferreira, D. D. (2016). Classification of lung sounds using higher-order statistics: A divide-and-conquer approach, *Computer methods and programs in biomedicine* **129**: 12–20.

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.

Sugiyama, M. (2007). Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis, *Journal of machine learning research* **8**(May): 1027–1061.

Wang, J. and Chang, C.-I. (2006). Independent component analysis-based dimensionality reduction with applications in hyperspectral image analysis, *IEEE transactions on geoscience and remote sensing* **44**(6): 1586–1600.

Weng, J., Zhang, Y. and Hwang, W.-S. (2003). Candid covariance-free incremental principal component analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **25**(8): 1034–1040.

Wold, S., Esbensen, K. and Geladi, P. (1987). Principal component analysis, *Chemometrics and intelligent laboratory systems* **2**(1-3): 37–52.