Classificação de condutores utilizando informações do sistema OBD-II

Pedro H. A. Ribeiro[1], José M. P. M Júnior[1]

[1]Grupo de Análise de Sistemas Inteligentes – Universidade Federal do Piauí (UFPI)
Campus Universitário Ministro Petrônio Portella, s/n Bairro Ininga
CEP: 64049-550 – Teresina – PI – Brasil

pedro.eeufpi@gmail.com, josemenezesjr@ufpi.edu.br

Abstract. *The application of machine learning techniques in the identification and classification of drivers using the OBD-II (On Board Diagnostic) system has recently been successfully performed. This paper proposes the classification of drivers by using OBD-II acquired information and machine learning techniques. This information refers to 10 conductors and 51 attributes. The evaluation of the techniques is based on the calculation of the average accuracy, median, maximum and minimum values, standard deviation and execution time of each applied computational technique. The analysis of the presented results shows that the classification was satisfactory, since the techniques allowed to obtain average accuracy above 96%, with standard deviation close to zero. Thus, such results allow the inference of the viability of the use of the techniques applied in later works.*

Resumo. *A aplicação de técnicas de aprendizagem de máquinas para a identificação e classificação de motoristas utilizando o sistema OBD-II (On Board Diagnostic) tem sido realizada recentemente com sucesso. Este trabalho propõe a classificação de condutores pela utilização de informações adquiridas do OBD-II e de técnicas de aprendizagem de máquina. Estas informações se referem a 10 condutores e 51 atributos. A avaliação das técnicas se baseia no cálculo da acurácia média, mediana, valores máximos e mínimos, desvio padrão e tempo de execução de cada técnica computacional aplicada. A análise dos resultados apresentados evidencia que a classificação foi satisfatória, uma vez que as técnicas permitiram a obtenção de acurácias médias acima de 96%, com desvio padrão próximo de zero. Assim, tais resultados permitem a inferência da viabilidade do uso das técnicas aplicadas em posteriores trabalhos.*

1. Introdução

Aprendizagem de máquinas tem sido utilizada com sucesso em tarefas de identificação e classificação de condutores de veículos utilizando o sistema OBD-II (*On Board Diagnostic*). Informações como a posição do acelerador, a velocidade do veículo e o consumo de combustível por quilômetro, por exemplo, podem ser utilizados para determinar se o condutor está autorizado ou não para conduzir o veículo (Sistema antifurto), como realizado em Ramos 2016, ou para determinar as ações comumente executadas por motoristas em diferentes situações no trânsito, de acordo com Fernandez et al 2016, ou, ainda, como o motorista dirige, em relação à sua experiência e habilidade, como realizado em Vaiana et al 2014.

Uma vez que existem nos veículos atuais a rede CAN (*Controller Area Network*), que permite a comunicação entre dispositivos eletrônicos nos automóveis modernos como o sistema de bloqueio e o sistema de freios antitravamento, dados em tempo real são enviados para o barramento CAN e, por meio do protocolo de comunicação OBD-II, disponibilizados para extração das grandezas de condução desejadas (Martinelli et al, 2018). Pode-se, então, utilizar essas informações para classificar condutores através técnicas de aprendizagem de máquina.

Aplicação semelhante foi realizada por Kwak et al 2016, que fez a classificação de motoristas pelas técnicas de Árvore de Decisão, *Random Forest*, *Multilayer Perceptron* e *K-Nearest Neighbor* aplicando no pré-processamento a exclusão de atributos idênticos ou redundantes, feita pelo método *InfoGainAttributeEval*, em WEKA. Já em Souza et al 2018, a classificação de condutores é feita mediante a aplicação de técnicas de dados e aprendizado de máquina (*Ranfom Forest, Multilayer Perceptron* e *K-Nearest Neighbor*) através da extração de características estatísticas e redução de dimensionalidade por Análise Discriminante de Fisher (FDA), Análise de Componentes Principais (PCA), Análise de Componentes Principais Incrementais (IPCA) e Análise de Componentes Independentes (ICA).

Pode-se atrelar à fase de pré-processamento de dados a dificuldade do problema, pois alguns dos atributos do banco são fortemente correlacionados, possuindo, ainda, escalas distintas, oriundas de cada sensor constituinte do veículo. Assim, o pré-processamento dos dados será baseado na matriz de correlação dos atributos e na normalização dos mesmos, bem como na influência deles no desempenho das técnicas aplicadas.

No tocante à construção deste artigo, na seção 3 são aplicadas técnicas de aprendizagem de máquinas. Após, o pré-processamento dos dados de condução, apresentado na seção 4, buscando-se a classificação dos condutores do veículo, avaliada por grandezas estatísticas, como a acurácia média, a mediana e o desvio padrão, além de outras.

2. Barramento CAN e Protocolo OBD-II

A rede CAN define um padrão de comunicação genérico através de um barramento que disponibiliza diversas variáveis que caracterizam a condução de motoristas, como consumo de combustível e velocidade do motor. No entanto, tal barramento é composto por várias unidades de controle eletrônico (ECU) que se comunicam entre si. Logo, necessita-se de algo que faça comunicação entre as unidades de controle. É neste ponto que, sendo mais específico que o CAN e especificamente criado como padrão para veículos desde 1996, mostra-se a importância do protocolo OBD-II, que permite a comunicação entre as ECU (Birnbaum et al, 2001).

3. Técnicas para Classificação

3.1. Árvore de Decisão - DT

Uma árvore de decisão é um mapa dos possíveis resultados de uma série de escolhas possíveis. Assim, permite que um indivíduo ou organização compare possíveis ações baseado em seus custos, probabilidades e benefícios. As arvores de decisão podem ser utilizadas, com excelente desempenho, na tarefa de classificação (Rokach et al,2008). A

partir de um conjunto de regras quantitativas e qualitativas, pode-se calcular valores desejados.

3.2. Random Forest - RF

Trata-se de uma técnica de aprendizagem de máquina usada para classificação, regressão, e outras aplicações que se utiliza da combinação de várias árvores de decisão, para o aumento da precisão e estabilidade dos resultados, corrigindo o hábito de supertreinamento (*overfitting*) dos algoritmos de árvores de decisão comuns. Desta forma, as amostras de treinamento serão distribuídas de forma a manter a mesma esperança para cada árvore criada, à medida que reduz a variância.

A técnica *Random Forest* tem o objetivo de efetuar a criação de várias árvores de decisão usando um subconjunto de atributos a partir do conjunto original (conjunto de dados de treinamento), o que possibilita uma melhor análise dos dados. Após a criação dos conjuntos de árvores de decisão é possível efetuar a classificação de qual possui melhor ganho de conhecimento para a solução de determinado problema (Neto, 2014).

3.3. Multilayer Perceptron - MLP

Buscando a similaridade com o cérebro humano, as MLP são técnicas computacionais compostas por camadas de neurônios, interligados por sinapses de pesos (Silva et al, 2010). Uma representação generalista desta rede com uma única camada escondida, Figura 1, é baseada no modelo *perceptron* proposto por Rosenblatt, que englobava algoritmos de treinamento supervisionado para a atualização dos pesos entre as sinapses (*backpropagation*).

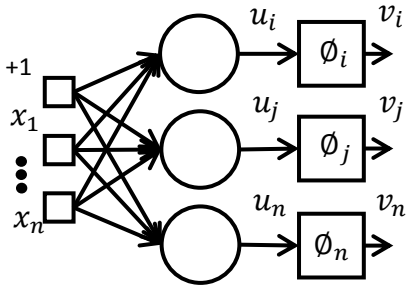

Figura 1. Representação generalista de uma rede neural MLP

A inserção de neurônios em sequência determina a quantidade de camadas ocultas, que podem ou não aumentar o desempenho do algoritmo e tem como finalidade a extração de características mais complexas do banco de dados. Porém, a maior complexidade do problema está relacionada com o processamento de informações na camada oculta, logo, com o tempo necessário para treinamento da rede (Bishop, 1995). O treinamento da rede é baseado na atualização dos pesos sinápticos entre os neurônios, que recebem a informação de entrada e transportam-na pela rede gerando uma saída estimada. Esta é comparada com o resultado desejado produzindo um a diferença, denominada custo (*feedback*) (Chollet, 2018).

4. Metodologia

Este trabalho utilizou a linguagem *Python* para implementação e aplicação das técnicas de aprendizagem de máquina no banco de dados disponibilizado por Kwak et al 2016 em conjunto com o *Hacking and Countermeasure Research Lab*, que consiste de informações de condução de 10 motoristas num percurso de 23 km na cidade de Seul, na Coréia do

Sul, obtidos a partir do barramento CAN do veículo Kia Soul. Esta aquisição foi permitida pelo uso da interface OBD-II. Cada motorista conduziu o veículo duas vezes, o que permitiu que as técnicas fossem aplicadas em dois bancos de dados diferentes, com 10 classes (motoristas de A a J) e 51 atributos, cada classe com a mesma quantidade de instâncias. Alguns dos atributos dos bancos de dados são apresentados na Tabela 1.

Tabela 1. Alguns atributos dos bancos de dados

Característica	Descrição
Valor do pedal de acelerador	Porcentagem do ângulo de abertura do pedal do acelerador
Velocidade do Veículo	Velocidade instantânea do veículo
Consumo de Combustível	Consumo instantâneo de combustível
Temperatura do líquido de arrefecimento do motor	Temperatura do líquido do motor a combustão
Velocidade de travamento do motor	Monitoramento da válvula de travamento

Analisando-se os bancos de dados, pode-se verificar a existência de atributos constantes para todos os condutores, desconsiderados, assim, para a análise dos algoritmos aplicados, e atributos com alta correlação, fato que pode afetar negativamente o desempenho das técnicas de classificação (Lee et al, 2005). Assim, aplicando-se na matriz de correlação 51x51 o operador matemático módulo e, a posteriori, somando-se todas as linhas coluna a coluna, construiu-se um vetor coluna. Neste, removeu-se do banco de dados os atributos cujos módulos do vetor coluna, representado pelo histograma da Figura 2, fossem maiores que 13. Assim, reduziu-se da quantidade de atributos para 25.

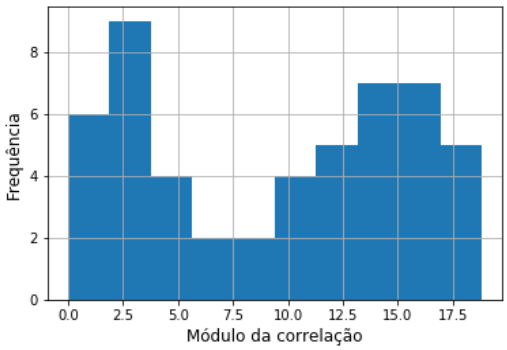

Figura 2. Histograma do módulo das correlações dos atributos do banco de dados

A exclusão de atributos mediante a correlação entre os mesmos pode ser explicada pelo fato de alguns atributos, como por exemplo, o fator de escala de torque (*Torque scaling factor*) e as velocidades das rodas (*Wheel velocity front left-hand, Wheel velocity rear right- Hand, Wheel velocity front right- Hand, Wheel velocity rear left-hand*) possuírem, respectivamente, valor constante para todas as classes e alta correlação (módulo em torno de 0,98).

5. Resultados

Como este trabalho tem o objetivo principal de classificar condutores utilizando técnicas de aprendizagem de máquina, obtiveram-se, variando-se os parâmetros de cada técnica (quantidade de árvores e de neurônios), os seguintes resultados.

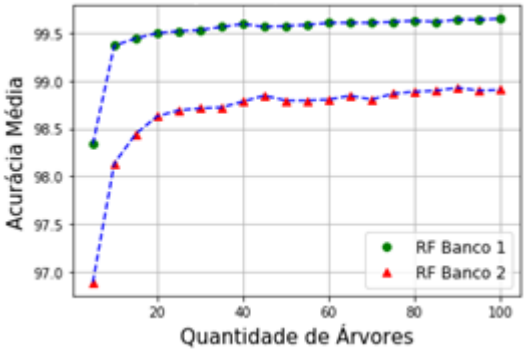

Figura 3. Acurácia do *Random Forest*

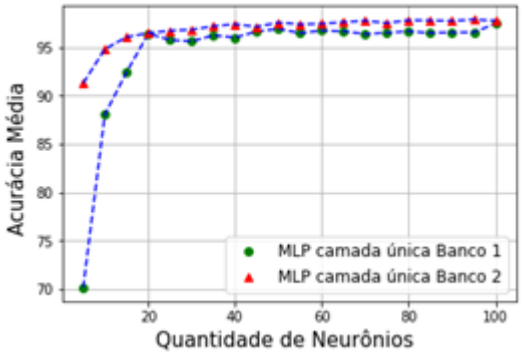

Figura 4. Acurácia da MLP camada única

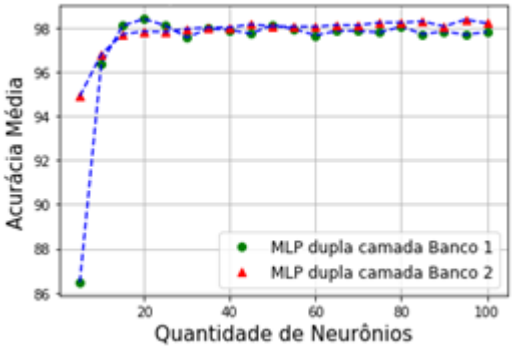

Figura 5. Acurácia da MLP dupla camada

Para as técnicas de decisão por árvore (Árvore de Decisão e *Random Forest*), Figura 3, percebe-se que o aumento da quantidade de árvores traz consigo o aumento da acurácia do algoritmo para os dois bancos de dados, sendo no segundo banco, um pouco maior que no primeiro, pelo fato de as árvores trabalharem com probabilidades e custos de acertos, que aumentam com a quantidade de árvores.

Comportamento semelhante acontece com as *MLP* de camada única da Figura 4, ou seja, o aumento da quantidade de neurônios eleva a acurácia do algoritmo até uma porcentagem máxima, ainda que para algumas quantidades de neurônios a acurácia decresça. Isto ocorre pela tendência de maior precisão na retropropagação do erro entre as camadas da rede. Em relação aos dois bancos de dados, verifica-se que a acurácia média no segundo banco de dados é maior. Já para as *MLP* camada dupla da Figura 5, tem-se um comportamento mais oscilante quando do aumento da quantidade de neurônios nas camadas ocultas, onde o número de neurônios da primeira camada oculta é o dobro da quantidade da segunda camada, e uma média de acurácia ligeiramente maior no banco 2.

Entretanto, nem sempre o aumento do número de camadas significa aumento da acurácia, pois ocorre também o aumento da complexidade do tratamento dos dados entre os neurônios, neste caso, prejudicial à acurácia da técnica.

A determinação das acurácias das técnicas de classificação usadas para cada banco de dados se baseou na análise estatística de cada algoritmo através do cálculo da média, mediana, valores máximos e mínimos, desvio padrão e tempo de classificação, uma vez que ao se usar técnicas heurísticas para a classificação, deve-se fornecer, para cada técnica usada, certa confiabilidade. Isto pode ser fornecido estatisticamente pelo cálculo da média, mediana, valores máximo e mínimo e desvio padrão de cada algoritmo, após repetições (30 para este trabalho) da mesma técnica para uma determinada configuração. Estes resultados podem ser analisados nas Tabelas 2 e 3.

Tabela 2. Análise estatística de resultados (Banco 1)

Técnica	Média (%)	Desv. Pad. (%)	Mediana (%)	Máx. (%)	Mín. (%)	Tempo (s)
Árv. Decisão	99,19	0,014	99,19	99,22	99,17	3,31
RF (100)	99,68	0,016	99,69	99,71	99,66	42,82
MLP (60)	96,80	0,32	96,67	97,08	95,90	1738,04
MLP (40-20)	98,57	0,23	98,62	98,82	97,99	2551,91

Tabela 3. Análise estatística de resultados (Banco 2)

Técnica	Média (%)	Desv. Pad. (%)	Mediana (%)	Máx. (%)	Mín. (%)	Tempo (s)
Árv. Decisão	98,24	0,05	98,17	98,30	98,09	4,22
RF (90)	98,98	0,02	98,96	99,02	98,94	47,61
MLP (95)	97,91	0,016	97,84	97,88	97,75	1886,83
MLP (194-97)	98,45	0,013	98,46	98,59	98,16	2267,32

A análise dos resultados apresentados permite a inferência de que apesar da acurácia média de uma técnica computacional ser alta nem sempre ela representa a mais específica, fato evidenciado pelo desvio padrão. Pode-se observar nas Tabelas 2 e 3 que os desvios para as técnicas utilizadas foram diferentes, mesmo para acurácias médias com valores próximos (variações em até 5% não são tão significantes em termos práticos). A partir disso, percebe-se a importância do cálculo não somente das acurácias médias, mas de outras grandezas estatísticas (desvio padrão, acurácias máxima, mínima e mediana).

O tempo computacional de cada algoritmo, por sua vez, é de grande importância, ao se pensar na exequibilidade deste trabalho em ramos industriais, afinal, quanto mais rápido e eficiente é a técnica computacional utilizada, mais atrativa será para as aplicações reais. Percebe-se nas Tabelas 2 e 3 que os algoritmos baseados na cognição (redes MLP) apresentaram os maiores tempos para a classificação, fato já esperado uma vez que as técnicas baseadas em decisão (Árvore de Decisão e *Random Forest*) não necessitam de cálculos extensos, como a retropropagação do erro nas redes MLP.

Em comparação com Kwak et al 2016, pela Tabela 4, as acurácias obtidas neste trabalho foram semelhantes, mesmo com metodologias diferentes entre os trabalhos.

Tabela 4. Acurácia de Kwak et al 2016.

Técnica	Acurácia (%)	Banco de Atributos
Árv. Decisão	98,4	*Our research's feature set + statistical feature*
RF	99,6	*Our research's feature set + statistical feature*
MLP	96,4	*Our research's feature set + statistical feature*

Para a obtenção dos resultados apresentados, utilizou-se da biblioteca *Scikit-learn* disponível na linguagem *Python* e das configurações a seguir, para cada técnica usada.

Configuração de técnicas para o Banco 1

- Árvore de Decisão: critério – *entropy;*
- *Random Forest (RF)*: 100 estimadores, critério – *entropy;*
- *Multilayer Perceptron (MLP)* camada escondida única: 60 neurônios, função de ativação *ReLU*, taxa adaptativa iniciando de 0,01 e momentum de 0,9;
- *Multilayer Perceptron (MLP)* dupla camada escondida: 40 por 20 neurônios, função de ativação *ReLU*, taxa adaptativa iniciando de 0,01 e momentum de 0,9.

Configuração de técnicas para o Banco 2

- Árvore de Decisão: critério – *entropy;*
- *Random Forest (RF)*: 90 estimadores, critério – *entropy;*
- *Multilayer Perceptron (MLP)* camada escondida única: 95 neurônios, função de ativação *ReLU*, taxa iniciando de 0,01 e momentum de 0,9;
- *Multilayer Perceptron (MLP)* dupla camada escondida: 194 por 97 neurônios, função de ativação *ReLU*, taxa iniciando de 0,01 e momentum de 0,9.

Percebe-se que alguns parâmetros das configurações dos algoritmos usados diferiram entre os bancos de dados. Pode-se explicar isso pelo fato de os dois bancos, apesar de serem dados referentes aos mesmos condutores (A a J), serem diferentes e representarem mais fielmente e em mais instâncias a condução dos mesmos motoristas num mesmo percurso, fato que ocorre corriqueiramente e que não significa que um motorista não identificado/autorizado esteja conduzindo o veículo.

6. Conclusões

O presente trabalho apresentou como objetivo avaliação da classificação de condutores pelo uso de técnicas de aprendizagem de máquina. A partir dos resultados apresentados, os algoritmos com acurácia maior e tempo de execução menor foram identificados, conforme apresentados nas Tabelas 2 e 3. Logo, percebeu-se que a acurácia nem sempre deve ser o único fator a ser analisado para uma técnica, desta forma, destacando o impacto da mediana, dos valores máximos e mínimos, do desvio padrão e do tempo de execução para a classificação de motoristas através do uso de técnicas computacionais, uma vez que existe a possibilidade de uma técnica possuir alta acurácia média, mas baixa especificidade (alto desvio padrão e maior tempo para realização da tarefa desejada).

Uma vez que a análise dos resultados apresentados evidenciou que as técnicas utilizadas apresentaram resultados satisfatórios para a classificação de motoristas (acima de 96% de média de acertos), tais técnicas podem ser aplicadas em um banco de dados

próprio do autor deste trabalho, a fim de verificar a adequação das mesmas aos dados. No entanto, tais resultados não descartam a possibilidade da aplicação de outras técnicas de aprendizagem de máquinas para classificação, buscando-se melhorar os resultados obtidos neste artigo.

7. Referências

Birnbaum, R., Truglia, J. (2001). Getting to know OBD II. Editora Ralph Birnbaum.

Bishop, C. (1995). Neural networks for pattern recognition. Oxford University Press.

Chollet, F. (2018). Deep learning with Python. Manning Publications. United States of America.

Fernandez, S., Ito, T. (2016) Driver classification for intelligent transportation systems using fuzzy logic. In: IEEE. Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on. [S.l.]. p. 1212–1216.

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

Lee, H. D., Monard,M. C., Voltolini, R.F., Wu, F. C. (2005). Avaliação Experimental e Comparação de Algoritmos de Seleção de Atributos Importantes com o Algoritmo FDimBF Baseado na Dimensão Fractal. Relatórios Técnicos do ICMC. N° 264.

Martinelli, F., Mercaldo, F., Orlando, A., Nardone, V., Santone, A., Sangaiah, A. K (2018). Human Behavior Characterization Driving Style Recognition in Vehicle System. Disponível em: https://doi.org/10.1016/j.compeleceng.2017.12.050.

Neto, C. D. G. (2014) Potencial de técnicas de mineração de dados para o mapeamento de áreas cafeeiras. INPE.

Ramos, C. A. (2016) Sistema Neural Antifurto Veicular. Programa de Pós Graduação em Ciência da Computação. Universidade Federal de Lavras, Minas Gerais.

Rokach, L., Maimon, O. (2008) Data mining with decision trees: Theory and applications. World Scientific Publishing. 2ª Ed. Series in Machine Perceptron Artificial Intelligence, Vol 81.

Silva, I. N., Spatti, D. H., Flauzino, R. A. (2010) Redes neurais artificiais para engenharia e ciências aplicadas. Artliber Editora Ltda, São Paulo, SP, Brasil.

Souza, A. G., Lacerda, W. S., Ferreira, D. D., Campos, G. L. (2018). Sistema de Identificação de Condutores Baseado em Métodos de Extração de Características Estatísticas e Técnicas de Redução de Dimensionalidade. XXII Congresso Brasileiro de Automática.

Vaiana, R. et al. (2014) Driving behavior and traffic safety: an acceleration-based safety evaluation procedure for smartphones. Modern Applied Science, v. 8, n. 1, p. 88, 2014.