Classificação do Modo de Transporte e Propósito de Viagem Baseada em Dados Socioeconômicos

Elton F. de S. Soares[1], **Carlos Alberto V. Campos**[1]

[1]Universidade Federal do Estado do Rio de Janeiro (UNIRIO)
Av. Pasteur, 458 – Rio de Janeiro – RJ – Brasil

elton.soares,beto@uniriotec.br

Abstract. *Socioeconomic characteristics, such as income and age, can influence the mobility patterns observed in the residents of a city, which in turn can be restricted by the options and quality of the modes of transportation available. Given this, one of the main tasks during the characterization of mobility data is the identification of the travel modes most frequently used during the movement of citizens, as well as the purposes of their trips. This work proposes a technique for identifying the modes of transportation used in a trip as well as its purpose, based only on socioeconomic variables. The proposed method is evaluated through cross-validation experiments using a public dataset from New York City. By comparing the confusion matrices obtained with different supervised learning algorithms, it is possible to conclude that the proposed method presents a performance similar to the other techniques in the classification of the purpose of the trip, with an overall accuracy of 67%, and higher for the classification of the travel mode, with an overall accuracy of 82%.*

Resumo. *Características socioeconômicas, como renda e idade, podem influenciar os padrões de mobilidade observados nos residentes de uma cidade que por sua vez podem ser restringidos pelas opções e qualidade dos meios de transporte disponibilizados. Dado isso, uma das principais tarefas durante a caracterização de dados de mobilidade é a identificação dos meios de transporte utilizados mais frequentemente durante o deslocamento dos cidadãos, assim como dos propósitos de suas viagens. Este trabalho, propõe uma técnica para identificação dos meios de transporte utilizados em uma viagem assim como o seu propósito, baseando-se somente em variáveis socioeconômicas. O método proposto é avaliado através de experimentos de validação cruzada utilizando uma base de dados pública da cidade de Nova Iorque. Através da comparação das matrizes de confusão da obtidas com diferentes algoritmos de aprendizado supervisionado é possível concluir que o método proposto apresenta um desempenho similar as demais técnicas na classificação do propósito da viagem, com acurácia de 67%, e superior para a classificação do modo de transporte, com acurácia de 82%.*

1. Introdução

O crescimento populacional nos centros urbanos é um fenômeno que gera desafios e oportunidades em diversas áreas, como saúde, meio ambiente e mobilidade. A eficiência dos

sistemas de transportes destes centros é vital para a mobilidade de seus residentes, além de ajudar na redução do impacto ambiental e melhoria da qualidade de vida nas grandes cidades [Calabrese et al. 2014].

Um dos pré-requisitos para o estabelecimento e manutenção de sistemas de transporte inteligentes é a caracterização da mobilidade urbana, que por sua vez necessita de ferramentas e técnicas para coleta, pré-processamento e interpretação dos dados de mobilidade de seus residentes [Zhang et al. 2011].

Características socioeconômicas, como renda e idade, podem influenciar os padrões de mobilidade observados nos residentes de uma cidade que por sua vez podem ser restringidos pelas opções e qualidade dos meios de transporte disponibilizados [Wang et al. 2017a]. Dentro desse contexto, uma das principais tarefas durante a caracterização de dados de mobilidade é a identificação dos meios de transporte utilizados mais frequentemente durante o deslocamento dos cidadãos, assim como dos propósitos de suas viagens.

Trabalhos recentes demonstraram que a utilização de variáveis socioeconômicas aumenta a acurácia dos modelos de classificação de modo de transporte baseados em dados de GPS [Wang et al. 2017a, Ye et al. 2018]. No entanto, ainda não foi proposta, muito menos avaliada, uma técnica para identificação dos meios de transporte utilizados em uma viagem, assim como o destino da viagem, baseando-se somente em variáveis socioeconômicas.

Dado isso, neste trabalho apresentamos as seguintes contribuições:

- A proposta de um método de classificação do modo de transporte e propósito de viagem baseado em variáveis socioeconômicas e utilizando a técnica de aprendizado de máquina supervisionado *Random Forest* [Breiman 2001].
- Avaliação do método proposto através de experimentos de validação cruzada utilizando uma base de dados pública da cidade de Nova Iorque e métricas obtidas a partir da matriz de confusão, como acurácia, precisão, cobertura e medida-F1 [Fawcett 2006].

O restante do artigo está estruturado da seguinte forma. A Seção 2 discute os principais trabalhos relacionados e a Seção 3 descreve o método proposto. Em seguida, a Seção 4 apresenta a avaliação experimental do método proposto e comparação de algoritmos de aprendizado de máquina. Por fim, a Seção 5 apresenta a conclusão e trabalhos futuros.

2. Trabalhos Relacionados

Nesta seção, apresentamos os trabalhos mais relevantes sobre o problema de identificação do modo de transporte e propósito da viagem.

Em um dos primeiros trabalhos a apresentar uma solução para este problema [Stopher et al. 2005], pesquisadores da Universidade de Sidney propuseram o uso de heurísticas com base em atributos extraídos dos registros de GPS, como velocidade média e velocidade máxima, combinados com informações de sistemas de informação geográfica (SIG), como trajetos de ruas e paradas de transporte público, para identificar os modos de transporte e os propósitos das viagens feitas por cidadãos de Sidney,

Austrália. Em um estudo posterior [Bohte and Maat 2009], realizado na Universidade de Tecnologia de Delft, os autores propuseram o uso de coordenadas GPS das extremidades da viagem combinadas com informações de ponto de interesse derivadas dos dados SIG para identificar o propósito da viagem e uma mistura de atributos derivados da velocidade coletada pelo GPS e dados de SIG para identificar os modos de transporte.

Em um estudo publicado em 2014 [Montini et al. 2014b], por pesquisadores do Instituto Federal de Pesquisa da Súiça, os autores propuseram uma técnica de predição de propósito de viagem utilizando um modelo de *Random Forest* treinado com dados de GPS e acelerômetro coletados por 156 participantes, participando de uma pesquisa de mobilidade de uma semana na Suíça, concluída em 2012 [1]. O trabalho seguinte [Montini et al. 2014a], destes mesmos autores, explorou o efeito do treinamento personalizado na precisão da predição do propósito da viagem e seu trabalho mais recente [Montini et al. 2015] comparou o desempenho da coleta de dados de mobilidade através de dispositivos GPS dedicados e smartphones. Em seus estudos, os autores utilizaram e aprimoraram um aplicativo de diário de mobilidade desenvolvido em um trabalho anterior [Schrammel et al. 2013] que executava a detecção do modo de transporte e a predição do propósito da viagem, embora cada tarefa fosse executada com base em um conjunto diferente de atributos.

Em 2015, um estudo realizado por pesquisadores da Universidade de Twente, Universidade da Cidade de Londres e uma empresa chamada Mobidot [Geurs et al. 2015] propôs uma solução que utilizava atributos extraídos por GPS, acelerômetro, Wi-Fi e rede celular combinados com atributos derivados de SIG para treinar um modelo Bayesiano para a detecção do modo de transporte. A identificação do propósito da viagem foi realizada com base em padrões históricos de mobilidade, obtidos através das rotas e locais mais frequentes para onde cada pessoa se deslocou.

Já em 2016, em um de nossos primeiros estudos nesta área [Quintella et al. 2016], implementamos o aplicativo de diário de viagem CityTracks na plataforma iOS com o objetivo de coletar dados do modo de transporte dos usuários de smartphones por meio de sensoriamento participativo e oportunista. Os dados coletados foram utilizados para desenvolver um algoritmo completo de detecção do modo de transporte que executa o pré-processamento e segmentação de dados, extração e sumarização de atributos e classificação do modo de transporte por meio de um modelo hierárquico de aprendizado de máquina. Em uma continuação deste trabalho [Soares et al. 2017], propusemos uma técnica para detecção do modo de transporte em tempo real que fora implementada em um aplicativo chamado CityTracks-RT[2] na plataforma Android e testada em campo com 19 voluntários na região metropolitana da cidade do Rio de Janeiro, Brasil. Com uma metodologia semelhante a nossa, pesquisadores da Universidade de Bologna coletaram dados de GPS, acelerômetro, giroscópio e do modo de transporte usado pelos usuários de smartphones para criar uma técnica de detecção do modo de transporte em tempo real utilizando *Cascading* de classificadores [Bedogni et al. 2016]. A técnica proposta também fora implementada em um protótipo na plataforma Android.

Em um estudo publicado durante o mesmo período, pesquisadores da Universidade de Tecnologia de Tampere apresentaram uma comparação de técnicas de extração de

[1] http://www.project-peacox.eu/

[2] https://github.com/eltonfss/CityTracks-RT

atributos e abordagens de aprendizado de máquina, incluíndo *Autoencoders* para detecção do modo de transporte usando GPS [Mäenpää et al. 2017]. O uso de *Autoencoders* também foi explorado em um outro estudo de diferentes autores [Wang et al. 2017b], em que fora avaliado uma técnica que usa *Sparse Autoencoders* para extrair atributos profundos no nível de ponto (APNPs) de atributos "artesanais" em nível de ponto e uma rede neural convolucional para agregar os APNPs e gerar atributos profundos no nível da trajetória (APNT). Como resultado deste estudo, os autores propuseram uma arquitetura de rede neural profunda para detectar o modo de transporte usando os atributos "artesanais" no nível da trajetória e os APNTs. Neste estudo, os atributos considerados "artesanais" são atributos extraídos através de técnicas de pré-processamento tradicionais, ou seja, atributos que não são extraídos através de redes neurais profundas utilizando aprendizado representacional.

Também em 2017, um estudo da Universidade de Minnesota [Ermagun et al. 2017] propôs o uso de serviços de pesquisa e descoberta on-line baseados em localização, como a API do Google Places, para melhorar a predição do propósito das viagens. Nesse trabalho foram aplicados os modelos de *Nested Logit* e *Random Forest* para identificar cinco propósitos de viagens coletadas durante o *2010 Travel Behavior Inventory em Minneapolis-St.* Outros trabalhos recentes deste tópico exploraram a inferência semi-automatizada de propósitos de viagem por meio de sensoriamento participativo [Seo et al. 2017], combinação de trajetórias de GPS, pontos de interesses e dados de mídia social [Meng et al. 2017], seleção de dados de diferentes estações do ano para treinamento e teste de modelos [Gong et al. 2017], detecção do propósito de viagem com redes neurais artificiais e com otimização de enxame de partículas [Xiao et al. 2016].

Em nosso trabalho mais recente [Soares et al. 2019], propusemos uma técnica de detecção do modo de transporte e propósito da viagem baseada em dados de GPS, utilizando uma mesma técnica de pré-processamento e dois modelos *Random Forest* para classificação. A técnica de *Random Forest* apresentou uma performance superior na avaliação experimental deste estudo, além de ter obtido bons resultados em trabalhos relacionados, incluindo um estudo de identificação do propósito da viagem baseado em GPS realizado por pesquisadores da Universidade de Shangai [Wang et al. 2017a]. Este fora o primeiro estudo a indroduzir a utilização de dados socioeconômicos para auxiliar na identificação do propósito da viagem a partir de dados de localização.

Conforme discutimos ao longo desta seção, diversas técnicas foram propostas nos últimos anos para permitir a identificação do modo de transporte e do propósito das viagens de habitantes de centros urbanos através de dados de sensores. No entanto, poucos trabalhos consideraram a utilização de dados socioeconômicos para auxiliar na tarefa de identificação e nenhum deles considerou a possibilidade de realização desta tarefa utilizando somente estes dados. Com base nisso, na próxima seção apresentaremos um método para predição dos modos de transporte e propósitos de viagens preferenciais dos habitantes de centros urbanos utilizando somente dados socioeconômicos dos mesmos.

3. Método Proposto

A método proposto no presente trabalho, denominado Método de Classificação do Modo de Transporte e Propósito da Viagem (MCMTPV), consiste no treinamento de dois mo-

delos de classificação a partir de atributos socioeconômicos de habitantes de centros urbanos utilizando a técnica de aprendizado de máquina supervisionado *Random Forest* [Breiman 2001]. A utilização desta técnica apresenta três vantagens principais para utilização no método proposto:

1. Ao selecionar variáveis e amostras de forma aleatória, para gerar diferentes árvores de decisão, o modelo final se torna mais robusto e resistente a ruídos nos dados.
2. É adaptável à diferentes bases de dados e consegue lidar com problemas de alta dimensionalidade, onde o número de atributos considerados na classificação é muito grande.
3. Consegue processar dados discretos ou contínuos sem a necessidade de normalização.

O treinamento dos modelos de classificação se dá em três etapas. Primeiramente, selecionam-se N subconjuntos de amostras do dataset original D para formar o dataset de treinamento $D_{treinamento}$ utilizando um método de amostragem em *bootstrap*:

$$D_{treinamento} = \{D_1, D_2, ..., D_N\} \tag{1}$$

Em uma segunda etapa, para cada subconjunto de amostras D_i em $D_{treinamento}$ é gerada uma árvore de decisão F_i utilizando um subconjunto aleatório dos atributos θ_i, garantindo que todos os $\{\theta_i, i = 1, 2...N\}$ sejam independentes e identicamente distribuídos. Dado isso, o conjunto de todas as árvores de decisão geradas é denominado F:

$$F = \{F_1, F_2, ..., F_N\} \tag{2}$$

Cada árvore de decisão em F é treinada utilizando o algoritmo de treinamento t, sem a aplicação de poda:

$$F_i = t(D_i, \theta_i), i = 1, 2, ..., N \tag{3}$$

Por fim, é gerado um modelo de *ensemble* E baseado no conjunto de árvores F, onde as classificações de novas amostras são realizadas através de uma votação. Dada uma nova amostra d, cada árvore F_i pertencente a F emite um voto através da função de classificação f_i e o resultado final da classificação do modelo E para amostra d, denominado, $f(d)$ é obtido pela seguinte equação:

$$f(d) = \max_{Y} \sum_{i=1}^{N} I(f_i(d) = Y) \tag{4}$$

Onde f_i representa a classificação gerada pela árvore F_i, Y representa uma classe de modo de transporte ou propósito de viagem e I representa a função característica de probabilidade [Van der Vaart 2000]. A Figura 1 ilustra o processo de treinamento dos modelos e a classificação de novas amostras proposto.

No método proposto, a classificação é feita utilizando dois modelos distintos. Um modelo é treinado para classificação do modo de transporte e o outro é treinado para

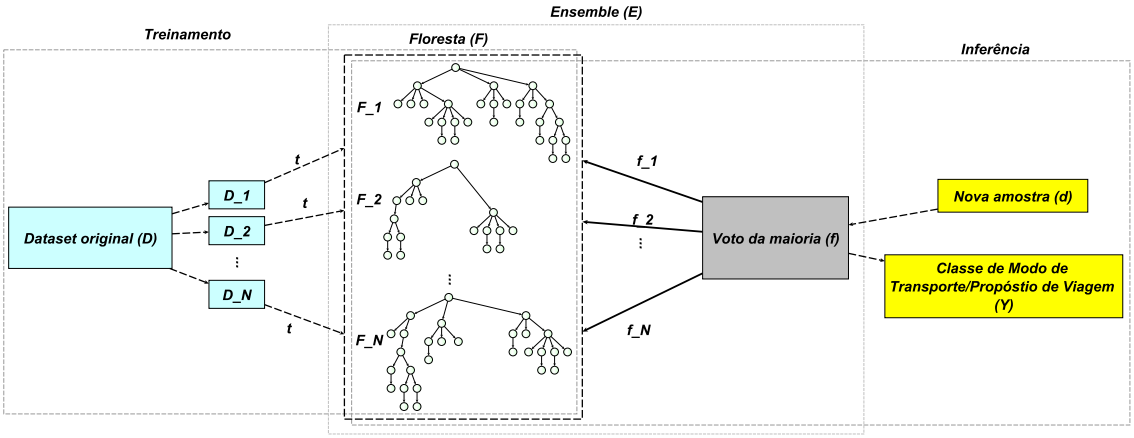


Figura 1. Treinamento de modelos e classificação do modo de transporte ou propósito de viagem de novas amostras.

classificação do propósito da viagem. Ambos são treinados e utilizados de acordo com as formalizações apresentadas e o processo ilustrado na Figura 1.

Apresentaremos na próxima seção uma avaliação experimental do método MCMTPV utilizando um dataset público com informações socioeconômicas dos usuários de um centro urbano.

4. Avaliação Experimental

A avaliação do método proposto será realizada através de experimentos de validação cruzada utilizando a base de dados New York City 2017 Citywide Mobility Survey [NYC Department of Transportation 2017]. Dentro desse contexto, faremos uma comparação da técnica proposta com outras técnicas de aprendizado de máquina do estado da arte, como Regressão Logística [Sperandei 2014], Árvore de Decisão [Song and Ying 2015] e *Support Vector Machine* [Ma and Guo 2014]. A análise e comparação de performance dos algoritmos considerarão a matriz de confusão do *fold* de maior acurácia, assim como, a precisão, cobertura e medida-F1 [Fawcett 2006] média de cada algoritmo por classe.

Dado isso, o objetivo desta avaliação é responder as seguintes perguntas:

Q1 - É possível identificar os modos de transporte utilizados por moradores de grandes centros urbanos a partir de modelos de classificação gerados com o método proposto?

Q2 - É possível identificar os propósitos das viagens de moradores de grandes centros urbanos a partir de modelos de classificação gerados com o método proposto?

Q3 - O algoritmo de aprendizado de máquina utilizado no método proposto permite a geração de modelos com mais acurácia que outras alternativas do estado da arte?

4.1. Base de dados: New York City 2017 Citywide Mobility Survey

Os dados desta base foram obtidos através de uma pesquisa mista conduzida pelo Departamento de Transportes da cidade de Nova Iorque em conjunto com a PSB, uma empresa de

pesquisa de marketing independente, durante sete semanas, compreendidas pelo período de 13 de Maio à 1 de Julho de 2017.

Um total de 3.603 residentes, maiores de 18 anos, da cidade de Nova Iorque participaram da pesquisa. Metade dos participantes foram recrutados pelo telefone utilizando a técnica de *random digit dialing* (RDD) baseada no código de área e uma lista telefônica. A outra metade foi recrutada pela internet através de listas obtidas pela PSB e identificadas pelo código postal.

Durante o período da pesquisa, foram registrados quatro dias em que as temperaturas ficaram acima de 90 graus *Farenheit* e quatro dias em que o nível de chuva excedeu uma polegada. Feriados compreendidos nesse período foram: dia das mães, dia dos pais, *Memorial Day*, *Puerto Rican Day Parade*, *Mermaid Parade*, *Eid al-Fitr* e *Pride Parade*.

A pesquisa foi dividida em duas partes: a pesquisa principal e os diários de viagem. A pesquisa principal avaliou comportamentos, atitudes e percepções dos meios de transporte em toda a cidade de Nova Iorque. Já os diários de viagem foram definidos como jornadas unidirecionais que começavam em uma localização de origem e terminavam em uma localização de destino.

Para avaliação do método proposto neste trabalho, são utilizados somente os dados obtidos através dos diários de viagem. O número das amostras contidas nesses dados é de 6.986 viagens, sendo 3.252 capturadas pelo telefone e 3.734 pela Internet. A margem de erro dos dados coletados é de $\pm 2,3\%$.

Para as amostras coletadas pelo telefone foram aplicados pesos baseados na *American Community Survey* de 5 anos sobre os seguintes fatores: idade, gênero, etnia, nível educacional e geografia. Já a amostragem pela Internet é um *oversample* das populações de bairros onde há dificuldade de obtenção de dados via telefone e está de acordo com a demografia de cada um desses bairros, não sendo representativa para toda a cidade de Nova Iorque. Os dados obtidos a partir dos diários de viagem são agregados por viagem e os percentuais indicam o percentual de viagens que apresentam cada característica.

Na Figura 2, apresentamos um mapa de correlação das variáveis selecionadas para nossos experimentos. Em um total de nove variáveis, incluindo o modo de transporte e o propósito da viagem, apenas renda e propósito da viagem apresentaram uma correlação acima de 0.3, tendo o modo de transporte apresentado uma correlação de 0.19 com a duração da viagem.

Na Figura 3, apresentamos os histogramas do número de amostras por modo de transporte e propósito de viagem, respectivamente. As classes menos frequentes foram agrupadas na classe "Outros" e é possível observar por estes gráficos o elevado desbalanceamento das amostras. Logo, fica evidente a necessidade de se utilizar técnicas de rebalançeamento durante nossos experimentos.

4.2. Experimentos

Os experimentos consistem em validações cruzadas de 10 *folds* onde para cada *fold* é feito o rebalanceamento do número de amostras por classe do dataset de treinamento utilizando a técnica de *Synthetic Minority Oversampling* (SMOTE) [Chawla et al. 2002]. Após o rebalanceamento é executado o algoritmo de aprendizado e o modelo gerado é avaliado no dataset de teste.

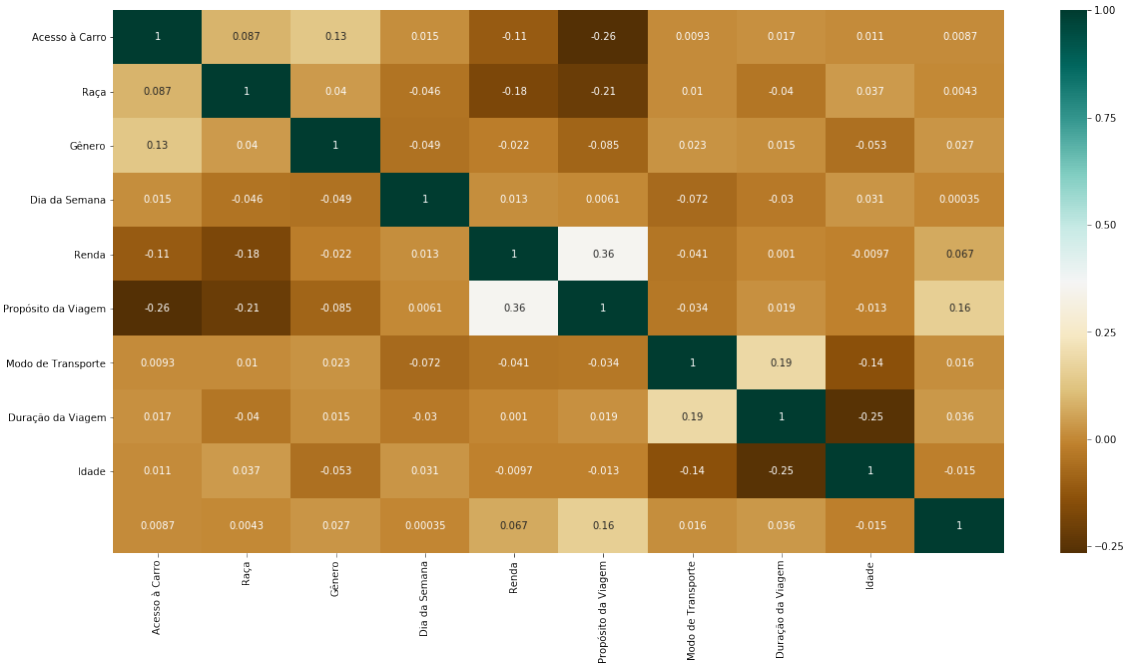


Figura 2. Mapa de correlação entre as variáveis selecionadas para os experimentos.

4.3. Métricas utilizadas

Para medir a performance da técnica proposta em comparação as outras técnicas, escolhemos usar as métricas: acurácia, precisão, cobertura e medida-F1, que serão definidas a seguir.

A acurácia é obtida pela razão do número de amostras classificadas corretamente como pertencentes ($Positivos\ Verdadeiros$) ou não pertencentes ($Negativos\ Verdadeiros$) a uma classe pelo total de classificações realizadas ($Total$) [Fawcett 2006]. Logo, $Acurácia$ é definida como $\frac{Positivos\ Verdadeiros + Negativos\ Verdadeiros}{Total}$.

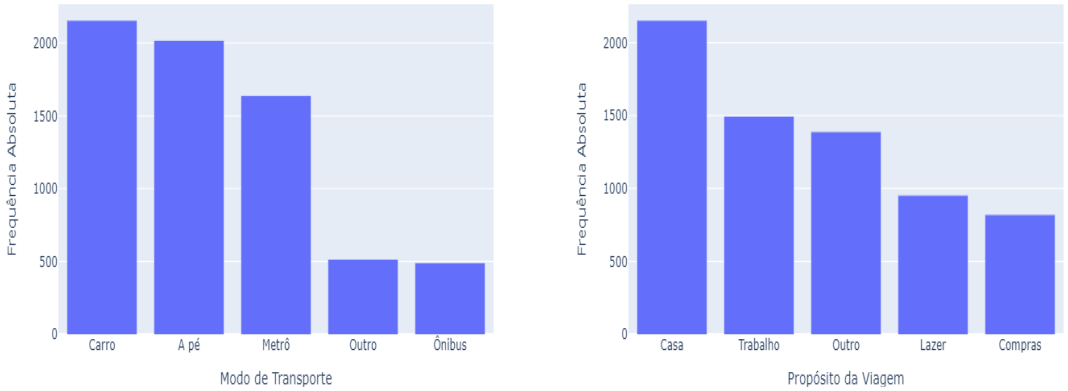
Já a métrica de precisão é obtida pela razão do número de amostras classificadas corretamente como pertencentes a uma classe pelo total de classificações realizadas para a mesma classe ($Positivos\ Verdadeiros$ + $Positivos\ Falsos$) [Fawcett 2006]. Logo, $Precisão$ é definida como $\frac{Positivos\ Verdadeiros}{Positivos\ Verdadeiros + Positivos\ Falsos}$.

A cobertura é obtida pela razão do número de amostras classificadas corretamente como pertencentes a uma classe pelo total de amostras pertencentes a mesma classe ($Positivos\ Verdadeiros$ + $Negativos\ Falsos$) [Fawcett 2006]. $Cobertura$ é definida como $\frac{Positivos\ Verdadeiros}{Positivos\ Verdadeiros + Negativos\ Falsos}$.

Por fim, a medida-F1 (F_1), é a média harmônica da precisão e cobertura de cada classe de modo de transporte [Fawcett 2006]. É definida como sendo igual a $2 \times \frac{Precisão \times Cobertura}{Precisão + Cobertura}$.

4.4. Resultados

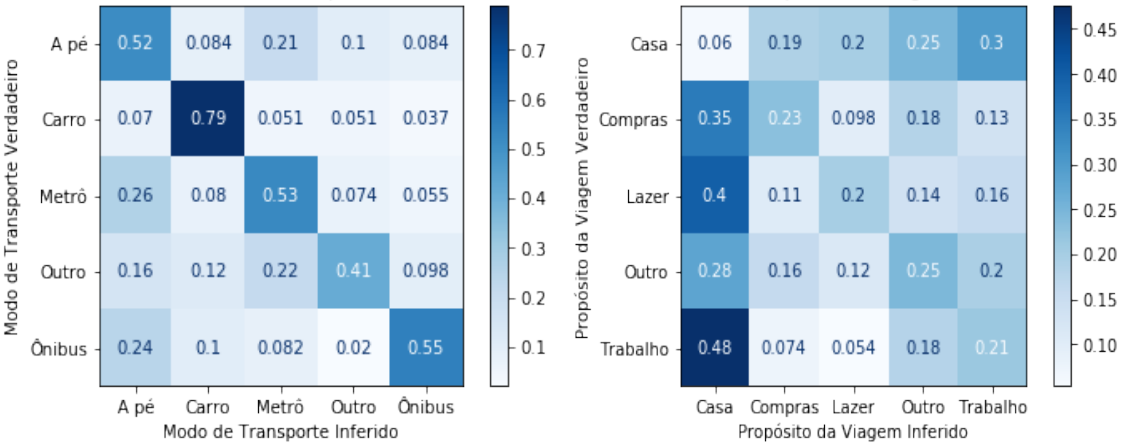
Na Figura 4 apresentamos as matrizes de confusão normalizadas das melhores *folds* obtidas com a técnica *Random Forest*. Na classificação do modo de transporte é possível



(a) (b)

Figura 3. Histogramas de número de amostras por modo de transporte (a) e propósito da viagem (b).

observar que o maior percentual de acertos foi obtido para a classe "Carro" e que em todas as outras classes o percentual foi superior a 50% com exceção da classe "Outros". As classes em que o classificador mais se confundiu foram "A pé" x "Metrô", tendo havido um percentual de amostras classificadas incorretamente superior a 20% para ambas as classes. No que diz respeito ao propósito da viagem, o percentual de acertos foi superior a 20% para a maioria das classes, sendo melhor que uma escolha aleatória, mas bastante abaixo do nível de acurácia desejado. O grande número de erros se deveu principalmente a um alto número de falsos positivos para as inferências da classe "Casa", com um percentual superior a 25% para todas as demais classes.

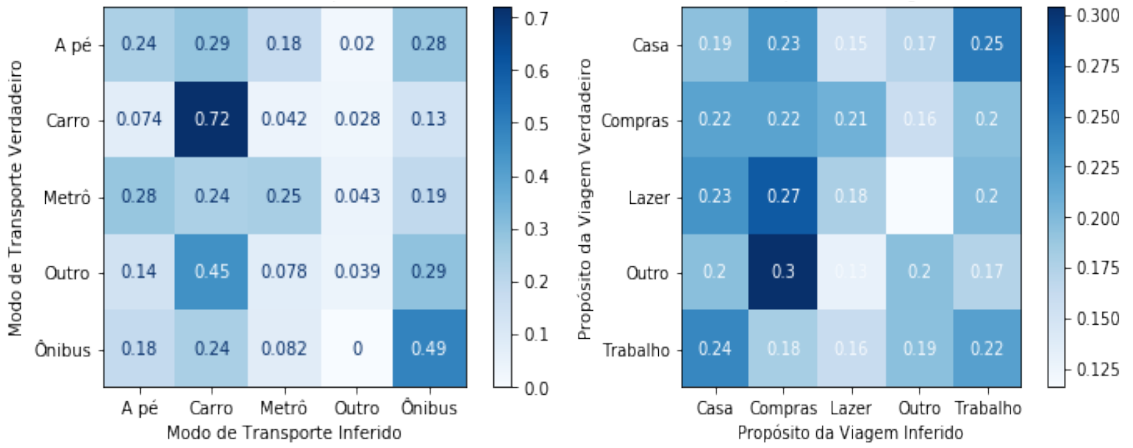


(a) (b)

Figura 4. Matrizes de confusão obtidas com as melhores *folds* da técnica *Random Forest* para classificação de Modo de Transporte (a) e Propósito de Viagem (b), respectivamente.

Na Figura 5 apresentamos as matrizes de confusão normalizadas das melhores *folds* obtidas com a técnica Regressão Logística. Na classificação do modo de transporte, esta técnica obteve um alto percentual de acerto na classe "Carro", mas teve percentual de acertos inferior a 50% para todas as outras classes. As classes em que o classificador

mais se confundiu foram “Carro” x “Outro”, onde houve um alto número de amostras classificadas como “Carro” que eram da classe “Outro”. Já na classificação do propósito da viagem a técnica também apresentou uma performance abaixo do esperado, com percentuais de acerto inferiores a 25% em todas as classes. As classes em que ocorreu maior número de falsos positivos foram as classes “Compras” x “Outro”, onde houve um alto número de amostras da classe “Outro” erroneamente classificadas como pertencentes a classe “Compras”.



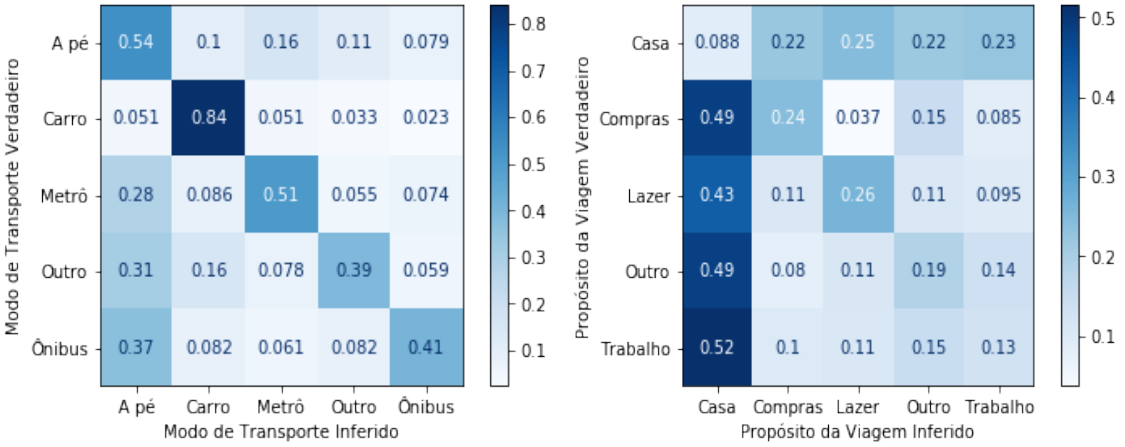
(a) (b)

Figura 5. Matrizes de confusão obtidas com as melhores *folds* da técnica de Regressão Logística para classificação de Modo de Transporte (a) e Propósito de Viagem (b), respectivamente.

Na Figura 6 apresentamos as matrizes de confusão normalizadas das melhores *folds* obtidas com a técnica Árvore de Decisão. Na classificação do modo de transporte, esta técnica obteve um percentual de acerto para a classe “Carro” superior a 80%, mas obteve um percentual de acertos para classe “Outro” inferior a 40%. O maior percentual de falsos positivos foi observado nas inferências da classe “A pé”, o que foi um dos principais causadores do desempenho inferior nas demais classes. Na classificação do propósito da viagem, o desempenho também foi abaixo do esperado, tendo havido um número elevado de inferências incorretas para a classe “Casa”.

Na Figura 7 apresentamos as matrizes de confusão normalizadas das melhores *folds* obtidas com a técnica *Support Vector Machine*. Na classificação do modo de transporte esta técnica foi a que apresentou o maior percentual de acertos para a classe “Metrô”. No entanto, é possível observar que houve um alto número de falsos positivos nas classificações desta mesma classe, o que prejudicou fortemente os resultados para as classes “A pé” e “Outro”.

Nas Tabelas 1 e 2 é possível confirmar as análises feitas a partir das matrizes de confusão. Nestas tabelas apresentamos os valores de acurácia, precisão, cobertura, e medida-F1 médias, por classe, para cada algoritmo nas tarefas de classificação do modo de transporte e propósito da viagem, respectivamente. É possível observar que a técnica de *Random Forest* obtém performance superior em todas as métricas de classificação do modo de transporte e que nenhuma das técnicas atinge um medida-F1 substancialmente superior a 20% na classificação de propósito da viagem, o que equivale a uma escolha



(a) (b)

Figura 6. Matrizes de confusão obtidas com as melhores *folds* da técnica de Árvore de Decisão para classificação de Modo de Transporte (a) e Propósito de Viagem (b), respectivamente.

aleatória.

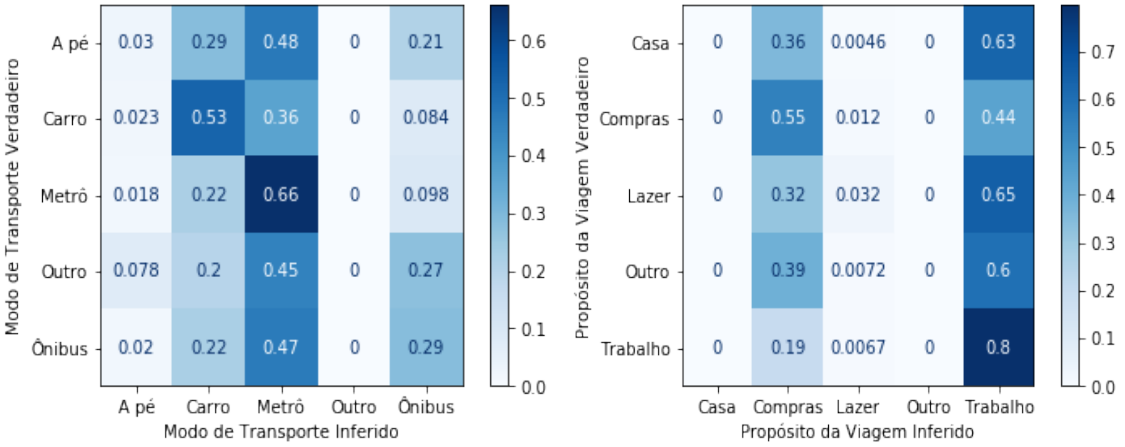
Tabela 1. Acurácia, precisão, cobertura, medida-F1 média por classe de modo transporte obtida com a melhor *fold* de cada algoritmo de aprendizado de máquina (Algoritmo).

Algoritmo	Acurácia	Precisão	Cobertura	Medida-F1
Regressão Logística	73.92%	34.78%	33.80%	30.58%
Árvore de Decisão	81.55%	54.00%	56.40%	53.8%
Support Vector Machine	72.08%	30.2%	22.2%	23.42%
Random Forest	**82.42%**	**56.00%**	**57.40%**	**56.00%**

Tabela 2. Acurácia, precisão, cobertura, medida-F1 média por classe de propósito da viagem obtida com a melhor *fold* de cada algoritmo de aprendizado de máquina (Algoritmo).

Algoritmo	Acurácia	Precisão	Cobertura	Medida-F1
Regressão Logística	68.10%	20.20%	20.60%	20.40%
Árvore de Decisão	67.21%	21.94%	22.48%	19.56%
Support Vector Machine	71.07%	27.64%	21.4%	16.8%
Random Forest	**67.59%**	**19.00%**	**21.96%**	**20.14%**

Dado isso, concluímos que a resposta para a pergunta Q1 é sim, os modelos de classificação gerados a partir do método proposto são capazes de identificar modo de transporte utilizados por moradores de grandes centros urbanos. Para a Q2, concluímos que a resposta também é sim, no entanto consideramos necessárias a realização de mais experimentos para uma melhor avaliação da acurácia dos modelos de classificação gerados, uma vez que esta foi inferior a expectativa. Por fim, consideramos que a resposta da Q3 é sim, a técnica de aprendizado de máquina utilizada no método proposto permitiu a geração de modelos de acurácia equivalente ou superior a outras alternativas do



(a) (b)

Figura 7. Matrizes de confusão obtidas com as melhores *folds* da técnica de *Support Vector Machine* para classificação de Modo de Transporte (a) e Propósito de Viagem (b), respectivamente.

estado da arte e, portanto, é considerada a melhor opção dentre as que avaliamos nestes experimentos. No entanto, a avaliação apresentada neste trabalho pode ser expandida para considerar um conjunto maior de técnicas de aprendizado, incluindo redes neurais e técnicas de aprendizado profundo, que requerem um ajuste mais fino dos hyperparametros e um maior número de amostras e iterações para a realização de um aprendizado efetivo.

5. Conclusões

Este trabalho abordou os problemas de identificação do modo de transporte e propósitos das viagens, realizadas por habitantes de centros urbanos, com base em dados socioeconômicos. Os experimentos realizados demonstraram que o método proposto apresentou um bom desempenho na classificação do modo de transporte, alcançando uma acurácia de 82% e um desempenho abaixo do esperado na classificação de propósito de viagem, inferior a 70%. A técnica de aprendizado de máquina utilizada permitiu a geração de modelos de classificação com acurácia superior que as demais alternativas do estado da arte analisadas e os resultados sugerem que é possível realizar a inferência do modo de transporte e propósito de viagem utilizando somente dados socioeconômicos.

Como trabalhos futuros consideraremos a realização de novos experimentos utilizando dados de mobilidade de outros centros urbanos, como do Reino Unido [Morris et al. 2013] e de Sidney[3], para uma avaliação mais robusta, assim como a avaliação de outras técnicas, como as redes neurais e de aprendizado profundo.

Referências

Bedogni, L., Di Felice, M., and Bononi, L. (2016). Context-aware android applications through transportation mode detection techniques. *Wireless Communications and Mobile Computing*, 16(16).

[3]https://www.transport.nsw.gov.au/data-and-research/passenger-travel/surveys/household-travel-survey-hts

Bohte, W. and Maat, K. (2009). Deriving and validating trip purposes and travel modes for multi-day gps-based travel surveys: A large-scale application in the netherlands. *Transportation Research Part C: Emerging Technologies*, 17(3):285–297.

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

Calabrese, F., Ferrari, L., and Blondel, V. D. (2014). Urban sensing using mobile phone network data: a survey of research. *ACM Computing Surveys*, 47(2):1–20.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Ermagun, A., Fan, Y., Wolfson, J., Adomavicius, G., and Das, K. (2017). Real-time trip purpose prediction using online location-based search and discovery services. *Transportation Research Part C: Emerging Technologies*, 77:96–112.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861 – 874.

Geurs, K. T., Thomas, T., Bijlsma, M., and Douhou, S. (2015). Automatic trip and mode detection with move smarter: First results from the dutch mobile mobility panel. *Transportation research procedia*, 11:247–262.

Gong, L., Kanamori, R., and Yamamoto, T. (2017). Data selection in machine learning for identifying trip purposes and travel modes from longitudinal gps data collection lasting for seasons. *Travel Behaviour and Society*.

Ma, Y. and Guo, G. (2014). *Support vector machines applications*. Springer.

Mäenpää, H., Lobov, A., and Lastra, J. L. M. (2017). Travel mode estimation for multi-modal journey planner. *Transportation Research Part C: Emerging Technologies*, 82:273–289.

Meng, C., Cui, Y., He, Q., Su, L., and Gao, J. (2017). Travel purpose inference with gps trajectories, pois, and geo-tagged social media data. In *IEEE International Conference on Big Data (BigData 2017), Boston, USA*, pages 1319–1324.

Montini, L., Prost, S., Schrammel, J., Rieser-Schüssler, N., and Axhausen, K. W. (2015). Comparison of travel diaries generated from smartphone data and dedicated gps devices. *Transportation Research Procedia*, 11:227–241.

Montini, L., Rieser-Schüssler, N., and Axhausen, K. W. (2014a). Personalisation in multi-day gps and accelerometer data processing. In *14th Swiss Transport Research Conference (STRC), Ascona, Switzerland*.

Montini, L., Rieser-Schüssler, N., Horni, A., and Axhausen, K. (2014b). Trip purpose identification from gps tracks. *Transportation Research Record: Journal of the Transportation Research Board*, (2405):16–23.

Morris, S., Humphrey, A., Pickering, A., Tipping, S., Templeton, I., and Hurn, J. (2013). National travel survey 2013. *The Department for Transport, NatCen Social Research, London, UK*.

NYC Department of Transportation (2017). Citywide mobility survey. Data retrieved from NYC Open Data, `https://data.cityofnewyork.us/`

Transportation/Citywide-Mobility-Survey-Trip-Diary/mpk5-48av.

Quintella, C. A. M. S., Andrade, L. C., and Campos, C. A. V. (2016). Detecting the transportation mode for context-aware systems using smartphones. In *IEEE Conference on on Intelligent Transportation Systems (ITSC'2016), Rio de Janeiro, Brazil*.

Schrammel, J., Busch, M., and Tscheligi, M. (2013). Peacox-persuasive advisor for co2-reducing cross-modal trip planning. In *8th International Conference on Persuasive Technology (PERSUASIVE), Sidney, Australia*.

Seo, T., Kusakabe, T., Gotoh, H., and Asakura, Y. (2017). Interactive online machine learning approach for activity-travel survey. *Transportation Research Part B: Methodological*.

Soares, E. F. d. S., Revoredo, K., Baião, F., de MS Quintella, C. A., and Campos, C. A. V. (2019). A combined solution for real-time travel mode detection and trip purpose prediction. *IEEE Transactions on Intelligent Transportation Systems*, 20(12):4655–4664.

Soares, E. F. S., Quintella, C. A. M. S., and Campos, C. A. V. (2017). Towards an application for real-time travel mode detection in urban centers. In *IEEE Vehicular Technology Conference (VTC-Fall'2017), Toronto, Canada*.

Song, Y.-Y. and Ying, L. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2):130.

Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia medica: Biochemia medica*, 24(1):12–18.

Stopher, P. R., Jiang, Q., FitzGerald, C., et al. (2005). Processing gps data from travel surveys. In *2nd international colloqium on the behavioural foundations of integrated land-use and transportation models: frameworks, models and applications*.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Wang, B., Gao, L., and Juan, Z. (2017a). Travel mode detection using gps data and socioeconomic attributes based on a random forest classifier. *IEEE Transactions on Intelligent Transportation Systems*.

Wang, H., Liu, G., Duan, J., and Zhang, L. (2017b). Detecting transportation modes using deep neural network. *IEICE Transactions on Information and Systems*, 100(5):1132–1135.

Xiao, G., Juan, Z., and Zhang, C. (2016). Detecting trip purposes from smartphone-based travel surveys with artificial neural networks and particle swarm optimization. *Transportation Research Part C: Emerging Technologies*, 71:447–463.

Ye, N., Gao, L., Juan, Z., and Ni, A. (2018). Are people from households with children more likely to travel by car? an empirical investigation of individual travel mode choices in shanghai, china. *Sustainability*, 10(12):4573.

Zhang, J., Wang, F.-Y., Wang, K., Lin, W.-H., Xu, X., and Chen, C. (2011). Data-driven intelligent transportation systems: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 12(4):1624–1639.