Uso de aprendizado supervisionado para análise de confiabilidade de dados de *crowdsourcing* sobre posicionamento de ônibus

Diego Vieira Neves, Felipe Cordeiro Alves Dias, Daniel Cordeiro

[1]Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

Abstract. *Intelligent Transportation Systems allows sensors and GPS devices to monitor public transport systems in Smart Cities. Capturing and processing this data should, in theory, allow systems to make the public transport more reliable and predictable for the citizens, which would improve the quality of life of the urban population and the environment. Insufficient or low-quality data, nevertheless, may prevent its use on such real-time systems. This work studies the use of data obtained from crowdsourcing as an alternative to augment this data. In order to mitigate the uncertainties introduced by the crowdsourced data, this work proposes a reliability model for crowdsourced data conceived for the São Paulo bus-based public transport system.*

Resumo. *Sistemas de Transportes Inteligentes permitem o uso de sensores e equipamentos de GPS para monitorar os sistemas de transportes públicos em Cidades Inteligentes. A captura e processamento desses dados permite, em tese, que o cidadão possa utilizar o transporte público com confiabilidade e previsibilidade, o que melhoraria a qualidade de vida da população urbana e o meio ambiente. Contudo, diversos fatores podem fazer com que esses dados sejam insuficientes ou de baixa qualidade para uso em tempo real. Este trabalho estuda o uso de dados obtidos via colaboração coletiva (crowdsourcing) como complemento dessas informações. Para mitigar as incertezas introduzidas pelo uso de crowdsourcing, este trabalho propõe um modelo de análise de confiabilidade dos dados coletados especializado para o sistema de transporte público (por ônibus) do município de São Paulo.*

1. Introdução

O conceito de Cidades Inteligentes (SC — *Smart Cities*) é utilizado para descrever cidades que promovem a integração de diversas tecnologias para solucionar os problemas ocasionados pelo crescimento populacional e, desta forma, criar respostas inovadoras e alinhadas às necessidades da população [De Santis et al. 2014, Batista et al. 2016]. Uma característica importante relacionada a SC que requer respostas inovadoras é a *mobilidade urbana*, que poderia ser melhorada com a disponibilização de serviços de transporte coletivo de qualidade [Figueiredo 2016].

Um problema observável nas cidades brasileiras é a qualidade dos seus serviços de transporte público, especialmente quando nos referimos ao modal ônibus. As prestadoras de serviços têm dificuldades para estabelecer e cumprir horários de itinerários das linhas de ônibus nas grandes metrópoles. A falta de um sistema confiável leva o usuário a não optar pela utilização desses serviços, o que agrava problemas urbanos sociais e ambientais.

As iniciativas em SC têm como objetivo resolver problemas como o mencionado anteriormente. Tais propostas utilizam dados coletados por meio de sistemas distribuídos, per-

tencentes as estruturas tecnológicas utilizadas por essas cidades como, por exemplo, os Sistemas de Transporte Inteligentes (ITS — *Intelligent Transport System*), que realizam a coleta de diversas informações por meio de sensores [Magalhães 2008]. Entretanto, as limitações decorrentes do uso de dispositivos físicos acoplados aos meios de transporte, tais como, indisponibilidade do sinal GPS ou falha na conexão 3G/4G, contribuem para que parte dos dados coletados pelos ITS não possa ser utilizada [Zhu et al. 2011, Pons et al. 2015].

Dados obtidos com *crowdsourcing* são eficazes no contexto de transporte público, pois permitem a obtenção de novas informações em tempo real [Cerotti et al. 2016]. Informações sobre a localização dos ônibus disponibilizadas voluntariamente pelos passageiros poderiam ser combinadas às informações obtidas pelos dispositivos acoplados aos ônibus urbanos, o que contribuiria para a criação de ITS eficientes e confiáveis [Figueiredo 2016, Batista et al. 2016].

Contudo, os dados informados por meio de *crowdsourcing* podem ser subjetivos, imprecisos e incorretos [Misra et al. 2014]. Portanto, cada nova informação obtida via *crowdsourcing* precisa ser analisada e sua confiabilidade assegurada antes de ser incorporada à base de dados de um ITS. Este trabalho propõe um novo modelo para a análise de confiabilidade desses dados por meio de métricas escolhidas para o contexto do modal ônibus, com técnicas de aprendizagem de máquina que permitem identificar de forma eficiente registros com informações não confiáveis (ausentes ou erradas).

2. Fundamentação Teórica

2.1. Cidades Inteligentes

O conceito de Cidades Inteligentes engloba iniciativas com o objetivo de atenuar e gerir eficientemente problemas urbanos que afetam a qualidade de vida da população, decorrentes do constante crescimento populacional [Caprotti 2016]. O termo SC tem sido utilizado para descrever cidades que buscam continuadamente o desenvolvimento urbano, mediante a políticas que estimulam a participação da população por meio de inteligências distintas, como a *(i)* humana, *(ii)* coletiva e *(iii)* artificial (por meio das TICs — Tecnologias da Informação e Comunicação) [Cury and Marques 2016, Batista et al. 2016].

No contexto de mobilidade urbana, um dos objetivos das SC é resolver os problemas relacionados à disponibilização de serviços de transportes com qualidade para a população [Figueiredo 2016]. Os ITS e as técnicas de *crowdsourcing* podem ser descritos com exemplos de soluções que contribuem para aumentar a qualidade dos serviços de transportes oferecidos e a atrair novos usuários [Sussman 2008, Misra et al. 2014, Cerotti et al. 2016].

2.2. Sistemas Inteligentes de Transporte

Os Sistemas de Transportes Inteligentes têm como objetivo coletar as informações sobre os veículos da rede de transporte público para utilizar esses dados em aplicações inteligentes [Figueiredo 2016]. Diferentes soluções em ITS são utilizadas para: *(i)* identificar a localização dos veículos através de sensores ou GPS, *(ii)* transmitir e receber grandes quantidades de dados, *(iii)* processar grandes quantidades de informações e *(iv)* utilizar essas informações para melhorar as condições do tráfego [Sussman 2008].

Um grupo de ITS relevante para esse trabalho são os Sistemas Avançados de Transporte Público (APTS - *Advanced Public Transportations Systems*), que podem ser descritos como um conjunto de aplicações que aumentam a qualidade, segurança e eficiência dos

sistemas de transporte públicos [Figueiredo 2005, Hwang et al. 2006]. Para facilitar o gerenciamento do transporte público, os APTS utilizam a estrutura de um sistema de Localização Automática de Veículos (AVL — *Automatic Vehicle Location*), que possibilita o rastreamento dos veículos em tempo real [Chowdhury and Sadek 2003].

2.3. Crowdsourcing

O termo *crowdsourcing* (*colaboração coletiva* ou *contribuição colaborativa*) se refere ao uso de uma rede distribuída de voluntários dispostos a resolver problemas, desenvolver novas tecnologias, contribuir com dados, etc. [Howe 2006]. As iniciativas baseadas neste conceito estão se tornando essenciais para as infraestruturas das SC, uma vez que possibilitam a captura de diversas informações que normalmente não poderiam ser capturadas por meio da utilização de abordagens e técnicas tradicionais [Cullina et al. 2015, Cerotti et al. 2016].

O uso de *crowdsourcing* em problemas de mobilidade urbana já foi investigado por outros trabalhos [Pedersen et al. 2013, Misra et al. 2014, Cullina et al. 2015, Cerotti et al. 2016]. Tais trabalhos se centraram no problema de como construir Sistemas de Informação ao Usuário (SIU) eficientes, confiáveis e em como permitir melhor interação entre usuários e gestores de serviços de transporte público. Os dados necessários para a criação destes sistemas são fornecidos pelos próprios usuários, por meio dos seus respectivos dispositivos móveis, o que possibilita a atualização constante dos dados. Nenhum dos trabalhos mencionados, entretanto, preocupou-se com a qualidade e confiabilidade dos dados fornecidos pelos usuários.

O aumento da capacidade computacional dos dispositivos móveis possibilitaram aplicações que utilizam técnicas de *crowdsourcing* [Pedersen et al. 2013]. Contudo, assim como ocorre com outras abordagens, os dados obtidos por meio de *crowdsourcing* também apresentam informações não confiáveis, que comprometem a confiabilidade dos dados coletados. Diversos autores [Mashhadi and Capra 2011, Allahbakhsh et al. 2013, Mousa et al. 2015, Daniel et al. 2018] abordam a necessidade da definição e adoção de métodos e técnicas para avaliar a qualidade e garantir a confiabilidade dos dados obtidos com o uso de *crowdsourcing*. Para mitigar as incertezas introduzidas pelo uso de *crowdsourcing*, este trabalho propõe o uso de técnicas de aprendizagem de máquina para criar um modelo de análise de confiabilidade e qualidade desses dados.

2.4. Técnicas de Aprendizagem de Máquina

Aprendizagem de Máquina (ML — *Machine Learning*) são técnicas que utilizam conceitos de inteligência artificial e/ou métodos estatísticos para realizar o reconhecimento de padrões [Mitchell 1997]. São utilizadas no desenvolvimento de sistemas inteligentes capazes de adquirir conhecimento de forma automática por meio da análise de um conjunto de dados [Ghotra et al. 2015]. Os algoritmos de classificação utilizados por essa abordagem são capazes de realizar aprendizagem interativa, mediante a análise de um conjunto de dados chamados de *amostra* [Bishop 2006]. Tais algoritmos podem ser utilizados para resolver problemas de classificação, regressão, *clustering* ou extração de regras, possibilitando, deste modo, a construção de modelos de predição [Ghotra et al. 2015].

Além dos algoritmos utilizados para criar o modelo, outros elementos considerados são a parametrização e complexidade computacional dos algoritmos, as variáveis que serão utilizadas pelo modelo, etc [Mitchell 1997, Ghotra et al. 2015]. A seleção

do melhor algoritmo para a criação de um modelo de predição é uma atividade considerada complexa, devido aos fatores que podem influênciar no desempenho do modelo desenvolvido [Ghotra et al. 2015]. As técnicas mais utilizadas por estudos correlatos [Wu et al. 2004, Wang et al. 2009, Biagioni et al. 2011, Altinkaya and Zontul 2013, Kormáksson et al. 2014] são:

- **Árvore de Decisão**: algoritmo composto por uma estrutura no formato de árvore, aonde cada nó interno representa um determinado teste em uma característica de um registro, e os arcos representam o resultado do teste realizado. A predição da variável-alvo é feita através de regras de decisão simples, inferidas pelos dados de treinamento [Ghotra et al. 2015, Witten et al. 2016].
- **K-Nearest Neighbour**: estrutura que realiza aprendizagem por analogia, ou seja, esse algoritmo relaciona cada um dos registros do conjunto de dados a um ponto em um espaço $m - dimensional$, sendo m o número de atributos de entrada que descrevem o conjunto de dados. Para classificar um novo registro, a similaridade com outros registros já conhecidos é calculada por meios da distância de tais registro ao novo registro [Witten et al. 2016].
- **Regressão logística**: método utilizado para entender a relação entre um conjunto de variáveis independentes (ou explicativas) e uma variável dependente (ou resposta) e construir um modelo que explique essa associação [Hosmer Jr et al. 2013].
- **Análise de discriminante linear**: técnica utilizada para identificar as características que discriminam uma determinada classe ou grupo de dados, e, assim, elaborar previsões a respeito de uma nova observação, identificando o grupo mais adequado a que ela deverá pertencer, em função de suas características [Hair et al. 2009].
- **Gaussian Naive Bayes**: algoritmo baseado no *Teorema de Bayes* e que é utilizado para calcular a probabilidade da ocorrência de um evento, baseando-se em probabilidades obtidas da análise dos eventos passados [Mitchell 1997, Witten et al. 2016].
- **Máquinas de Vetores Suporte**: método supervisionado para predição de rótulos utilizando técnicas de regressão e classificação. Baseia-se na construção de hiperplanos ou espaços dimensionais infinitos [Vapnik 1998].

3. Análise de confiabilidade de dados de *crowdsourcing*

3.1. Procedimentos de coleta e análise dos dados

O modelo de análise de dados de *crowdsourcing* para o transporte público utiliza duas bases de dados: a da *SPtrans*, que contém dados disponibilizados pela empresa São Paulo Transporte S. A. (SPTrans[1]); e a *base de dados crowdsourcing*, disponibilizada pela *startup* Scipopulis[2], especializada em mobilidade urbana.

A primeira base de dados contém os registros referentes aos trajetos realizados pelos ônibus da rede de transporte público da cidade São Paulo. Esses dados foram coletados pelo sistemas AVL instalados nos veículos. A SPTrans disponibiliza os dados coletados em tempo real para os desenvolvedores, acrescidos de informações como: identificação da linha, identificação e localização dos veículos, horários em que foram realizados os registros, entre outras.

Já a segunda base contém dados privados disponibilizados pela *startup* Scipopulis,

[1]São Paulo Transporte S. A.: `http://www.sptrans.com.br/`
[2]Scipopulis: `http://scipopulis.com/`

com informações disponibilizadas voluntariamente pelos usuários do aplicativo Coletivo[3]. Este aplicativo possui uma funcionalidade que permite que usuários informem se um determinado ônibus já passou enquanto o usuário esperava pelo seu ônibus. Quando o passageiro informa que o ônibus chegou no ponto de parada onde está o usuário, o sistema registra informações como (*i*) horário (data e hora atual em que registro foi realizado), (*ii*) localização (latitude e longitude obtida pelo GPS do usuário), (*iii*) identificador do ponto de parada e (*iv*) número da linha do ônibus. A Figura 1(a) apresenta uma visão geral da distribuição dos dados coletados por *crowdsourcing*.

(a) (b)

Figura 1. (a) Mapa de distribuição de localizações das contribuições coletivas, (b) Coleta de dados de posicionamento dos ônibus.

Os dados coletados são referentes aos trajetos realizados por 785 linhas de ônibus da cidade de São Paulo, durante o período de 1º de novembro à 31 de dezembro de 2016. As bases contém mais de 25 milhões registros, totalizando 12 GB de dados. A Figura 1(b) apresenta de forma esquemática como os dados são coletados. Contudo, é importante ressaltar que nem todos os registros contidos nas bases foram utilizados.

Os dados coletados passaram pelo seguinte processo de preparação e seleção, antes de serem utilizados: (i) *consolidação dos dados* em uma única base dos dados referentes aos trajetos realizados pelos ônibus; (ii) *estruturação e normalização* para remoção de dados duplicados, com falta de informação; (iii) *classificação manual* da confiabilidade do dado de acordo com o seguinte protocolo de validação:

- **Distância do usuário em relação ao ponto de ônibus** — (a) diretamente proporcional à probabilidade do dado ser confiável, se inferior a 300 metros ou (b) não confiável, caso contrário;
- **Distâncias anteriores e posteriores do veículo em relação ao ponto de parada informado pelo usuário** — (a) diretamente proporcional à probabilidade do dado ser confiável, se inferior a 10 quilômetros ou (b) não confiável, caso contrário;
- **Velocidade média entre as localizações anteriores e posteriores do veículo em relação a localização do ponto de ônibus informado pelo usuário** — (a) diretamente proporcional à probabilidade do dado ser confiável, se superior 10 km/h e inferior a 80 km/h ou (b) não confiável, caso contrário;
- **Tempos e distâncias dos registros realizados pelos usuários em relação à localização e aos tempos anteriores e posteriores do AVL** — (a) diretamente proporcional à probabilidade do dado ser confiável, se a diferença de tempo entre o registro capturado pelo sistema AVL em relação ao registro informado pelo usuário for inferior a 15 minutos ou (b) não confiável, se superior 15 minutos, e

[3]Aplicativo Coletivo: `https://www.facebook.com/appcoletivo`

(c) diretamente proporcional à probabilidade do dado não ser confiável, quando o tempo e a distância entre os registros capturados pelo equipamento AVL em relação ao registro informado pelo usuário forem inversamente proporcionais.

Por fim, após a execução das etapas anteriores chegamos ao conjunto de dados final contendo 971 registros que servirão como base para o modelo proposto. O conjunto de dados selecionado considera as seguintes informações: identificadores da linha, do veículo, do ponto de parada; tipo de trajeto realizado (ida ou volta); distância do usuário em relação ao ponto de parada obtida por meio das coordenadas de latitude e longitude capturadas a partir dos dispositivos móveis dos passageiros; os tempos em que foram realizados os registros pelos usuários (data e hora); distância da localização anterior do veículo em relação ao ponto de parada obtida por meio das coordenadas de latitude e longitude capturadas a partir dos sistemas AVL instalados nos veículos; os tempos em que foram capturadas as informações do sistema AVL (data e hora); velocidade média necessária para o deslocamento do veículo entre a localização anterior registrada pelo sistema AVL até a nova localização informada pelo usuário; e classificação da qualidade da informação (confiável = 1 e não confiável = 0).

3.2. Descrição do modelo de análise da confiabilidade

O modelo proposto tem como objetivo realizar, de forma automatizada, a análise de confiabilidade dos dados fornecidos pelos usuários (com *crowdsourcing*) sobre o horário de passagem dos ônibus nos pontos de paradas. Para cada registro informado por um usuário, o modelo deverá ser capaz de identificar inconformidades e/ou anomalias que possam influenciar ou comprometer a qualidade dos dados coletados.

Este trabalho avalia o uso de 6 algoritmos de predição diferentes: (i) Árvore de Decisão, (ii) *K-Nearest Neighbour*, (iii) Regressão logística, (iv) Análise de discriminante linear, (v) *Gaussian Naive Bayes*, e (vii) Máquinas de Vetores Suporte. Para a implementação dos algoritmos foi utilizada a biblioteca *Scikit-Learn*[4] [Pedregosa et al. 2011] e em cada modelo implementado, a confiabilidade do conjunto de dados foi analisada considerando-se 4 variáveis, sendo uma variável resposta (y) e 3 variáveis independentes (X), à saber:

- **Classificação**: variável resposta, indica a qualidade da informação dos registros analisados;
- **Distância do usuário**: variável independente que indica a distância entre a localização do usuário em relação ao ponto de parada informado pelo passageiro;
- **Distância anterior**: variável independente que indica a distância entre o ponto de parada informado pelo usuário em relação à localização anterior do veículo;
- **Velocidade**: variável independente que indica a velocidade média necessária para o deslocamento entre a localização anterior do veículo e o ponto de parada informado pelo usuário.

A divisão entre o conjunto de dados de treinamento e testes foi feita utilizando-se validação cruzada (*cross-validation*) [Kohavi 1995], que é usada para avaliar a capacidade de generalização do modelo (classificação de dados desconhecidos) [Witten et al. 2016]. Essa técnica ajuda a evitar os problemas de *overfitting* e *underfitting*, garantido um conjunto

[4] *Scikit-Learn* é uma biblioteca de código open source que possui diversos algoritmos de aprendizado de máquina implementados em linguagem de programação Python.

de dados adequado para o modelo desenvolvido [Witten et al. 2016]. Em particular, foi utilizado o método *K-fold cross-validation*, que consiste em dividir o conjunto de dados aleatoriamente em k segmentos chamados de *folds*. Os modelos desenvolvidos foram treinados e testados com a interação de 10-*folds*, conforme recomendações de diversos autores [Kohavi 1995, Witten et al. 2016].

Avaliou-se também o desempenho do algoritmo em termos de tempo de execução e de avaliação do modelo. Na literatura podem ser encontradas inúmeras medidas que são utilizadas individualmente ou em conjunto para avaliação de modelos de predição [Kohavi 1995, Witten et al. 2016]. Neste estudo, consideramos as seguintes medidas de desempenho: Matriz de Confusão, Acurácia, Precisão, Sensibilidade, Especificidade, *F-score* e *MCC* (*Matthews correlation coefficient*). O tempo de execução é apresentado como a média de 30 execuções.

4. Análise dos resultados

A Tabela 1 retrata as matrizes de confusão geradas para cada um dos 6 modelos utilizados, bem como os indicadores de desempenho para cada modelo. A matriz de confusão consolida a quantidade de registros realmente pertencentes a cada uma das classes (Não confiável e Confiável). Com isso, é possível calcular os indicadores de cada algoritmo como, por exemplo, precisão, sensibilidade, especificidade, *F-score*, MCC e acurácia.

Tabela 1. Indicadores de desempenho dos modelos

Algoritmos	Matriz de confusão			Indicadores de desempenho geral					Indicadores de treinamento		Indicadores de validação	
Árvores de Decisão	Classes	Não confiável	Confiável	Precisão	Sensibilidade	Especificidade	F-score	MCC	Acurácia	Tempo médio de execução (s)	Acurácia	Tempo médio de execução (s)
	Não confiável	88	0	1.00	0.99	1.00	0.99	0.99	0.9882 ± 0.02	0.040699	0.9966 ± 0.01	0.003127
	Confiável	1	203									
K-Nearest Neighbors	Classes	Não confiável	Confiável	Precisão	Sensibilidade	Especificidade	F-score	MCC	Acurácia	Tempo médio de execução (s)	Acurácia	Tempo médio de execução (s)
	Não confiável	71	1	0.98	0.80	0.99	0.88	0.85	0.9117 ± 0.04	0.053361	0.9349 ± 0.05	0.003896
	Confiável	18	202									
Regressão Logística	Classes	Não confiável	Confiável	Precisão	Sensibilidade	Especificidade	F-score	MCC	Acurácia	Tempo médio de execução (s)	Acurácia	Tempo médio de execução (s)
	Não confiável	76	4	0.95	0.85	0.98	0.89	0.86	0.9190 ± 0.06	0.113797	0.9418 ± 0.05	0.009664
	Confiável	13	199									
Gaussian Naive Bayes	Classes	Não confiável	Confiável	Precisão	Sensibilidade	Especificidade	F-score	MCC	Acurácia	Tempo médio de execução (s)	Acurácia	Tempo médio de execução (s)
	Não confiável	74	0	1.00	0.83	1.00	0.91	0.88	0.9352 ± 0.05	0.026670	0.9486 ± 0.04	0.002109
	Confiável	15	203									
Análise Discriminante Linear	Classes	Não confiável	Confiável	Precisão	Sensibilidade	Especificidade	F-score	MCC	Acurácia	Tempo médio de execução (s)	Acurácia	Tempo médio de execução (s)
	Não confiável	21	0	1.00	0.24	1.00	0.38	0.42	0.7570 ± 0.06	0.043261	0.7671 ± 0.05	0.002278
	Confiável	68	203									
Máquinas de Vetores Suporte	Classes	Não confiável	Confiável	Precisão	Sensibilidade	Especificidade	F-score	MCC	Acurácia	Tempo médio de execução (s)	Acurácia	Tempo médio de execução (s)
	Não confiável	5	0	1.00	0.05	1.00	0.10	0.19	0.7186 ± 0.06	0.938610	0.7123 ± 0.06	0.107141
	Confiável	84	203									

Os resultados mostram que os algoritmos de Árvores de Decisão, *Gaussian Naive Bayes* e Regressão Logística apresentaram os melhores indicadores. O modelo desenvolvido utilizando o algoritmo de Árvores de Decisão previu 100% dos registros classificados como confiáveis e 99,65% dos registros considerados como não confiáveis. A acurácia total do modelo foi de 99,66%. Além disso, o modelo apresentou excelentes indicadores

de desempenho com 100% de precisão, 97,75% de sensibilidade, 100% de especificidade, 98,86% de *F-score* e 98,39% de MCC.

Os algoritmos de *Gaussian Naive Bayes* e Regressão Logística apresentaram valores de acurácia similares. Porém, a partir da análise dos demais indicadores, identificamos que o algoritmo de *Gaussian Naive Bayes* apresentou o melhor desempenho. Este modelo obteve acurácia de 94,86%, e previu 100% dos registros classificados como confiáveis e 83,15% dos registros considerados como não confiáveis, além de apresentar indicadores de desempenho com 100% de precisão, 83,15% de sensibilidade, 100% de especificidade, 90,80% de *F-score* e 87,99% de MCC. Já o algoritmo de regressão logística apresentou acurácia de 94,18% e foi capaz de prever 98,03% dos registros classificados como confiáveis e 85,39% dos registros considerados como não confiáveis, e obteve indicadores de desempenho com 95% de precisão, 85,39% de sensibilidade, 98,03% de especificidade, 89,94% de *F-score* e 86,10% de MCC.

O modelo criado com o *K-Nearest Neighbour* identificou 99,50% dos registros classificados como confiáveis e 79,77% dos registros considerados como não confiáveis. A acurácia total do modelo foi de 93,49%. Esse modelo apresentou indicadores com 98,61% de precisão, 79,78% de sensibilidade, 99,51% de especificidade, 88,20% de *F-score* e 84,67% de MCC. Os algoritmos de Análise de Discriminante Linear e Máquinas de Vetores Suporte tiveram o menor desempenho dentre os outros modelos. O modelo com o algoritmo de Análise de Discriminante Linear obteve acurácia total 76,71% e identificou 100% dos registros classificados como confiáveis, contudo o modelo identificou apenas 23,59% dos registros considerados como não confiáveis. Esse algoritmo apresentou indicadores com 100% de precisão, 23,60% de sensibilidade, 100% de especificidade, 38,18% de *F-score* e 42,04% de MCC. Do mesmo modo, o modelo com o algoritmo de Máquinas de Vetores Suporte apresentou acurácia total de 71,23% e, embora tenha identificado 100% dos registros classificados como confiáveis, classificou apenas 5,61% dos registros como não confiáveis, e apresentou os seguintes indicadores, 100% de precisão, 5,62% de sensibilidade, 100% de especificidade, 10,64% de *F-score* e 19,93% de MCC.

5. Conclusão e Trabalhos Futuros

No contexto de transporte público, um dos problemas existentes é a pontualidade dos ônibus em relação aos horários de parada pré estabelecidos. A integração eficiente das diversas tecnologias existentes nas SC tem permitido a criação e identificação de inúmeras oportunidades para solucionar (ou amenizar) problemas como o mencionado, que afetam a qualidade dos serviços de transporte disponibilizado para a população. Contudo, a predição dos tempos de chegadas de ônibus é uma atividade que depende de diversos fatores aleatórios (ambiente estocástico). Como consequência, a precisão das predições realizadas pode ser prejudicada, uma vez que os erros de predição são ocasionados por fatores que não podem ser controlados como, por exemplo: atrasos em intersecções sinalizadas, número de passageiros em pontos de parada, etc.

A utilização de técnicas de *crowdsourcing* possibilita a obtenção de novos dados ou até mesmo a correção de informações existentes. No entanto, assim como em outras abordagens, dados de *crowdsourcing* também apresentam informações não confiáveis que podem comprometer a confiabilidade das atividades de predições realizadas. Este trabalho investigou o uso de aprendizado supervisionado para analisar a confiabilidade dos dados obtidos por *crowdsourcing*. Analisamos o desempenho de 6 algoritmos de ML em termos

de eficiência computacional e em qualidade de classificação. A análise dos resultados obtidos mostram que os modelos desenvolvidos utilizando os algoritmos de Árvores de Decisão, *Gaussian Naive Bayes* e Regressão Logística foram os mais adequados para análise de dados de mobilidade urbana. Acreditamos que a qualidade da classificação e o bom desempenho computacional apresentado pelos algoritmos os tornariam candidatos à serem utilizados na prática.

Como trabalhos futuros, pretendemos analisar a confiabilidade de outras formas de obtenção de dados como, por exemplo, as informações de mídias sociais e analisar o uso desses algoritmos na análise de fluxos de dados obtidos em tempo real.

6. Agradecimentos

Os autores agradecem à empresa Scipopulis pelos dados fornecidos. Esta pesquisa é parte do INCT da Internet do Futuro para Cidades Inteligentes financiado pelo CNPq, proc. 465446/2014-0, CAPES, proc. 88887.136422/2017-00 e FAPESP, proc. 2014/50937-1.

Referências

Allahbakhsh, M., Benatallah, B., Ignjatovic, A., Motahari-Nezhad, H. R., Bertino, E., and Dustdar, S. (2013). Quality control in crowdsourcing systems: Issues and directions. *IEEE Internet Computing*, 17(2):76–81.

Altinkaya, M. and Zontul, M. (2013). Urban bus arrival time prediction: A review of computational models. *International Journal of Recent Technology and Engineering*, 2(4):164–169.

Batista, D. M., Goldman, A., Hirata, R., Kon, F., Costa, F. M., and Endler, M. (2016). Interscity: Addressing future internet research challenges for smart cities. In *7th International Conference on the Network of the Future (NOF)*, pages 1–6.

Biagioni, J., Gerlich, T., Merrifield, T., and Eriksson, J. (2011). Easytracker: automatic transit tracking, mapping, and arrival time prediction using smartphones. In *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems*, pages 68–81. ACM.

Bishop, C. M. (2006). Pattern recognition and machine learning (information science and statistics) springer-verlag new york. *Inc. Secaucus, NJ, USA*.

Caprotti, F. (2016). *Eco-cities and the transition to low carbon economies*. Springer.

Cerotti, D., Distefano, S., Merlino, G., and Puliafito, A. (2016). A crowd-cooperative approach for intelligent transportation systems. *IEEE Transactions on Intelligent Transportation Systems*.

Chowdhury, M. A. and Sadek, A. W. (2003). *Fundamentals of intelligent transportation systems planning*. Artech House.

Cullina, E., Conboy, K., and Morgan, L. (2015). Measuring the crowd: a preliminary taxonomy of crowdsourcing metrics. In *Proceedings of the 11th International Symposium on Open Collaboration*, page 7. ACM.

Cury, M. J. F. and Marques, J. A. L. F. (2016). A cidade inteligente: uma reterritorialização/smart city: A reterritorialization. *Redes*, 22(1).

Daniel, F., Kucherbaev, P., Cappiello, C., Benatallah, B., and Allahbakhsh, M. (2018). Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Comput. Surv.*, 51(1):7:1–7:40.

De Santis, R., Fasano, A., Mignolli, N., and Villa, A. (2014). Smart city: fact and fiction. *Munich Personal RePEc Archive*.

Figueiredo, G. D. S. (2016). Técnicas de simulação para apoio à decisão em planejamento urbano. Master's thesis, Universidade Federal do Rio de Janeiro.

Figueiredo, L. M. B. (2005). *Sistemas inteligentes de transporte*. PhD thesis, Univ. do Porto.

Ghotra, B., McIntosh, S., and Hassan, A. E. (2015). Revisiting the impact of classification techniques on the performance of defect prediction models. In *Proceedings of the 37th International Conference on Software Engineering-Volume 1*, pages 789–800. IEEE Press.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., and Tatham, R. L. (2009). *Análise multivariada de dados*. Bookman Editora.

Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*, volume 398. John Wiley & Sons.

Howe, J. (2006). The rise of crowdsourcing. *Wired magazine*, 14(6):1–4.

Hwang, M., Kemp, J., Lerner-Lam, E., Neuerburg, N., and Okunieff, P. (2006). Advanced public transportation systems: state of the art update 2006. Technical Report FTA-NJ-26-7062-06.1, U.S. Department of Transportation.

Kohavi, R. (1995). The power of decision tables. In *European conference on machine learning*, pages 174–189. Springer.

Kormáksson, M., Barbosa, L., Vieira, M. R., and Zadrozny, B. (2014). Bus travel time predictions using additive models. In *IEEE Intl. Conference on Data Mining*, pages 875–880. IEEE.

Magalhães, C. T. d. A. (2008). Avaliação de tecnologias de rastreamento por gps para o monitoramento do transporte público por ônibus. Master's thesis, Univ. Federal do Rio de Janeiro.

Mashhadi, A. J. and Capra, L. (2011). Quality control for real-time ubiquitous crowdsourcing. In *Proceedings of the 2Nd International Workshop on Ubiquitous Crowdsouring*, UbiCrowd '11, pages 5–8, New York, NY, USA. ACM.

Misra, A., Gooze, A., Watkins, K., Asad, M., and Le Dantec, C. (2014). Crowdsourcing and its application to transportation data collection and management. *Transportation Research Record: Journal of the Transportation Research Board*, 2414:1–8.

Mitchell, T. M. (1997). Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, 45(37):870–877.

Mousa, H., Mokhtar, S. B., Hasan, O., Younes, O., Hadhoud, M., and Brunie, L. (2015). Trust management and reputation systems in mobile participatory sensing applications. *Comput. Netw.*, 90(C):49–73.

Pedersen, J., Kocsis, D., Tripathi, A., Tarrell, A., Weerakoon, A., Tahmasbi, N., Xiong, J., Deng, W., Oh, O., and de Vreede, G.-J. (2013). Conceptual foundations of crowdsourcing: A review of is research. In *Hawaii International Conference on System Sciences*. IEEE.

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.

Pons, I., Monteiro, J., and Speicys Cardoso, R. (2015). Big data para análise de métricas de qualidade de transporte: metodologia e aplicação. Technical report, ANTP.

Sussman, J. S. (2008). *Perspectives on intelligent transportation systems (ITS)*. Springer.

Vapnik, V. N. (1998). *Statistical learning theory*, volume 1. Wiley New York.

Wang, J.-n., Chen, X.-m., and Guo, S.-x. (2009). Bus travel time prediction model with ν-support vector regression. In *Intelligent Transportation Systems, 2009. ITSC'09. 12th Intern. IEEE Conf. on*, pages 1–6. IEEE.

Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Wu, C.-H., Ho, J.-M., and Lee, D.-T. (2004). Travel-time prediction with support vector regression. *IEEE transactions on intelligent transportation systems*, 5(4):276–281.

Zhu, T., Ma, F., Ma, T., and Li, C. (2011). The prediction of bus arrival time using global positioning system data and dynamic traffic information. In *Wireless and Mobile Networking Conference (WMNC), 2011 4th Joint IFIP*, pages 1–5. IEEE.