Sistema de Previsão do Tempo de Chegada dos Ônibus Baseado em Dados Históricos Utilizando Modelos de Regressão

Alternative Title: Prediction System of Bus Arrival Time Based on Historical Data Using Regression Models

Kássio R. Coquita
Centro de Informática - Universidade Federal de Pernambuco (UFPE)
50.733-970 – Recife/PE – Brasil
krc3@cin.ufpe.br

Arley R. R. Ristar
Centro de Informática - Universidade Federal de Pernambuco (UFPE)
50.733-970 – Recife/PE – Brasil
arrr2@cin.ufpe.br

Adriano L. I. Oliveira
Centro de Informática - Universidade Federal de Pernambuco (UFPE)
50.733-970 – Recife/PE – Brasil
alio@ cin.ufpe.br

Patrícia C. A. R. Tedesco
Centro de Informática - Universidade Federal de Pernambuco (UFPE)
50.733-970 – Recife/PE – Brasil
pcart@cin.ufpe.br

RESUMO

Os Sistemas Inteligentes de Transporte são aplicações de tecnologias da informação e comunicação que visam à melhoria da área de transporte. O fornecimento de informações sobre a chegada do ônibus nas paradas é muito importante e útil aos passageiros e gestores de trânsito. Este artigo apresenta a proposta de um sistema de previsão do tempo de chegada do ônibus na parada onde o usuário está localizado. Para isso, um estudo experimental na linha de ônibus Campina do Barreto de número 722 em Recife-PE foi realizado, comparando o modelo de Máquina de Vetor de Suporte para Regressão (SVR) e a rede neural do tipo Máquina de Aprendizado Extremo (ELM) na estimativa do tempo gasto para o ônibus percorrer paradas adjacentes. Os experimentos foram realizados utilizando os dados de log de GPS de ônibus na Região Metropolitana do Recife, e os resultados mostraram que para este sistema o SVR obteve uma performance melhor, quando comparado com a ELM.

Palavras-Chave

Sistemas Inteligentes de Transporte, Previsão, Regressão, SVR, ELM.

ABSTRACT

Intelligent Transportation Systems are applications of information and communication technologies aimed at improving the transportation area. Providing information about bus arrival time on the bus stop is very important and useful to passengers and transit managers. This paper presents a proposed bus arrival time prediction system at the bus stop where user is located. For this, an experimental study on bus route named Campina do Barreto No. 722 in Recife-PE, was performed by comparing the regression model for Support Vector Machine (SVR) and the neural network Extreme Learning Machine (ELM) to estimate the time it takes to go through adjacent bus stops. The experiments were performed using the bus GPS log data in the metropolitan region of Recife, and the results showed that for this application the SVR significantly outperforms ELM.

Categories and Subject Descriptors

C.4 [**Performance of Systems**]: measurement techniques, modeling techniques, performance attributes; D.4.8 [**Operating Systems**]: Performance - modeling and prediction; G.3 [**Probability and Statistics**]: robust regression, statistical computing.

General Terms

Algorithms and Performance.

Keywords

Intelligent Transportation Systems, Prediction, Regression, SVR, ELM.

1. INTRODUÇÃO

O tráfego nas grandes cidades brasileiras tem piorado a cada dia que passa. Na realidade brasileira, os congestionamentos nas vias têm levado o transporte público a se tornar ineficiente, dado o tempo gasto para se locomover de um ponto a outro. O transporte público seria uma alternativa para a redução do fluxo de veículo nas vias se não fosse a baixa qualidade do serviço que as empresas oferecem. Um dos exemplos da baixa qualidade desse serviço é o longo tempo de espera que os passageiros têm que enfrentar nos pontos de ônibus.

De acordo com [3], esse tipo de problema contribui para a relutância dos usuários em utilizarem esse meio de transporte. Prover a informação do tempo de chegada do ônibus pode permitir que os passageiros planejem melhor suas viagens, diminua o tempo de espera e ansiedade nas paradas, além de poder diminuir o risco de assaltos na parada, já que o usuário só precisaria ir para o ponto de ônibus quando o mesmo estivesse se aproximando.

Em um cenário em que as informações sobre o trânsito são de grande utilidade para os usuários, a área de Sistema de Transporte Inteligente ganha importância e surge para ajudar a melhorar a vida das pessoas. Além disso, prover informações de grande relevância para os usuários de transporte coletivo urbano

(ônibus), melhorando a qualidade do serviço é mais uma contribuição desse tipo de sistema.

Neste artigo, foram utilizados os dados históricos coletados a partir de dispositivos GPS instalados nos ônibus da Região Metropolitana do Recife. A linha de ônibus Campina do Barreto de número 722 foi selecionada para treinamento e validação dos modelos de regressão escolhidos. Os dados históricos de GPS foram submetidos a um pré-tratamento para deixá-los com as informações desejadas (não contidas nos dados originais), e posteriormente, utilizá-los para criar os modelos de previsão com a Máquina de Vetor de Suporte para Regressão, ou *Suport Vector Regression* (SVR) e com a Máquina de Aprendizado Extremo, ou em inglês, *Extreme Learning Machine* (ELM).

O restante deste artigo está organizado da seguinte forma: na Seção 2 são discutidos alguns trabalhos relacionados, já na Seção 3 e 4 são introduzidos os conceitos que cerceiam este projeto. Na Seção 5, é descrita a proposta do sistema. A descrição sobre os dados e o seu processo de pré-tratamento é realizada na Seção 6. A configuração do experimento é exposta na Seção 7. Os resultados preliminares obtidos com a implementação dos modelos de regressão podem ser vistos na Seção 8. Por fim, na Seção 9 é apresentada as conclusões e perspectivas para o trabalho futuro.

2. TRABALHOS RELACIONADOS

Foram encontrados na literatura alguns trabalhos que calculavam a previsão de tempo de chegada dos ônibus nas paradas, tais como em [5], que na tentativa de realizar previsões sobre o tráfego de maneira mais eficiente, estendeu vários métodos de previsão de séries temporais existentes capazes de lidar com os dados gravados em intervalos de tempo irregulares. Essa característica permitiu que os autores trabalhassem com dados de fonte intermitentes, tais como o Sistema de Posicionamento Global (GPS). Os resultados mostram que o melhor desempenho em termos da média absoluta do erro relativo foi obtido por meio de um modelo de rede neural.

Em [3] foi desenvolvido um algoritmo de previsão de aprendizagem não-supervisionada baseado em um modelo de dados históricos. Essa base histórica continha dados de localização e velocidade dos ônibus obtidos através do sensor de GPS instalados neles. Esses dados foram utilizados para treinar uma Rede Neural do tipo *backpropagation* para prever a velocidade média e o tempo de chegada em determinadas seções da estrada. Os resultados experimentais mostraram que o algoritmo proposto obteve uma melhor precisão na previsão**,** quando comparado com outros métodos baseados em dados históricos do tempo de viagem.

No estudo de [6], foi apresentada uma abordagem que combinava dados históricos e em tempo real para prever o tempo de chegada dos ônibus. Esse trabalho foi dividido em duas fases, em que na primeira, uma rede neural com Função de Ativação de Base Radial foi utilizada para aprender e aproximar a relação não-linear em dados históricos. Já na segunda fase, o Filtro de Kalman foi aplicado para ajustar a previsão realizada na primeira etapa utilizando dados *online*. Os resultados obtidos pela combinação desses modelos foram comparados com a Regressão Linear Múltipla, Rede Neural do tipo *backpropagation* e Rede Neural com Função de Ativação de Base Radial sem o ajuste *online*. Essa comparação mostrou que a abordagem proposta no estudo obteve um desempenho melhor na predição quando comparada aos outros modelos.

Assim, pôde-se observar na literatura consultada, um evidente e crescente interesse no que concerne ao tema proposto neste projeto. Esse levantamento bibliográfico também permitiu um maior entendimento das técnicas, ajudando na elaboração de estratégias para utilização das mesmas, ou até mesmo, na etapa de pré-processamento dos dados, de modo a contribuir efetivamente para o desenvolvimento deste trabalho.

Além disso, diante dos trabalhos citados, não foi encontrada uma pesquisa que aplicasse as Máquinas de Vetor de Suporte para Regressão ou as Máquinas de Aprendizado Extremo na previsão do tempo gasto pelo ônibus para percorrer paradas adjacentes, bem como, alguma que utilizasse a mesma característica dos dados para que pudesse ser realizada uma comparação mais detalhada entre elas e este trabalho em relação à performance de predição.

A presente pesquisa foi conduzida seguindo os conceitos referentes aos Sistemas Inteligentes de Transporte e aos métodos de regressão: Máquina de Vetor de Suporte para Regressão e Máquina de Aprendizado Extremo, conforme será discutido nas próximas seções.

3. SISTEMAS INTELIGENTES DE TRANSPORTE

Os Sistemas Inteligentes de Transporte, ou em inglês, *Intelligent Transportation Systems* (ITS) são aplicações de tecnologia para a área de transporte que integram informação, métodos de comunicação e tecnologias, a fim de auxiliar o sistema de transporte de uma região, integrando pessoas, estradas e veículos [1].

Um dos principais objetivos dos ITS é monitorar o tráfego das vias para otimizar as viagens, fornecedo suporte à mobilidade dos passageiros e veículos, de modo a aumentar a eficiência e a segurança nas redes de transportes atuais.

Uma das aplicações dos ITS é controlar os horários de chegada dos ônibus nas paradas, tendo em vista que o tempo de chegada é influenciado por vários fatores, como por exemplo, condições meteorológicas, o congestionamento do tráfego, as catástrofes naturais, e etc. Tais fatores resultam em atrasos no cronograma e inconveniência para os passageiros devido ao tempo de espera nas paradas de ônibus [4].

Neste artigo, é proposta uma abordagem na área de Sistemas de Transporte Inteligente para previsão do tempo de chegada do ônibus nas paradas, fazendo uso de técnicas de regressão e tendo como base um conjunto de dados históricos de logs de GPS.

4. OS MÉTODOS DE REGRESSÃO

Em problemas de regressão é dado um conjunto de treinamento $\{(x_1, y_1), ..., (x_n, y_n)\} \subset \chi \times \mathbb{R}$, onde χ denota o espaço dos padrões de entrada, por exemplo, $\mathbb{R}^d$. O principal objetivo da regressão é encontrar uma função $f(x)$ que melhor modele esses dados de treinamento [7]. No caso deste trabalho, há o interesse em construir e comparar modelos de regressão para prever o tempo gasto para o ônibus percorrer paradas adjacentes tendo como base os dados históricos.

4.1 Máquina de Vetor de Suporte para Regressão

Um dos modelos utilizados nesta pesquisa foi o ε-SVR, tipo de Máquina de Vetor de Suporte para Regressão, que define a função de perda ε-insensível. Esse tipo de função define uma

margem, ou faixa, em torno dos valores reais, como pode ser visto na Figura 1.

A principal ideia desse método é que os erros menores que certos limiares $\varepsilon > 0$ sejam ignorados. Ou seja, erros gerados por pontos que estejam na faixa do tubo são considerados zero. Em contrapartida, erros causados por pontos localizados fora da faixa são medidos pelas variáveis ξ e ξ^*, como mostrado na Figura 1.

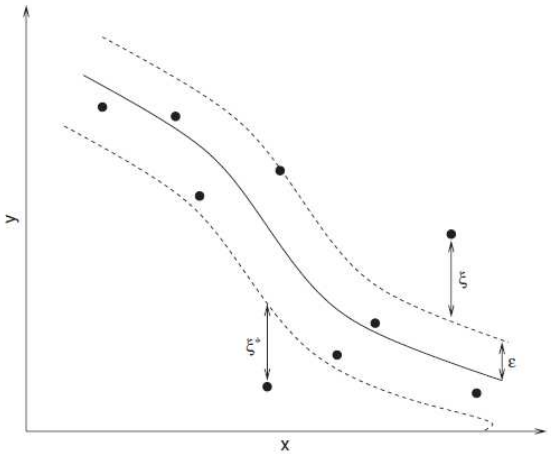

Figura 1. Regressão usando ε-SVR

4.2 Máquina de Aprendizado Extremo

Outro método para regressão adotado nesta pesquisa foi uma rede neural conhecida como Máquina de Aprendizado Extremo, ou em inglês, *Extreme Learning Machine* (ELM). A principal característica desta rede está na sua capacidade de aprendizado sem a necessidade de iterativa configuração dos parâmetros, ou seja, suas características são ajustadas aleatoriamente.

A ELM mostra que as camadas ocultas são importantes, mas podem ser ajustadas aleatoriamente e de forma independente do conjunto de dados de treinamento. Ao contrário dos métodos convencionais de aprendizagem que devem ter contato com os dados de treinamento antes de gerar os parâmetros dos neurônios ocultos, a ELM pode gerar aleatoriamente os parâmetros dos neurônios ocultos antes de ver os dados de treinamento [2].

Quando comparados os esforços desses métodos em relação à configuração, a ELM possui uma maior simplicidade, uma vez que requer como parâmetros apenas os números de neurônios e a função de ativação. Já o SVR necessita, em geral, dos parâmetros de custo (C), o nível de precisão da função aproximada (ε) e a largura (γ), em caso de kernel RBF.

5. A PROPOSTA DO SISTEMA

O sistema proposto terá a função de informar ao usuário o tempo de chegada do ônibus, tendo como entrada a linha do veículo, a parada em que o passageiro espera o veículo, o tipo do trajeto (ida ou volta), se a requisição está sendo realizada na hora de pico (07:00 às 09:00 e 17:00 às 19:00), e se é um dia de semana, além da informação da localização do ônibus desejado.

A Figura 2 mostra um diagrama esquemático do trajeto de um ônibus. A rota é dividida em segmentos s entre paradas adjacentes e i é o número da parada (i = 1, 2, 3,..., n).

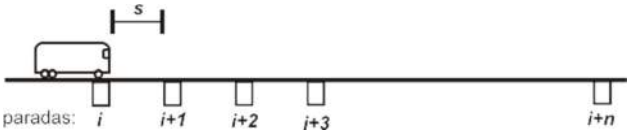

Figura 2. Esquema de uma rota de ônibus com várias paradas.

Quando um usuário requisita o tempo de chegada de um ônibus até a sua posição, o sistema irá verificar a localização do veículo desejado e contar o número de paradas entre o ônibus e o usuário. Caso o ônibus não esteja estacionado em uma parada, o sistema preverá o tempo que falta para ele atingir a próxima parada $P_{inicial}$ e somará com os tempos previstos que o veículo levará da parada $P_{inicial}$ para a parada P_{final}. O esquema funcional do sistema pode ser visto pela Figura 3.

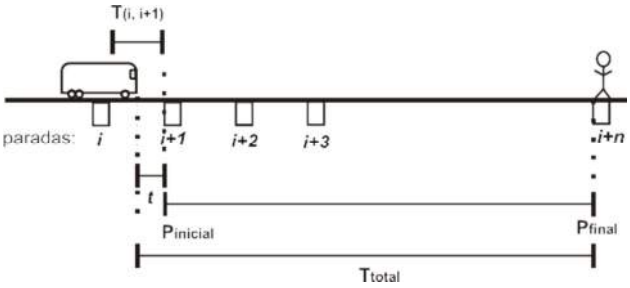

Figura 3. Esquema do sistema de previsão de tempo de chegada do ônibus.

Portanto, o tempo de chegada do ônibus na parada em que o usuário está localizado é então calculado de acordo com a Equação (1):

$$T_{total} = \sum_{i=P_{inicial}}^{P_{final}} \left(T_{(i,i+1)}\right) + t \qquad (1)$$

Na qual:

T_{total} = é o tempo total previsto que o ônibus levará da sua localização até chegar na parada P_{final};

$T_{(i,\, i+1)}$ = é o tempo previsto que o ônibus leva da parada i até a próxima parada ($i + 1$);

t = é o tempo que falta para o ônibus atingir a parada mais próxima.

Assim, fazendo uso do modelo de regressão apropriado para prever o tempo que o ônibus levará para percorrer paradas adjacentes, o sistema poderá usá-lo para a calcular o tempo total que o veículo consumirá para chegar até o usuário.

6. DADOS

O estudo utilizou os dados de GPS de ônibus da linha Campina do Barreto de número 722, fornecidos pela empresa Grande Recife Consórcios e Transporte (GRCT). Os dados foram coletados durante 8 dias, no período de 16 de agosto de 2012 até 23 de agosto de 2012. A linha Campina do Barreto (722), que é realizada pela empresa São Paulo, possui um trajeto que vai do bairro Cajueiro até o ponto de retorno no Cais de Santa Rita, em Recife. A escolha dessa linha de ônibus foi motivada por ser uma linha mais central na cidade do Recife, ter uma frequência de saída um pouco menor que a média (menos pontos coletados) e a linha possuir alguns trechos com ida e volta juntas (passam pela mesma rua em sentido contrário) e outros diferentes.

A empresa GRCT disponibilizou um arquivo de texto com os dados de todas as linhas e de todos os ônibus da região metropolitana do Recife, que são enviados para um sevidor pelo equipamento de GPS instalado nos ônibus. Esse arquivo, chamado nesta pesquisa de arquivo-fonte, são *dumps* de um banco de dados e cada informação é armazenada entre ponto e vírgulas numa ordem específica. Uma linha desse arquivo-fonte pode ser visto a seguir:

- SPA;722;541;2012-08-21;00:00:04.0000000;-8,022242;-34,88397;0;2012-08-21 00:00:11.7370000

A ordem das informações no formato acima pode ser constatada de acordo com as seguintes posições: (I) sigla da empresa de ônibus; (II) o número da linha; (III) o prefixo do veículo; (IV) a data da informação armazenada no banco de dados; (V) o horário que a informação foi armazenada no banco de dados; (VI) a latitude; (VII) a longitude; (VIII) a velocidade instantânea e (IX) o *timestamp* de aferição do dispositivo.

Vale destacar que o formato das coordenadas (latitude e longitude) utilizado por este trabalho é o padrão WGS 84.

Para o presente trabalho, será utilizado o *timestamp* de aferição do dispositivo, presente na posição IX, como o atributo de tempo. Os campos IV e V serão ignoradas, pois devido aos problemas de conexão para o envio das informação ao servidor de banco de dados, a data e hora que a informação foi armazenada difere da data e hora de aferição.

Nesse arquivo-fonte estão contidos quase 49 milhões de registros referentes a todas as linhas de ônibus e ao período citado acima. No entanto, nem todas as informações contidas no arquivo foram necessárias para o uso neste trabalho, uma vez que apenas os registros referentes à linha 722 foram utilizados.

Além desse arquivo-fonte, foram usados dados das localizações das paradas de ônibus. Essas localizações também foram fornecidas pela empresa de transporte e serão importantes para dividir o trajeto da linha de ônibus escolhida em segmentos. Para a linha de ônibus escolhida, foram contabilizadas 55 paradas, das quais 31 fazem parte do trajeto de ida e 24 de volta.

Assim, diante dessa quantidade de dados disponibilizada foi fundamental realizar um pré-processamento neles, para selecionar as informações necessárias, formatá-las, e remover as possíveis inconsistências contidas no arquivo.

6.1 Pré-processamento dos Dados

Nesta seção serão descritos os procedimentos para seleção, formatação e inserção dos registros no banco de dados para poderem ser utilizados. Além disso, serão explicados os procedimentos de obtenção das viagens realizadas pelos ônibus da linha 722 selecionada para este estudo.

Esta etapa é fundamental para o trabalho, pois nem todas as informações contidas no arquivo-fonte são necessárias, e também, pela possibilidade de encontrar inconsistências, uma vez que o projeto faz uso de dados reais. Além disso, esta seção de pré-processamento serve para explicar em qual banco de dados as informações tratadas serão armazenadas e qual aplicação utilizar para poder visualizá-las.

Conforme mencionado na seção anterior, as informações do arquivo-fonte que serão necessárias neste trabalho são: (I) a sigla da empresa de ônibus; (II) o número da linha; (III) o prefixo do veículo; (VI) a latitude; (VII) a longitude, e (VIII) a velocidade instantânea. Vale lembrar que as informações contidas na posição IV e V não serão utilizadas.

Com a definição das informações necessárias, a primeira etapa do pré-processamento consistiu em selecioná-las no arquivo-fonte e armazená-las em um arquivo-intermediário.

Esse resultado intermediário foi armazenado em um banco de dados espacial, para poder melhor manipular os dados resultantes. O SGBD escolhido foi o PostgreSQL, por ser uma ferramenta de *software* livre e possuir uma boa extensão para trabalhar com dados espaciais (PostGIS).

De acordo com as informações selecionadas, foi necessário criar tabelas para armazenar (a) as empresas de ônibus, (b) os ônibus que cada empresa possui, (c) as linhas de ônibus, e (d) o histórico de registro. Essa tabela de histórico armazena as informações dos dados que estão contidas no arquivo-intermediário, com colunas para a linha do ônibus, o ônibus que realizou o trajeto, a velocidade instantânea do ônibus, o *timestamp* e a geolocalização em formato POINT do PostGIS. Não foi criada uma chave primária ou identificador único para esta tabela, porém existem chaves estrangeiras para as linhas e os ônibus. As duas chaves estrangeiras em combinação com o *timestamp* são teoricamente únicos, mas, como esses dados são reais, é possível que haja ruídos entre eles.

De fato, alguns ruídos foram encontrados nos dados armazenados, pois como o envio de informações para o servidor não obedece a um intervalo regular de tempo, devido aos problemas de conexão, foi necessário remover os registros duplicados.

Após essa primeira fase, os dados da tabela histórico referentes à linha de número 722 do dia 16 de agosto de 2012 até 23 de agosto de 2012, foram selecionados e copiados para uma tabela temporária. A criação dessa tabela temporária se dá na tentativa de obter uma melhor performance no momento da utilização dos dados, já que do total de registros, apenas 58.694 foram obtidos com essa seleção.

É válido ressaltar que uma linha de ônibus pode ser executada por vários ônibus, e a linha 722 conta com uma frota de 6 ônibus para realizar o trajeto.

Os dados desses ônibus da frota da linha 722 podem ser visualizados na Figura 4, em que é possível verificar o trajeto formado pelos pontos espaciais referentes à linha 722. A ferramenta utilizada para gerar essa figura foi a Quantum GIS (ou QGIS), ferramenta com licença de software livre e que se integra muito bem ao PostgreSQL.

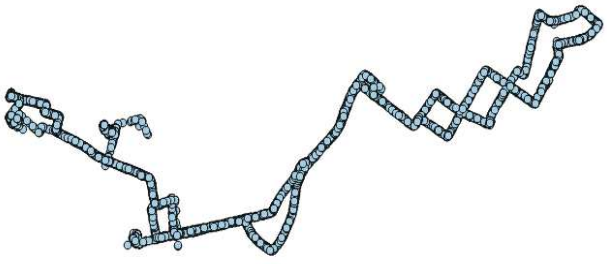

Figura 4. Visualização dos pontos de GPS do trajeto da linha 722.

Já na Figura 5, os dados podem ser visualizados utilizando a ferramenta Google Maps, com destaque para o trajeto de ida (em vermelho) e o de volta (em cinza).

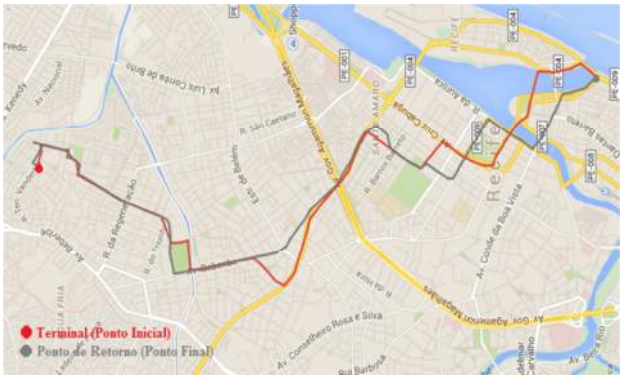

Figura 5. Visualização da linha 722 pelo Google Maps.

Posterior ao processo de seleção dos dados históricos dos logs dos ônibus, foi necessário submetê-los a alguns procedimentos, de modo a obter as informações desejadas (fatores de interesse) que se adequem às necessidades do sistema proposto descrito na Seção 5.

Portanto, com o objetivo de extrair desses métodos as previsões do tempo que o ônibus levará para percorrer paradas adjacentes, foram considerados os seguintes fatores de interesse: parada *i*, parada *i + 1*, tipo de trajeto, hora de pico, se é dia de semana, e o tempo de viagem entre as paradas, sendo esse último a variável dependente. Um exemplo desses fatores de interesse pode ser visto na Tabela 1.

Tabela 1. Exemplo dos fatores de interesses da linha 722.

Parada (i)	Parada (i + 1)	Tipo de Trajeto	Hora de Pico	Dia de Semana	Tempo de Viagem (seg)
1	2	ida	sim	sim	25.2
2	3	ida	sim	sim	48.0
3	4	ida	não	não	120.0
20	21	volta	sim	sim	78.0
21	22	volta	não	não	234.0
22	23	volta	não	não	30.0

Para poder chegar à esses fatores de interesse foi de suma importância identificar as coordenadas dos pontos iniciais e finais do trajeto de ida e volta, ou seja, onde estão localizados o terminal e o ponto de retorno da linha de ônibus escolhida.

Essa informação está disponível na página de consulta de itineráios no site da empresa fornecedora dos dados utilizados. Dessa forma, esses dados foram adquiridos manualmente e aplicados como pontos de referência para a divisão dos trajetos entre as viagens.

Após a determinação do ponto inicial e final da rota, foram identificadas todas as paradas de ônibus que estavam no trajeto de ida ou volta desta linha, conforme pode ser visto na Figura 6. Essas paradas de ônibus foram usadas para separar em partes menores os trajetos, e com isso obter os segmentos listados na Tabela 1. Com a realização desse processo, foram extraídos um total de 10.662 segmentos para a ida e volta.

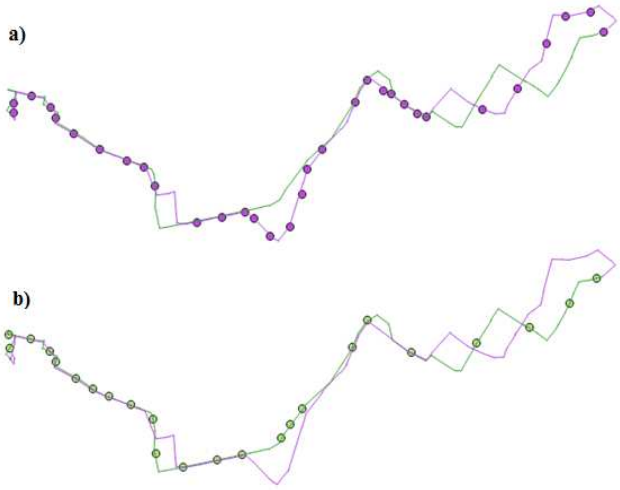

Figura 6. a) paradas de ônibus do trajeto de ida; b) paradas de ônibus do trajeto de volta.

É preciso mencionar que o tempo gasto por um ônibus para percorrer as paradas adjacentes foi calculado a partir do *timestamp* de pontos de GPS do ônibus. Como esses pontos de GPS nem sempre têm exatamente a mesma localidade das paradas, foi necessário ajustar o tempo que o ônibus gastaria para percorrer o segmento. A Figura 7 exibe o esquema realizado para calcular o tempo de viagem de um segmento (entre as paradas *i* e *i+1*).

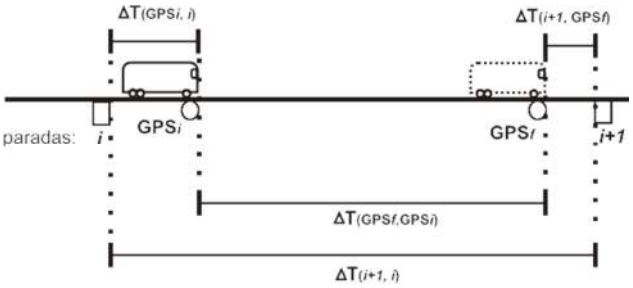

Figura 7. Esquema do cálculo do tempo de viagem de um segmento.

Portanto, o tempo que um ônibus leva para percorrer um segmento, definido entre as paradas *i* e *i +1* é então calculado de acordo com a Equação (2):

$$\Delta T_{(i+1,i)} = \Delta T_{(GPS_f,GPS_i)} + \Delta T_{(GPS_i,i)} + \Delta T_{(i+1,GPS_f)} \quad (2)$$

Na qual:

GPS_i = é o ponto de GPS inicial que o ônibus está localizado;

GPS_f = é o ponto de GPS final que o ônibus está localizado;

$\Delta T_{(i+1,i)}$ = é o tempo de viagem entre duas paradas adjacentes realizadas pelo mesmo ônibus;

$\Delta T_{(GPS_f,GPS_i)}$ = é a diferença de tempo que um ônibus leva para percorrer os pontos de GPS final e inicial;

$\Delta T_{(GPS_i,i)}$ = é a diferença de tempo que um ônibus leva para percorrer a parada *i* e seu ponto de GPS inicial;

$\Delta T_{(i+1,GPS_f)}$ = é a diferença de tempo que um ônibus leva para percorrer o seu ponto de GPS final e a parada *i+1*.

Assim, diferentemente das infomações contidas no arquivo-intermediário, foram obtidos os fatores de interesse independente, tais como: as paradas *i* e *i + 1*, o tipo do trajeto, se a medição foi realizada no horário de pico, e se a medição foi obtida em dia de semana, além do fator dependente, que é o tempo de viagem entre duas paradas adjacentes (em segundos).

Vale lembrar que o horário de pico considerado na pesquisa é das 07:00 às 09:00, pela manhã, e das 17:00 às 19:00 à noite. Além disso, esses fatores e seus respectivos valores, conforme mostrado na Tabela 1, serão chamados de dados históricos do tempo de viagem para cada segmento entre as paradas de ônibus adjacentes.

7. CONFIGURAÇÕES DO EXPERIMENTO

A regressão realizada neste trabalho utilizou o conjunto de dados históricos do tempo de viagem, mecionado na seção anterior. Este conjunto de dados possui cinco variáveis independentes – parada *i*, parada *i + 1*, tipo de trajeto, hora de pico e se é dia de semana – e uma variável dependente – tempo de viagem entre duas paradas adjacentes.

Neste projeto, o objetivo principal é estimar o tempo de viagem do ônibus, em segundos, baseado em dados históricos.

O tempo de viagem do ônibus em um determinado segmento (entre as paradas *i* e *i + 1*) pode ser denotado como $T_{(i,\ i+1)}$, sendo $1 \leq i \leq n$, para o ônibus que está percorrendo a rota através da parada *i* (*i* = 0,1, 2, 3,..., *n*). Dessa forma, $T_{(i,\ i+1)}$ pode ser obtido através das variáveis: parada *i*, parada *i +1*, tipo de trajeto, hora de pico e se é dia de semana.

As simulações dos algoritmos SVM do tipo "eps-regression" e ELM para regressão foram realizadas utilizando as bibliotecas LibSVM e elmNN do R, respectivamente. Além disso, os dados utilizados foram normalizados entre 0 e 1, conforme indicado pelas bibliotecas.

Vale destacar que a medida de performance utilizada neste trabalho é a média percentual do erro absoluto, ou em inglês, *mean absolute percentage error* (MAPE). A MAPE é dada pela Equação (3), onde $\{x_i\}$ é a observação real, $\{\hat{x}_i\}$ é o valor previsto e *N* é o valor do número de dados.

$$\text{MAPE} = \frac{100}{N} \times \sum_{i=1}^{N} \left| \frac{x_i - \hat{x}_i}{x_i} \right| \quad (3)$$

Para estimar os parâmetros ideais para o modelo, foi utilizada a validação cruzada do tipo *k-fold* em combinação com o *grid search*. Na validação cruzada *k-fold*, o conjunto de dados é particionado em *k* subconjuntos mutuamente exclusivos, onde *k* é o número total de subconjuntos. O conjunto de treinamento é construído utilizando *k* – 1 subconjuntos para estimar os parâmetros, e o subconjunto restante é utilizado para validação ou teste do modelo. Este processo é repetido *k* vezes, onde a cada iteração são empregados diferentes dados para previsão. Neste estudo foi utilizado um *k* igual a 10.

O método de *grid search*, foi aplicado na variação dos valores dos parâmetros necessários ao SVR e ELM. Nas simulações realizadas para o SVR foram considerados os parâmetros *C*, ε e γ para o kernel RBF. Nesse método, os parâmetros assumiram os seguintes valores: *C* = {10, 100, 1000}, ε = {10^{-3}, 10^{-4}, 10^{-5}} e γ = {1, 10^{-1}, 10^{-2}}. Dessa forma, considerando as diferentes combinações para os parâmetros, foram executadas 27 simulações com o kernel RBF.

Já nas simulações utilizando o método ELM foram variados os parâmetros referentes às funções de ativação e aos números de neurônios. Nesse método os parâmetros assumiram os seguintes valores: função de ativação = {*symmetric hard-limit*, base radial}, nº de neurônios = {10, 15, 20, 25, 30, 35, 40, 45, 50}, além de um *k-fold* igual a 10. Logo, considerando as diferentes combinações para os parâmetros, foram executadas 18 simulações no total. Vale salientar que para cada combinação de parâmetros da ELM a validação cruzada (10 *k-folds*) foi executada 30 vezes.

8. ANÁLISE COMPARATIVA DOS RESULTADOS

A análise dos resultados produzidos pelos métodos de regressão foi realizada pela comparação entre as medidas de performance (MAPE) obtidas para cada modelo.

A Tabela 2 compara as medidas de performance para o SVR considerando o kernel RBF, e variando os parâmetros *C*, ε e γ. Já a Tabela 3, compara o desempenho dos parâmetros para a rede ELM, alterando os valores do número de neurônio e da função de ativação.

Ainda na Tabela 2 é possível verificar que os valores da MAPE ao longo da alternância dos parâmetros não variou muito. Para os resultados obtidos com o ELM, exibidos na Tabela 3, é possível verificar que a MAPE para função de ativação *symmetric hard-limit* aumentou à medida que o número de neurônios também foi aumentando. Além disso, o mesmo aconteceu com a MAPE para o ELM com função de ativação de base radial.

Tabela 2. Comparação do desempenho do SVR de acordo com a variação dos parâmetros

Método/Parâmetros	MAPE	Desvio Padrão
(C = 10, $\varepsilon = 10^{-3}$, γ= 1.0)	3.514	0.857
(C = 10, $\varepsilon = 10^{-3}$, $\gamma = 10^{-1}$)	2.484	1.331
(C = 10, $\varepsilon = 10^{-3}$, $\gamma = 10^{-2}$)	2.101	0.646
(C = 10, $\varepsilon = 10^{-4}$, γ= 1.0)	3.246	1.049
(C = 10, $\varepsilon = 10^{-4}$, $\gamma = 10^{-1}$)	2.470	1.373
(C = 10, $\varepsilon = 10^{-4}$, $\gamma = 10^{-2}$)	1.958	0.434
(C = 10, $\varepsilon = 10^{-5}$, γ= 1.0)	3.103	1.007
(C = 10, $\varepsilon = 10^{-5}$, $\gamma = 10^{-1}$)	2.390	1.211
(C = 10, $\varepsilon = 10^{-5}$, $\gamma = 10^{-2}$)	2.045	0.499
(C = 100, $\varepsilon = 10^{-3}$, γ= 1.0)	2.624	1.713
(C = 100, $\varepsilon = 10^{-3}$, $\gamma = 10^{-1}$)	2.710	1.822
(C = 100, $\varepsilon = 10^{-3}$, $\gamma = 10^{-2}$)	2.159	0.629
(C = 100, $\varepsilon = 10^{-4}$, γ= 1.0)	2.564	1.762
(C = 100, $\varepsilon = 10^{-4}$, $\gamma = 10^{-1}$)	2.677	1.879
(C = 100, $\varepsilon = 10^{-4}$, $\gamma = 10^{-2}$)	2.344	0.772
(C = 100, $\varepsilon = 10^{-5}$, γ= 1.0)	2.576	1.598
(C = 100, $\varepsilon = 10^{-5}$, $\gamma = 10^{-1}$)	2.771	2.204
(C = 100, $\varepsilon = 10^{-5}$, $\gamma = 10^{-2}$)	2.341	0.885
(C = 1000, $\varepsilon = 10^{-3}$, γ= 1.0)	5.716	11.657
(C = 1000, $\varepsilon = 10^{-3}$, $\gamma = 10^{-1}$)	2.698	1.720
(C = 1000, $\varepsilon = 10^{-3}$, $\gamma = 10^{-2}$)	2.564	1.547
(C = 1000, $\varepsilon = 10^{-4}$, γ= 1.0)	5.729	11.559
(C = 1000, $\varepsilon = 10^{-4}$, $\gamma = 10^{-1}$)	2.827	1.794
(C = 1000, $\varepsilon = 10^{-4}$, $\gamma = 10^{-2}$)	2.552	1.526
(C = 1000, $\varepsilon = 10^{-5}$, γ= 1.0)	5.797	11.633
(C = 1000, $\varepsilon = 10^{-5}$, $\gamma = 10^{-1}$)	2.812	1.893
(C = 1000, $\varepsilon = 10^{-5}$, $\gamma = 10^{-2}$)	2.533	1.489

Tabela 3. Comparação do desempenho do ELM de acordo com a variação dos parâmetros

Método/Parâmetros	MAPE	Desvio Padrão
(nh = 10, fun = symmetric hard-limit)	6.39	2.16
(nh = 15, fun = symmetric hard-limit)	7.38	2.32
(nh = 20, fun = symmetric hard-limit)	8.50	2.31
(nh = 25, fun = symmetric hard-limit)	9.75	2.56
(nh = 30, fun = symmetric hard-limit)	9.86	2.23
(nh = 35, fun = symmetric hard-limit)	10.48	1.90
(nh = 40, fun = symmetric hard-limit)	11.35	1.75
(nh = 45, fun = symmetric hard-limit)	12.52	3.27
(nh = 50, fun = symmetric hard-limit)	12.88	2.04
(nh = 10, fun = radial basis)	8.35	1.85
(nh = 15, fun = radial basis)	10.97	1.23
(nh = 20, fun = radial basis)	13.25	3.27
(nh = 25, fun = radial basis)	14.56	3.30
(nh = 30, fun = radial basis)	19.02	4.67
(nh = 35, fun = radial basis)	25.31	9.98
(nh = 40, fun = radial basis)	43.17	19.93
(nh = 45, fun = radial basis)	85.74	38.99
(nh = 50, fun = radial basis)	153.63	122.06

Os resultados mostraram que o melhor valor da MAPE para o SVR foi de 1,958%, em que a combinação de parâmetros foi: C = 10, $\varepsilon = 10^{-4}$ e $\gamma = 10^{-2}$. Já para o ELM, o melhor resultado da MAPE foi de 6,39%, quando combinada a função de ativação *symmetric hard-limit* e o número de neurônios igual a 10.

Como o SVR obteve o melhor desempenho na previsão, um segundo experimento foi realizado, refinando seus parâmetros com base nos melhores valores alcançados anteriormente. Esses parâmetros assumiram os seguintes valores: $C = \{10, 50, 90\}$, $\varepsilon = \{10^{-3}, 5x10^{-4}, 5x10^{-5}\}$ e $\gamma = \{10^{-2}, 5x10^{-2}, 10^{-3}\}$.

Esse segundo experimento foi executado na tentativa de alcançar melhores resultados dos que já se tinham para o SVR. Ele foi realizado sob os mesmos procedimentos do experimento anterior, ou seja, com uma validação cruzada do tipo *k-fold*, com k = 10 e um *grid search* variando os novos valores dos parâmetros C, ε e γ.

Com a realização desse novo experimento, a MAPE do SVR foi reduzida de 1,958% para 1,492%, em que a combinação de parâmetros foi: C = 50, $\varepsilon = 10^{-3}$ e $\gamma = 10^{-3}$.

Após esse processo de determinação dos parâmetros, o SVR e o ELM foram reconstruídos utilizando os parâmetros que geraram o melhor desempenho para cada método. Além disso, um modelo clássico de regressão linear foi implementado pela função "lm" do R, para comparar os resultados. Essa comparação com o modelo linear se faz necessária, para saber se um método de regressão mais simples conseguiria obter uma performance melhor na previsão.

Para o treinamento e teste, foram selecionados 75% e 25% dos dados, respectivamente. Com essa execução dos métodos, a medida de desempenho (MAPE) obtida para cada método foi de 1,009% para o SVR, 4.499% para o ELM e 4.844% para o Linear.

Na Figura 8, parte da amostra foi selecionada para exibir um contraste entre os dados reais e as previsões realizadas por cada método.

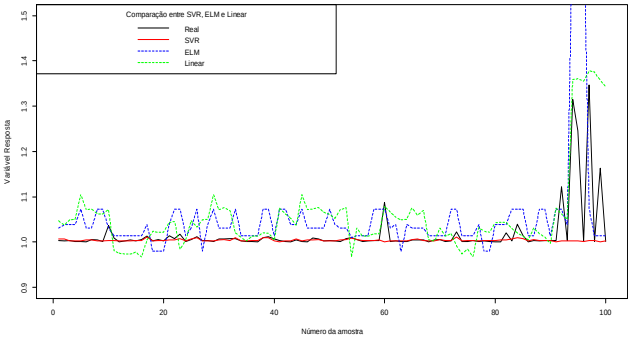

Figura 8. Comparação entre os métodos SVR, ELM e Linear para previsão

Assim, comparando os resultados das medidas de desempenho entre os métodos SVR, ELM e Linear aplicados para prever o tempo que o ônibus levará para percorrer paradas adjacentes, é possível verificar que o SVR obteve uma performance superior ao ELM e ao modelo clássico de regressão Linear.

9. CONCLUSÃO

O presente trabalho avaliou a performance de dois modelos para previsão do tempo de chegada do ônibus na parada. Foram investigados os métodos de Máquina de Vetor de Suporte para Regressão (SVR), bem como *Extreme Learning Machine* (ELM). Tais métodos utilizaram, para prever o tempo de chegada, as variáveis independentes: identificação das paradas adjacentes (i e $i + 1$), o tipo do trajeto, se a medição foi realizada no horário de pico e se é dia de semana. Essas informações foram extraídas à partir de dados históricos de GPS de ônibus cedidos pela empresa de transporte público do Recife, Grande Recife Consórcios e Transporte.

Os resultados de cada método foram comparados levando em consideração a média percentual do erro absoluto, ou MAPE. De acordo com as simulações, o método SVR obteve uma média percentual de 1.009%, que é menor que as médias de 4.499% e 4.844% obtidas pelos modelos ELM e Linear, respectivamente.

Para versões futuras deste trabalho, ficará a possibilidade de expandir e melhorar o sistema proposto. Uma das primeiras possibilidades a ser adicionada ao projeto, poderia ser a implementação da funcionalidade para calcular o tempo de espera do ônibus nas paradas. Outra possibilidade futura está na utilização da velocidade instantânea para melhorar a precisão da previsão. Além disso, características inerentes a linha do ônibus como, por exemplo, a quantidade de sinais de trânsito e quantidade de cruzamentos na via, podem ser consideradas.

10. AGRADECIMENTOS

Este trabalho foi apoiado pelo INES - Instituto Nacional de Ciência e Tecnologia para Engenharia de Software (http://www.ines.org.br/), financiado pelo CNPq, processo 573964/2008-4.

11. REFERÊNCIAS

[1] An, S.-H., Lee, B.-H., Shin, D-R. 2011. A Survey of Intelligent Transportation Systems. Communication Systems and Networks (CICSyN), 2011 Third International Conference, pp.332-337.

[2] Huang, G., Wang, D. H., Lan, Y. 2011. Extreme learning machines: a survey, em: International Journal of Machine Learning and Cybernetics, vol. 2, issue 1, p. 107-122.

[3] Pan, J., Dai, X., Xu, X., Li, Y. 2012. A Self-Learning Algorithm for Predicting Bus Arrival Time Based on Historical Data Model. Cloud Computing and Intelligent Systems (CCIS), IEEE 2nd International Conference, vol. 3, Hangzhou, p. 1112-1116.

[4] Raut, R. D., Goyal, V. K. 2012. Public transport Bus Arrival Time Prediction with Seasonal and Special Emphasis on Weather Compensation changes using RNN, em: International Journal of Advanced Research in Computer and Communication Engineering, vol. 1, issue 6.

[5] Ye, Q., Szeto, W. Y. e Wong, S. C. 2012. Short-Term Traffic Speed Forecasting Based on Data Recorded at Irregular Intervals, em: Intelligent Transportation Systems, vol. 13, no. 4.

[6] Wang, L., Zuo, Z. e Fu J. Bus Arrival Time Prediction Using RBF Neural Networks Adjusted by Online Data, 9th International Conference on Traffic & Transportation Studies, (2014), p. 67-75.

[7] Oliveira, A. L. I. Estimation of Software Project Effort with Support Vector Regression, Neurocomputing 69, (2006), p. 1749–1753.