Universidade de São Paulo

Instituto de Matemática e Estatística

Geração de Dados Sintéticos para Testes de Modelos de Machine Learning

Proposta de Trabalho

Rafael de Oliveira Magalhães
Orientadores: Ana Cristina Vieira de Melo
Flávio Soares Correa da Silva

São Paulo, 22 de abril de 2024

1 Introdução

Este trabalho é continuação de um projeto de iniciação científica cuja tema era testes de sistemas baseados em aprendizado de máquina, portanto será introduzido o assunto deste projeto e o que foi realizado para uma compreensão completa da relação com os dados sintéticos.

Atualmente, o emprego do aprendizado de máquina na resolução de problemas tem apresentado melhores resultados em comparação com abordagens tradicionais de inteligência artificial ou estatísticas. Entretanto, este tipo de solução não pode ser amplamente utilizado, sobretudo em aplicações críticas ou que exijam um alto grau de confiabilidade, dada a característica de se comportar como uma "caixa preta", ou seja, o funcionamento desses modelos não é transparente, o que impede a compreensão exata e não permite garantir a confiabilidade dos resultados.

Neste contexto, estudou-se ferramentas e métodos de teste para sistemas baseados em aprendizado de máquina, a fim de possiblitar uma maior confiabilidade por meio dos testes, da adequação dos dados de treinamento e da adaptabilidade tanto da rede neural quanto dos testes às variações nos dados.

Com este propósito, abordou-se o problema de previsão de tráfego, o qual é um problema de relevância prática e interessante pois representa um fenômeno do mundo real que é sensível à mudanças de padrões sociais (como a mudança do trabalho presencial para o remoto) ou à fatores ambientais e temporais (como a ocorrência de queimadas ou de feriados). Além disso, sendo um problema do mundo real, é habitual dos dados disponíveis apresentarem irregularidades ou dados faltantes e com erros, atributos que formam um fator comum com outros problemas, o que permite estender os resultados para outros domínios. E esta baixa qualidade dos dados disponíveis foi a principal dificuldade enfrentada durante o projeto, pois limitou os testes realizados, no entanto não impediu de observar que a escolha dos dados tem muito mais influência no resultado em comparação com a escolha do modelo.

Tendo em vista este cenário, o uso de dados sintéticos como ferramenta de teste apresentou-se como uma possível solução, pois permite controlar características dos dados, simular cenários menos usuais de se constatar em dados reais e gerar quantidades expressivas de amostras em um curto período e com baixo custo, o que não ocorre com os dados reais para este problema, os quais exigem grande investimento para a instalação e manutenção de sensores e um grande período de tempo para a construção de um conjunto de dados.

Além dos motivos citados anteriormente, os dados sintéticos são interessantes pois podem servir como substituto aos dados reais em situações nas quais os dados reais disponíveis são insuficientes ou de baixa qualidade. Também, para o estudo deste tema, será necessário primeiramente desenvolver métodos de geração destes dados, o que já é relevante como área de estudo independente para este problema. Portanto, a fim de

atingir o objetivo principal que é o uso de dados sintéticos como ferramentas de testes, estes e outros objetivos secundários também serão estudados.

2 Objetivos e Cronograma

Dado o contexto detalhado na seção anterior, os objetivos deste trabalho serão:

- Desenvolver métodos para a geração de dados sintéticos para o problema de previsão de tráfego.
- Estudar diferentes modelos de machine learning para este problema, e o respectivo funcionamento de cada um deles.
- Desenvolver e estudar critérios de avaliação da qualidade dos dados sintéticos, e de comparação com reais, a fim de averiguar a possibilidade da substituição dos dados reais por sintéticos.
- Estudar as características dos dados mais influentes para este problema, e para cada modelo, permitindo adaptar a geração dos dados sintéticos em função destes fatores, e também aumentando a compreensão do funcionamento dos modelos de machine learning, o que é importante ao selecionar dados reais.

Detalhando os tipos de geração de dados sintéticos, optou-se por duas abordagens, a primeira se assemelhará a uma simulação de um sistema viário, de modo que este método implemente as características do mundo real na geração, enquanto o segundo método será mais artificial, gerando os dados puramente através de funções matemáticas e estatísticas de modo a replicar características e propriedades dos dados e não do mundo real.

Em relação os modelos de machine learning, selecionou-se um modelo de GNN, descrito no artigo [1], um modelo de rede neural totalmente conectada e um modelo de regressão linear. Como se tratam de modelos com funcionamentos distintos, apenas esses três serão suficiente para a comparação dos dados.

Por fim, pensou-se em três abordagens para a avaliação dos dados, as quais serão utilizadas de forma conjunta:

- Comparação dos Modelos: Os modelos de machine learning selecionados serão utilizados também como ferramenta, treinando os modelos com cada dado gerado e comparando os valores das métricas.
- Estatística: Serão avaliados métricas estatísticas e realizados testes de hipótese.
- Gráfica: Tendo em vista a multidimensionalidade dos dados, esta abordagem gerará imagens correspondentes aos dados, a fim de permitir uma análise visual por padrões, também permitindo o estudo e manipulação dos dados por técnicas de processamento de imagens.

Um cronograma aproximado pode ser visto a seguir:

	Mês										
	1	2	3	4	5	6	7	8	9	10	11
Desevolvimento de métodos de geração dos dados sintéticos	X	X	X	X							
Estudo dos modelos			X	X	X						
Desenvolvimento de critérios para avaliação dos dados		X	X	X	X	X	X				
Estudo das características mais influentes dos dados e dos modelos				X	X	X	X	X	X	X	
Escrita da Monografia						X	X	X	X	X	X

Comentando o cronograma, algumas atividades já estão sendo desenvolvidas deste o início do ano e planeja-se iniciar a escrita da monografia com bastante antecedência, pois já há material resultante do projeto de iniciação científica que será necessário descrever detalhadamente para um entendimento completo do trabalho atual.

3 Referências Bibliográficas

[1] ZHANG, Y.; CHENG, T.; REN Y. *Graph deep learning method for short-term traffic forecasting on large road networks. Computer-Aided Civil and Inf*, v. 34, p. 877–896, 2019. https://doi.org/10.1111/mice.12450.

[2] Yotam Alexander; Nimrod De La Vega; Noam Razin; Nadav Cohen. *What Makes Data Suitable for a Locally Connected Neural Network? A Necessary and Sufficient Condition Based on Quantum Entanglement.* Tel Aviv University. https://doi.org/10.48550/arXiv.2303.11249.

[3] Tom Verbin; Noam Razin; Nadav Cohen. *What Makes Data Suitable for a Locally Connected Neural Network? A Necessary and Sufficient Condition Based on Quantum Entanglement.* Tel Aviv University. https://doi.org/10.48550/arXiv.2211.16494.

[4] CALIFORNIA DEPARTMENT OF TRANSPORTATION. *Caltrans Performance Measurement System (PeMS).* 2023. Disponível em: <https://pems.dot.ca.gov/>

[5] JIANG, W.; LUO, J. *Graph neural network for traffic forecasting: A survey. ScienceDirect - Expert Systems with Applications*, v. 207, 2022. https://doi.org/10.1016/j.eswa.2022.117921

[6] SHAYGAN, M.; et al. *Traffic prediction using artificial intelligence: Review of recent advances and emerging opportunities. ScienceDirect - Transportation Research Part C*, v. 145, 2022. https://doi.org/10.1016/j.trc.2022.103921