Relatório final de Iniciação Científica

Predizendo níveis de pobreza utilizando imagens de satélites e formulários socieconômicos.

João Pedro Donaire Albino

Orientador: Prof. Dr. Clayton Pereira

Departamento de Computação, Faculdade de Ciências,
Universidade Estadual Paulista Júlio de Mesquita Filho
Av. Eng. Luiz Edmundo Carrijo Coube, 14-01 - Vargem Limpa,
CEP 17033-360 - Bauru - SP.

25 de Junho de 2019

1 Introdução

A pobreza é uma das chagas complexas e importantes de nossa sociedade atualmente, pois se trata de um problema dificilmente controlado e trabalhado por métodos efetivos de combate [1]. Atrelado a isso também temos uma escassez de dados relevantes que possam mensurar com consistência indicadores de qualidade de vida e de poder aquisitivo da população. Esses dois fatores desencadeiam um notável esforço para mapear e diagnosticar o cenário da pobreza em determinadas regiões, assim perpetuando o problema da má distribuição de renda no país [2].

Por outro lado, atualmente, a gestão de dados no Brasil melhorou, porém ainda temos [PROCURAR SOBRE GESTÃO].

TEXTO

1.1 Origem do Tema

A necessidade da aplicação mapeamento sobre pobreza sempre foi algo necessário e a aplicação de métodos para diminuir custos é algo que pode auxiliar na sua realização. Por tanto, trazer métodos tecnológicos para é algo que poderia auxiliar em tal tema, facilitando a aplicação de recursos financeiros, normalmente governamentais, para problemas que de fato poderiam ser combatidos com investimentos.

1.2 Justificativa da Pesquisa

Em setembro de 2015, líderes dos 193 países membros da Organização das Nações Unidas (ONU) aprovaram um plano global de desenvolvimento sustentável, com o objetivo de melhorar os indicadores econômicos, sociais e ambientais para as próximas gerações [5]. Algo relevante para a proposta de pesquisa em questão é averiguar que a primeira dessas metas configura-se em eliminar todas as formas de pobreza no mundo. Essa decisão foi tomada ao levar em consideração que cerca de 705,5 milhões de pessoas vivem atualmente na extrema pobreza [6].

Para exemplificar através de dados locais, uma das principais dificuldades do Brasil reside em direcionar recursos para as população mais pobre [7]. Mesmo que através de formulários e pesquisas se torna possível identificar regiões com altos índices de pobreza, sofremos em obter dados relevantes que mensuraram, indicam e agrupam regiões por níveis de renda ou outras indicadores importantes.

Outro ponto relevante para se levar em consideração na proposta do trabalho diz respeito à quantidade de investimentos necessários para realizar uma pesquisa intensiva bem esclarecedora sobre indicadores de pobreza. Se levarmos em consideração países que não possuem recursos abundantes, obter dados relevantes se torna algo difícil. Segundo o próprio Instituto Brasileiro de Geografia e Estatística (IBGE), o orçamento do Censo de 2010 realizado no país fora calculado em R$ 1,677 bilhão [8].

Analisando tais fatos, chega-se em uma investida para inferir indicadores precisos utilizando dados *open-source* gratuitos já disponíveis na *Internet*. Esse é um dos tópicos que a Tecnologia de Dados que há algum tempo vem sendo trabalhado e será tratado dentro desta pesquisa.

1.3 Formulação do problema

É possível correlacionar intensidade luminosa com riqueza econômica utilizando dados open-source e Machine Learning?

1.4 Objetivos da Pesquisa

A pesquisa tinha como escopo principal criar um sistema utilizando Python e R totalmente automatizado para extrair, estruturar, carregar e classificar imagens de satélites noturnos do Estado de São Paulo, sendo estes de origem open-source.

1.5 Estrutura do Trabalho

O capítulo 1 discute-se sobre a introdução da pesquisa e objetivos da mesma. Nos dois seguintes trabalha-se apoio teórico e bibliográfico sobre o tema bem como metodologias de pesquisa aplicadas ao mesmo. O capítulo 4 apresenta alguns dados coletados no desenvolvimento da pesquisa e o 5, por sua vez, as considerações finais sobre o desenvolvimento da pesquisa.

2 Rerenferencial Teórico

2.1 Kmeans

2.2 SVM

3 Metodologia

Toda a pesquisa foi baseada em 4 macro principais estágios de análise de dados: Extração, Estruturação, Classificação e Predição. Além disso, para detalhar tem-se 7 microestágios correspondentes.

Divide-se a pesquisa em:

1. Extração de dados de uma origem open-source;
2. Estruturação das imagens no formato PGM;
3. Extração dos dados para uma planilha CSV;
4. Estruturação dos dados socioecônomicos abertos [CONFIRMAR COMO CHAMAM DADOS ABERTOS];
5. Concepção de classes baseados nos dados socioeconômicos;

6. Predição com modelos de Machine Learning;

7. Validação dos dados.

Na primeira etapa da pesquisa foi necessário um modelo para obter elementos em imagens de satélite e sócioeconômicos. A linguagem escolhida desde o início do projeto para construir o sistema foi o Python em conjunto com o R. Tal escolha baseou-se na alta demanda da mesma atualmente e na quantidade de bibliotecas que auxiliam nos estudos de Machine Learning. Foram utilizados o seguinte banco de dado imagético: National Geophysical Data Center Nighttime Lights Time Series para compor as imagens de dados de satélites noturnos. Além disso, todas as imagens foram computadas em TIFF, por padrão da origem. Porém foram convertidas para PGM como falaremos posteriormente.

Para o segundo momento, foi necessário definir padronização dos formatos dos dados. Nessa etapa, foi estabelecido que seria utilizado PGM como dados de imagem, PORQUE [EXPLICAR AQUI]. Além disso, todos os dados seriam estruturados em planilhas csv, por conta de [EXPLICAR MOTIVAÇÃO AQUI. PROCURAR ARTIGO NA NET].

Para o terceiro momento, extrai-se a intensidade luminosa dos mapas noturnos colocados em CSV com 198 pixels [CONFIRMAR CONVERSÃO]. Nesse caso, cada imagem é convertida em uma linha com 198 colunas. Depois disso foi necessário criar classes dos dados. Para tanto, usaram-se os índices de Produto Interno Bruto (PIB) de 2010 obtidos na pesquisa do IBGE, e foram auferidos classes utilizando KMEANs e PIB para comparar com o classificador SVM. Por fim, utilizamos o SVM baseado em 3 classes: menor, médio e maior. Ambas se baseiam em cidades com mínimo, médio e máximo das cidades de São Paulo.

Na primeira etapa da pesquisa é necessário um modelo bem treinado e preparado para identificar elementos em imagens de satélite, possibilitando no final um sistema útil para identificação de níveis de pobreza. Portanto, em uma primeira instância, deve-se utilizar um modelo de CNN treinado a fim de identificar características básicas de imagens. A linguagem escolhida para construir o sistema será o Python. Tal escolha baseia-se na alta demanda da mesma atualmente e na quantidade de bibliotecas que auxiliam nos estudos de *Machine Learning*.

Para o segundo momento, deve-se desenvolver uma aprendizagem em relação a imagens de mapas diurnos e noturnos. Primeiro, extrai-se a intensidade luminosa do mapas diurnos para, em seguida, identificar a correlacionar essa aprendizagem às imagens de mapas diurnos.

Serão utilizados os seguintes bancos de dados imagéticos: Mapas de satélites diurnos (Google Static Maps API) e de satélites noturnos (National Geophysical Data Center, Version 4 DMSP-OLS Nighttime Lights Time Series).

Em um terceiro momento, será necessário correlacionar as atribuições possíveis com os mapas aos dados socioeconômicos já obtidos pelos questionários. Para a análise desses dados, a pesquisa utilizará dados de formulários brasileiros de pesquisa de consumo fornecidos por instituições que fornecem dados abertos, como o próprio IBGE. Por fim, o produto final do projeto

Por fim, será feito uma avaliação final dos dados obtidos. Será avaliado se a classificação pelas imagens de fato condizem com os dados brutos obtidos pelos formulários.

4 Apresentação e Análise de dados

No primeiro mês de projeto, foi estruturada a origem dos dados que seriam utilizados, para que pudesse assim ser definido modelo e classificação dos dados. Para tanto, todas as procuras foram guiadas em encontrar dados abertos. Na pesquisa [CITA PESQUISA DE STANDFORD], eles utilizaram dados do portal National Geophysical Data Center para predizer dados da Nigéria e Uganda. A justificativa inicial foi vinculada às facilidades e qualidades dos dados do portal, portanto, partiu-se do mesmo pressuposto no projeto. E a escolha beneficiou em automatizar o processo da obtenção grande quantidades dos dados.

IMAGEM MOSTRANDO DADO SENDO OBTIDO E SALVO NAS PASTAS

Nos outros meses 2, foi focada a criação de algoritmos para realizar tal tarefas. A funcionalidade básica do algoritmo se reunimia na imagem seguinte.

FLUXOGRAMA MOSTRANDO A OBTENÇÃO DOS DADOS

Primeiramente, esperamos obter dados de 10 km por 10 km em arquivos TIFF. A escolha da quilometragem assegura que a localização obtida no mapa estivesse na análise. A contabilização realizada foi uma avaliação de um ponto como:

1. km lon = 10*(360/23903.297);
2. km lat = 10*(360/40075.00)

Realizou-se a obtenção de 3.075 imagens realizando o download baseado nas máximas e mínimas longitudes e latitudes do estado de Estado de São Paulo. Por fim, foram obtidos dados de todos satélites, porém utilizaram-se imagens apenas do ano de 2010 os satélites. Obtidos os dados, foi escolhido convertê-los para uma análise de pixels em 198. Depois de elencadas as imagens, foi preciso rotular os dados com as referentes cidades. Para tanto utilizamos a API do Google Maps, com ele foi possível enviar as latitutes e longitudes para obter o nome da área referente.Com isso, foi possível identificar XX de XX nos municípios do estado de São Paulo. Após tal fato, foi possível escolher a maneira mais aconselhável para predizer os dados. Primeiramente, elencaram-se em duas classes para usar como dados rotulados. A primeira classe foi escolhida baseada na classificação World Bank (https://blogs.worldbank.org/opendata/new-country-classifications-income-level-2017-2018, a qual divide como PIB: - Baixo: Renda menor ou igual a U$3955 - Médio: Renda entre U$3955 e U$12235 - Alto: Maior ou igual a U$12235 Com essa classificação foram obtidos os seguintes dados:

IMAGEM COM OS DADOS DA CLASSIFICAÇÃO COM PIB

Na segunda classe procurou-se rotular os dados otimizando com um classificador. KMEANs trouxe as seguintes classes para análise:

MOSTRAR DADOS COM O KMEANS

Com os dados rotulados, utilizou-se o SVM para classificar as imagens auferidas. Essas 3 classes baseadas na quantidade de pixels, estabelecemos 3 cidades que seriam usadas como base para o classificador. Do mais baixo para o maior tivemos: - Menor = Lucianópolis com uma soma de 0 pixels; - Médio = Pirangi com 4558.0 pixels; - Alto = São Paulo com a soma de 114735.0 pixels.

O resultado dessa classificação foi conferido em conjunto com os outros dados rotulados. Por final, foi possível o resultado de 68,74% utilizando KMEANs e 54% utilizando os classificadores da ONU.

5 Cronograma

O objetivo desta sessão é formalização do plano de distribuição das etapas do projeto. O projeto tem a previsão de um ano de realização. A Tabela 1 apresenta esse cronograma.

ID	Tarefas	Meses											
		1	2	3	4	5	6	7	8	9	10	11	12
1	Estudo sobre Machine Learning: CNN												
2	Estudo sobre Transfer Learning												
3	Estudo sobre utilização de LSMS												
4	Construção de modelo de CNN												
5	Implementação do primeiro treinamento de imagens												
6	Aperfeiçoamento do modelo de CNN												
7	Aplicação do modelo CNN em imagens de satélites												
8	Aperfeiçoamento do modelo de CNN												
9	Construção do modelo de obtenção de dados da Avaliação de Padrões de Vida (LSMS)												
10	Aplicação do modelo CNN nos dados da Avaliação de Padrões de Vida												
11	Validação dos resultados												
12	Escrita de relatório parcial												
13	Escrita de artigos científicos e relatório final												

Figura 1: Cronograma de atividades da presente proposta

6 Conclusões

No final deste período de pesquisa, considerando a quantidade de dados obtidos, os modelos aplicados e a acurácia auferida, avalia-se com um bom resultado da mesma. Primeiramente, foram utilizados cerca de 3.075 itens de 10 km x 10 km no formato PGMs , algo que compactua com cerca de 644 cidades do estado de São Paulo, correspondendo XX% de todas. Já em um segundo momento, estruturaram-se os dados de maneira automatizada assim possibilitando utilizar dois classificadores de dados e utilizar ao menos um método de predição. Por fim, alcançando cerca de XX% de correspondência dos dados.

COLOCAR FIGURA MOSTRANDO RESULTADOS

Porém, pode-se dizer que a pesquisa possui alguns pontos de melhora. Primeiramente, na utilização mais teórica dos dados de base, trazendo conhecimentos socioeconômicos mais refinados para então usar outros índices de desempenho econômico que iriam além do PIB. Além disso, seria muito importante utilizar mais de uma método de Machine Learning para validar os resultados esperados. Pensa-se em continuar com os estudos para implementação da OPF (COLOCAR REFERÊNCIA), por exemplo, que provou-se ser aconselhável para uso de dados como PGM. Por fim,

Referências

[1] G. Alejandro, "Medição da pobreza: o que tem na linha?" *Centro de Pobreza internacional*, 2004.

[2] R. Barros, R. Henriques, and R. Mendonça, "A estabilidade inaceitável: Desigualdade e pobreza no brasil," *Instituição de Pesquisa Econômica Aplicada*, p. 29, 2001.

[3] J. Blumenstock, G. Cadamuro, and R. On, "Predicting poverty and wealth from mobile phone metadata," *Science*, 2015.

[4] J. Henderson, A. Storeygard, and D. Weil, "Measuring economic growth from outer space," *American Economic Review*, 2012.

[5] "Onu: Países chegam a acordo sobre nova agenda de desenvolvimento pós-2015," 2015, [Online; Acessado 20 de Julho 2018].

[6] "Global extreme poverty," 2017, [Online; Acessado 10 de Julho 2018]. [Online]. Available: https://ourworldindata.org/extreme-poverty

[7] R. Lazarotto, "Disitribuição de renda no brasil - uma análise pós-plano real," p. 83, 2009.

[8] "Operação censitária," 2018, [Online; Acessado 21 de Julho 2018]. [Online]. Available: https://censo2010.ibge.gov.br/materiais/guia-do-censo/operacao-censitaria.html

[9] J. Henderson, A. Storeygar, and D. Weil, "Measuring economic growth from outer space," *American Economic Review*, 2012.

[10] L. Torrey and J. Shavlik, "Transfer learning," p. 22, 2009.