Instituto Federal do Rio Grande do Norte
Campus Natal - Central
Diretoria de Gestão e Tecnologia da Informação
Tecnologia em Análise e Desenvolvimento de Sistemas

Explorando Regiões Densas Interessantes e discrepâncias em dados espaciais

Francisco Bento da Silva Júnior

Natal-RN
Dezembro, 2018

Francisco Bento da Silva Júnior

Explorando Regiões Densas Interessantes e discrepâncias em dados espaciais

Trabalho de conclusão de curso de graduação do curso de Tecnologia e Análise em Desenvolvimento de Sistemas da Diretoria de Gestão e Tecnologia de Informação do Instituto Federal do Rio Grande do Norte como requisito parcial para a obtenção do grau de Tecnologo em Análise e Desenvolvimento de Sistemas.

Banco de Dados:
Análise de Dados

Orientador

Dr. Plácido Antonio de Souza Neto

TADS – Curso de Tecnologia em Análise e Desenvolvimento de Sistemas
DIATINF – Diretoria Acadêmica de gestão e Tecnologia da Informação
CNAT – Campus Natal - Central
IFRN – Instituto Federal do Rio Grande do Norte

Natal-RN

Dezembro, 2018

Trabalho de Conclusão de Curso de Graduação sob o título *Título* apresentada por Nome completo do autor e aceita pelo Diretoria de Gestão e Tecnologia da Informação do Instituto Federal do Rio Grande do Norte, sendo aprovada por todos os membros da banca examinadora abaixo especificada:

Nome completo do orientador e titulação
Presidente
DIATINF – Diretoria Acadêmica de Gestão e Tecnologia da Informação
IFRN – Instituto Federal do Rio Grande do Norte

Nome completo do examinador e titulação
Examinador
Diretoria/Departamento
Instituto

Nome completo do examinador e titulação
Examinador
Diretoria/Departamento
Universidade

Natal-RN, data da defesa (dia, mês e ano).

Homenagem que o autor presta a uma ou mais pessoas.

Agradecimentos

Agradecimentos dirigidos àqueles que contribuíram de maneira relevante à elaboração do trabalho, sejam eles pessoas ou mesmo organizações.

Citação

Autor

Explorando Regiões Densas Interessantes e discrepâncias em dados espaciais

Autor: Francisco Bento da Silva Júnior
Orientador(a): Titulação e nome do(a) orientador(a)

Resumo

O resumo deve apresentar de forma concisa os pontos relevantes de um texto, fornecendo uma visão rápida e clara do conteúdo e das conclusões do trabalho. O texto, redigido na forma impessoal do verbo, é constituído de uma sequência de frases concisas e objetivas e não de uma simples enumeração de tópicos, não ultrapassando 500 palavras, seguido, logo abaixo, das palavras representativas do conteúdo do trabalho, isto é, palavras-chave e/ou descritores. Por fim, deve-se evitar, na redação do resumo, o uso de parágrafos (em geral resumos são escritos em parágrafo único), bem como de fórmulas, diagramas e símbolos, optando-se, quando necessário, pela transcrição na forma extensa, além de não incluir citações bibliográficas.

Palavras-chave: Palavra-chave 1, Palavra-chave 2, Palavra-chave 3.

Explorando Regiões Densas Interessantes e discrepâncias em dados espaciais

Author: Francisco Bento da Silva Júnior
Supervisor: Titulação e nome do(a) orientador(a)

Abstract

O resumo em língua estrangeira (em inglês *Abstract*, em espanhol *Resumen*, em francês *Résumé*) é uma versão do resumo escrito na língua vernícula para idioma de divulgação internacional. Ele deve apresentar as mesmas características do anterior (incluindo as mesmas palavras, isto é, seu conteúdo não deve diferir do resumo anterior), bem como ser seguido das palavras representativas do conteúdo do trabalho, isto é, palavras-chave e/ou descritores, na língua estrangeira. Embora a especificação abaixo considere o inglês como língua estrangeira (o mais comum), não fica impedido a adoção de outras linguas (a exemplo de espanhol ou francês) para redação do resumo em língua estrangeira.

Keywords: Keyword 1, Keyword 2, Keyword 3.

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

Sumário

1 Introdução

Nos últimos dez anos, a busca por termos como *big data*, análise de dados e visualização de dados tem aumentado notoriamente. Existem muitas razões para esse fenômeno, um deles é que com o avanço do poder computacional nós agora lidamos com imensas quantidades de dados, que crescem diariamente, de diversas fontes, em vários formatos e num incrível curto espaço de tempo. Dessa forma, novos desafios vêm surgindo na área de análise de dados: Como processar essas imensas quantidades rápida e eficientemente? Como visualizar esse montante de dados? Como *limpar* o conjunto de dados sem perder pontos importantes?

No que se refere ao campo da análise de dados e o analista está lidando com grandes *datasets*, é muito comum encontrar vários pontos com atributos muito distantes do resto de seu conjunto. Isto acontece porque quanto maior é o dataset, mais facilmente se pode encontrar pontos atípicos que serão mais distantes da distribuição normal. Esse comportamento é importante ser estudado pelo analista para que seja descoberto mais informações sobre o dataset em si e com isso possa ser tomado decisões mais precisas e proposto melhores afirmações. Geralmente, essa preocupação sobre dados anômalos não era tão relevante para a maioria das pesquisas, mas isso vem mudando a partir de que informações importantes podem ser descobertas com essa análise de pontos incomuns.

Esse tipo específico de dado com essas características é chamado de Outliers e é muito importante que os atuais analistas de dados deem mais atenção para esses dados, pois informações importantes podem estar escondidas entre esses conjuntos particulares.

Por exemplo, se pegarmos um conjunto de dados isolado sobre os fluxo de táxis de Nova Iorque de 2011 e analisarmos a frequência de corridas no ano inteiro, irão aparecer muitos pontos fora dessa curva média e isso indicaria um comportamento anômalo nessa coleção.

Geralmente, a primeira tarefa a se fazer nessa situação era remover esses pontos irregulares e continuar o processamento com o resto do conjunto, mas se pegarmos outro

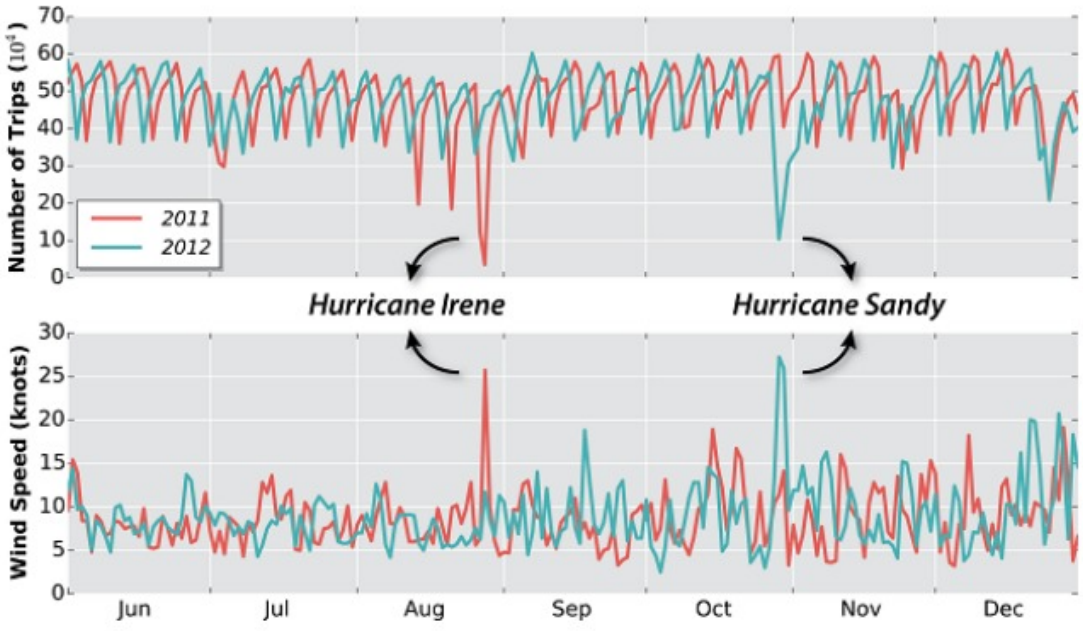

Figura 1: Figura tirada do (FREIRE et al., 2016) mostrando a relação entre o número de corridas de táxis e a velocidade do vento

conjunto de dados isolado sobre a velocidade do vento na região de Nova Iorque e no mesmo intervalo de tempo, nós iremos perceber algum picos de alta velocidade indicando furacões no mesmo momento e na mesma região da queda das corridas de táxi, como apresentado na Figura 1. Análises como essa provam a importância de detectar, estudar e interpretar esse outliers para acrescentar o conhecimento obtido de um dataset.

1.1 Contextualização

Hoje em dia nós estamos mais e mais conectados com múltiplas aplicações que acessam um imenso montante dos nossos dados existentes e ainda gera mais dados para melhorar suas análises sobre nós por diversos motivos. Ferramentas como Google Maps, Uber e Waze possuem muitos dados espaciais em tempo real sobre o nosso comportamento em relação ao tráfego (carros, transporte público, táxis, etc.), local de trabalho, locais de viagem frequentes, etc.

Quando tratamos usuários comuns, é muito comum que ele se perca frente à tanta massa de dados espaciais e isso vai prejudicar sua possível análise sobre o conjunto, mesmo a mais simples. Esse problema comum ainda não tem uma solução definitiva, então pesquisas atuais tentam indicar possíveis estratégias para mitigar esse problema e se aproximar de uma solução funcional. Essas abordagens são baseada em: agrupar um grande conjunto

de dados por um ou mais atributos específicos e resumir esses grupos baseado nesses atributos para conseguir simples *insights* sobre esses conjuntos, filtrar o dataset para reduzir os dados visíveis e focar em dados específicos para uma análise mais precisa (mas não vasta), e muitas outras estratégias para reduzir a complexidade da análise.

Junto desses problemas mais comuns, existe um importante que acontece antes do primeiro passo da análise que é: *O que fazer quando partes do dataset parecem irregular ou com dados corrompidos?*. Existem técnicas que ajudam na limpeza dessas partes de forma que não prejudique a análise, mas estudos recentes demonstram a importância desses dados *"anormais"* e o quanto um analista pode aprender estudando mais precisamente esse conjunto (FREIRE et al., 2016).

Nesse ambiente complexo de análise de dados espaciais com bastante variáveis e possibilidades, um usuário pode facilmente falhar numa dessas tarefas e comprometer seriamente o resultado de suas análises. Combinando todos esses detalhes, nós sugerimos uma abordagem que leve em consideração o *feedback* do usuário (capturando o movimento do mouse) e, baseado nesse feedback, nós nos tornaremos apto a analisar o interesse do usuário e dentro disso nós podemos detectar, estudar e propor ações para serem tomadas quando um dado considerado um outlier apareça nessa região de interesse do usuário.

1.2 Objetivos

Nesta seção estão definidos os objetivos gerais e específicos do trabalho.

1.2.1 Objetivos Gerais

- Introduzir o problema da análise e visualização em grandes conjuntos de dados espaço temporal atualmente.
- Explicar nossa abordagem proposta para detecção de outliers espaciais em grandes datasets utilizando o conceito de IDR e capturando o feedback do usuário.
- Apresentar nossos resultados utilizando a proposta para detecção de outliers no nosso ambiente espaço-temporal e os benefícios desse experimento.

1.2.2 Objetivos Específicos

- Analisar as pesquisas mais recentes na área de detecção de outliers em dados espaço-temporal.
- Apresentar nossa ferramenta proposta para análise e visualização de dados espaço-temporal.
- Comparar as pesquisas apresentadas destacando os prós e contras de cada pesquisa.
- Descrever o conceito de IDR utilizado na nossa ferramenta para mapear a preferência do usuário em um ambiente espaço-temporal.
- Resumir os algoritmos de detecção de outliers mais conhecidos para dados genéricos e espaciais.
- Mostrar o nosso algoritmo escolhido explicando as razões dessa escolha.
- Aplicar nosso conceito de IDR e nosso algoritmo de detecção de outlier escolhido num ambiente de dados espaço-temporal.
- Apresentar os resultados de nossa aplicação e indicar nossos trabalhos futuros.

1.3 Organização do Trabalho

O documento é organizado do seguinte modo: Seção 2 resume as pesquisas existentes no campo da análise e visualização de dados comparando com nossa ferramenta proposta. Seção 3 descreve os conceitos de IDRs (*Interesting Dense Region*) e Outliers com os algoritmos existentes para sua detecção e nosso algoritmo escolhido para detectar outliers em nossa plataforma. Seção 4 explica como aplicamos as IDRs e detecção de outliers na ferramenta GeoGuide. Seção 5 apresenta duas aplicações utilizando datasets distintos para demonstrar a aplicabilidade disso em problemas do mundo real e as vantagens dessa abordagem, mostrando e discutindo os resultados da detecção de outliers. Por fim, a conclusão e algumas direções para trabalhos futuros são dados na Seção 6.

2 Background

Neste capítulo nós vamos apresentar os trabalhos existentes sobre análise de dados e detecção de outliers com algoritmos e estratégias de como usar essa abordagem para melhorar a análise dos dados e aumentar a quantidade e a qualidade das informações que podemos obter dos datasets.

2.1 Trabalhos Relacionados

2.1.1 GeoGuide

Visando melhorar a análise de dados espaciais e a abordagem por orientação para esse tipo de dado, o GeoGuide (OMIDVAR-TEHRANI et al., 2017) é um framework interativo que visa destacar para o analista um subconjunto de k pontos espaciais interessantes, baseado nos feedbacks *implícito* (ex.: rastreamento do mouse) e *explícito* (ex.: pontos clicados) do analista, que podem não terem sido vistos dado o montante de informação aparente na sua tela. Esse framework leva em consideração duas importantes métricas para poder destacar um subconjunto. A primeira é a **relevância** de cada ponto para o ponto selecionado pelo analista considerando os atributos não espaciais desses pontos. O segundo é a **diversidade** geográfica para que assim possa expandir a área de análise do usuário em busca de possíveis novas regiões de seu interesse.

Todo esse processo pode ser utilizado em datasets espaciais genéricos, contanto que cada ponto do conjunto tenha duas características: atributos geográficos (ex.: latitude e longitude) e atributos *metadados* de domínio do dataset. Por exemplo, a plataforma Airbnb[1] tem datasets abertos sobre as casa disponíveis para alugar e cada uma delas tem atributos geográficos e *preço, nome do hospedados e disponibilidade* como seus atributos metadados que são específicos para cada tipo de dataset como podemos ver na Figura 2. Utilizando essa abordagem, o GeoGuide é o primeiro framework interativo eficiente

[1] *http://www.airbnb.com*

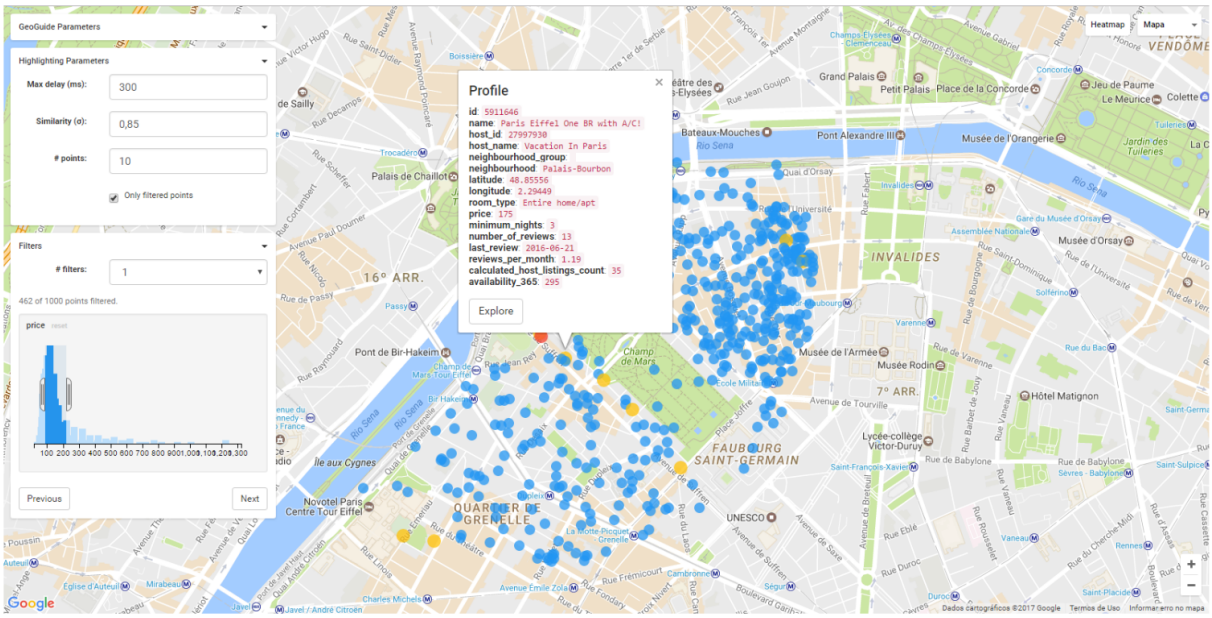

Figura 2: Imagem do GeoGuide no dataset do Airbnb - Cidade de Pais

para destaque de dados espaciais, combinando o feedback do analista com as métricas de relevância e diversidade para mostrar a ele um subconjunto de pontos interessantes que podem não ter sido deixado de lado durante sua análise.

2.1.2 Outliers

No campo da estatística, Outliers são quando se encontram "valores aberrantes" num determinado conjunto de dados, ou seja, quando alguém acha um valor atípico ou muito fora da distribuição normal daquele conjunto. Por exemplo, quando um pesquisador quer monitorar a temperatura de sua CPU durante um certo intervalo de tempo e foi percebido que a variação de temperatura foi entre 34 °C e 45 °C com máxima de 48 °C e mínima de 27 °C e no meio dessa amostragem são encontrados registros pontuais de 0 °C, isso é caracterizado como um outlier a, muito provavelmente, será interpretado como um defeito do equipamento que realizou a coleta da temperatura da CPU.

Entretanto, existem diversas formas de interpretar um Outlier além de um erro da coleta, como: um dado que pertença à uma população diferente da amostra, um dado defeituoso, um dado que esteja numa área em que certa teoria não é válida ou até, quando a amostra é muito grande, é normal haver pequenos quantidades de outliers naquele grupo. Em casos em que são provados que não é culpa de um equipamento defeituoso de coleta ou que não foi uma falha humana, é extremamente importante entender o porquê daquele outlier, pois não é interessante para o pesquisador simplesmente removê-lo ou ressignificá-lo definindo-lhe um novo valor, já que essa mudança pode comprometer a validade da

pesquisa e, caso aconteça, é de extrema importância que tudo isso seja documentado para registro dessas alterações.

Assim como as tecnologias da informação melhoram e aumentam continuamente seu poder computacional, uma grande variedade de algoritmos para detecção de outliers tem surgido e vem sendo aplicado em diferentes contextos com diversas características, sendo que a escolha para um desses algoritmos é baseada no domínio do problema. Na próxima seção nós apresentaremos alguns desses algoritmos com uma breve explicação sobre cada um.

2.2 Algoritmos

2.2.1 Z-Score

Z-Score é um dos métodos mais simples para detecção de outliers em um dataset. É um método parametrico que leva em consideração somente um atributo por execução e é necessário a entrada de um valor limite (geralmente um valor entre 2,5 e 3,5) para poder definir se um determinado dado pode ser considerado como um outlier ou não. Esse método é adequado para pequenos datasets que seguem a distribuição Gaussiana.

2.2.2 DBSCAN

É um algoritmo de clusterização de dados espaciais baseado em densidade (ESTER et al., 1996) pode ser aplicado em datasets os quais não se pode presumir qual a sua distribuição e que aceita datasets multidimensionais (com 3 ou mais dimensões). Entretanto, é necessário a entrada de um parametro (MinPts) que definirá quantos pontos são minimamente necessários para se formar um *cluster*. Portanto, se o tamanho do conjunto mudar, esse parâmetro terá que ser atualizado, caso contrário, o DBSCAN pode se tornar ineficiente.

2.2.3 Isolation Forests

É um algoritmo de detecção de outliers (LIU; TING; ZHOU, 2008) que usa um dos conceitos de aprendizagem de máquina que são as chamadas árvores de decisão. É unidimensional (só leva em consideração um atributo por vez) e são necessários poucos parâmetros (isso facilita a configuração e uso do algoritmo). Por não precisar escalar seus valores para

sua execução, isso o torna um algoritmo robusto para grandes datasets.

2.2.4 FDC

É uma abordagem de algoritmo baseado em profundidade (JOHNSON; KWOK; NG, 1998) para detecção de outliers em datasets 2D que se baseia no conceito do algoritmo ISODEPTH. O FDC computa os primeiros k contornos de profundidade (os pontos que podem ser considerados outliers) restrigindo a uma pequena parte do dataset completo. Dessa forma, é mais eficiente por não ter que calcular no dataset completo e, portanto, mais fácil de escalar para grandes datasets de duas dimensões.

2.2.5 HOD

É um método de detecção de outlier baseado em distância (XU et al., 2016) que surge para superar os métodos baseados em estatísticas já que na vasta maioria dos datasets a distribuição de probabilidade não é conhecida. Dessa forma, o método busca por outliers baseado na sua distância em relação aos seus vizinhos e se essa distância é maior que um parâmetro de entrada predefinido, então esse ponto é considerado um outlier. Entretanto, se já existe um cluster de outliers no dataset, isso pode afetar a detecção em algoritmos baseado em distância, para isso que serve o conceito HOD (*Hidden Outlier Detection*) do algoritmo que visa encontrar outliers mesmo quando eles estão agrupados em quantidades suficiente para formação de um cluster.

2.2.6 ORCA

É um algoritmo baseado em distância (BAY; SCHWABACHER, 2003) que otimiza um algoritmo simples de loop aninhado (que são algoritmos de complexidade exponencial os quais são extremamente ineficientes quando utilizados em grandes datasets) pela remoção de possíveis *non-outliers* durante sua execução. Dessa forma, ao invés de processar o dataset por completo, ele vai removendo cálculos que serão desnecessários, já que o ponto não vai ser considerado um outlier. A partir dele, novas pesquisas têm surgido para refinar mais esse conceito.

2.2.7 Linearization

É um algoritmo baseado em distância (ANGIULLI; PIZZUTI, 2002) que detecta outliers calculando a soma das distâncias de um ponto em relação aos seus vizinhos, chamando isso de *peso*, e definindo os outliers como os pontos com os maiores pesos no dataset. Dessa forma, é um algoritmo eficiente de complexidade linear tanto em relação ao número de pontos como ao número de dimensões. Para ajudar a calcular esses outliers mais eficientemente, o algoritmo utiliza o conceito da *curva de Hilbert*

2.2.8 RBRP

É um algoritmo de alta perfomance para datasets multidimensionais que é baseado nas distâncias entre os pontos para poder definir quais são os outliers no dataset (GHOTING; PARTHASARATHY; OTEY, 2006). Sua diferença para os outros algoritmos baseado em distância é que ele se torna mais eficiente quando executado em dataset com múltiplas dimensões, sua escalabilidade é aproximadamente linear para o número de dimensões e logarítmica para o número de pontos no dataset.

2.2.9 LOF

É um algoritmo baseado em densidade que adiciona um novo conceito na busca por outliers: o *Local Outlier Factor (LOF)* (BREUNIG et al., 2000), que é um grau de propensidade que aquele ponto tem de ser um outlier e dessa forma o processo de definição de outlier não é mais binário, mas sim algo gradual. Com isto, a abordagem não é mais sobre um ponto ser outlier ou não, mas sim o "quão outlier" esse ponto é em relação ao dataset. O *outlier factor* é local no sentido que somente os vizinhos daquele ponto é levado em consideração para definir seu fator.

2.2.10 ABOD

É um algoritmo de detecção de outlier baseado em ângulo (KRIEGEL; HUBERT; ZIMEK, 2008) que é focado em datasets de altas dimensões, diferente de outros algoritmos baseado em distãncia que acabam prejudicados quando o dataset tem muitas dimensões. Sua abordagem é baseado no cálculo do grau do ângulo entre os diferentes vetores de um ponto com os seus vizinhos. Com isto, pontos mais centralizados dentro do cluster terão esse grau calculado com um valor maior, já os pontos mais próximos da borda terão esse

Algorithms	**I**	**II**	**III**	**IV**	**V**
Z-Score	Sim	Baseado em Modelo	Não	Não	Sim
DBSCAN	Não	Baseado em Densidade	Não	Sim	Sim
Isolation Forests	Não	Baseado em Profundidade	Não	Não	Não
FDC	Não	Baseado em Profundidade	Sim	Não	Não
Hidden Outlier Detection	Não	Baseado em Distância	Sim	Sim	Sim
ORCA	Não	Baseado em Distância	Não	Sim	Sim
Linearization	Não	Baseado em Distância	Sim	Sim	Sim
RBRP	Não	Baseado em Distância	Sim	Sim	Sim
LOF	Não	Baseado em Densidade	Não	Sim	Sim
ABOD	Não	Altas Dimensões	Não	Sim	Não

Tabela 1: Comparação dos Algoritmos de Detecção de Outlier apresentados

valor um pouco menor e o possíveis outliers terão esse grau com valores muito pequenos, já que eles geralmente estarão distantes do cluster em uma direção particular.

Baseado nos algoritmos apresentados, nós organizamos cada um de acordo com a resposta para nossas questões propostas sobre algoritmos de detecção de outliers no geral. As questões são: **I** *É paramétrico?*; **II** *Qual é a abordagem?*; **III** *É escalável em termos de performance?*; **IV** *É escalável em termos de múltiplas dimensões?* e **V** *Ele recebe algum argumento?*. O resultado é apresentado na Tabela Table 1.

Cada questão tem uma motivação específica para estar na Tabela 1. A questão **I** é sobre a distribuição de probabilidade do dataset. Se o algoritmo é paramétrico, então nós podemos assumir a distribuição de probabilidade do dataset baseado em um conjunto fixo de parâmetros. A questão **II** serve para classificar cada algoritmo baseado na abordagem que ele utiliza para indicar se um dado é um outlier ou não, as opções são: textitBaseado em Modelo, *Baseado em Densidade*, *Baseado em Profundidade*, *Baseado em Distância* e *Altas Dimensões*. A questão **III** é relativa a performance computacional de cada algoritmo, se o tempo de execução do algoritmo não é comprometido de acordo com o crescimento do dataset, significa que ele é escalável em termos de performance. A questão **IV** é sobre a performance de cada algoritmo também, mas nesse caso é relacionado ao crescimento de dimensões do dataset. Por fim, a questão **V** indica se o algoritmo requer algum argumento de entrada para processar o dataset, isso é importante pois se o algoritmo requer muitos argumentos pode se tornar um problema de precisão quando o dataset aumentar.

3 Outliers

3.1 Outliers

Nesta seção nós iremos abordar mais detalhadamente as características dos pontos que são considerados *discrepantes* em um determinado conjunto de dados e quais as influências que isso pode trazer para uma pesquisa. Além disso, vamos falar do tratamento que propomos para o GeoGuide e qual a utilidade disso e também sobre o algoritmo escolhido e as motivações que levaram à essa escolha.

3.1.1 Dados espaciais discrepantes

Quando tratamos de temas como a captura e análise de dados no nosso cotidiano, o cenário atual se encontra numa situação nunca antes vista na história, já que nunca tivemos tanto volume de dados sendo gerado a todo segundo e das mais diversas formas e fontes. Caso a gente se aprofunde ainda mais nesse meio e foque nos dados espaciais, encontraremos uma variedade de características específicas desse nicho e também problemas específicos os quais buscamos solução.

Por exemplo, podemos destacar aqui 3 exemplos distintos de datasets que estão disponíveis atualmente na internet e são referentes a dados espaciais, cada um com sua particularidade:

3.1.1.1 Táxis de Nova Iorque

Desde de 2011 que as empresas de táxis de Nova Iorque, visando entender melhor seu funcionamento, conhecer mais sobre seu mercado e evitar as ocorrências de fraudes, investem em uma estrutura para coleta e armazenamento dos dados referentes a cada corrida de táxi que acontece na cidade. O resultado disso é que anualmente são armazenados e disponibilizados gigabytes de dados referentes a cada uma dessas corridas e que podem ser utilizados para análises de qualquer natureza e finalidade.

Entretanto, não é tão simples o processo de análise de volumes desse porte e antes de mesmo dessa análise começar, vários fatores tem que serem levados em consideração, como, por exemplo, a estrutura desses dados e a confiabilidade dele. No caso desse dataset, falhas podem ocorrer nos equipamentos responsáveis pela coleta, na próprio gerenciamento e organização desses dados, pois nesse processo necessita de intervenção humana e isso por si só é um fator de risco em qualquer análise.

Imagine que num determinado táxi em um certo dia e horário, todas as corridas tiveram o dado referente a distância percorrida pelo táxi com um valor negativo, o que fazer nessa situação? Ou então que a data iníc

io de uma corrida seja superior à data de término da mesma? Ou até sejam valores próximos, mas sua distância seja extremamente alta para esse tempo? Todos esses problemas podem acontecer nesse contexto e cada um deles devem ser tratados propriamente e com o devido cuidado. Esses tipos de problemas são os problemas no processo de *limpeza dos dados* e é de uma importância crucial que eles sejam mitigados, pois nessas circustâncias as futuras análises estariam todas comprometidas com a presença de *ruídos* que afetam diretamente uma boa análise.

3.1.1.2 Hospedagens de Paris

De maneira similar aos datasets sobre viagens de táxis, existem também grandes volumes de dados espaciais sobre hospedagens ao redor do mundo disponíveis na internet através da plataforma Airbnb, que é um serviço online para divulgações de hospedagens, com um diferencial, pois os lugares são divulgados por pessoas comuns e que, geralmente, vivem no local que querem divulgar.

Esses datasets podem ser utilizados para as mais diversas análises e com os mais diferentes propósitos. A sua estrutura consiste em dados espaciais (latitude e longitude) referente à hospedagem e os dados sobre a própria localidade em si, como por exemplo: o nome do anfitrião, o preço diário daquele local, a quantidade mínima de noites para poder alugar aquele lugar, a quantidade de dias que aquele imóvel está disponível ano e etc.

Por exemplo, existe uma plataforma online que utilizou esses datasets para entender mais sobre esse mercado de hospedagens e se realmente, como proposto pela empresa Airbnb, esse novo mercado é uma alternativa as industrias de hotéis que existem pelo mundo. Quando analisado os dados e os volumes são agrupados pelas cidades ao redor do mundo, se percebe que esse novo mercado não é tão disruptível assim e que, na realidade, a

maioria dos locais disponíveis para hospedagens são imóveis completos, gerando assim uma nova indústria mais moderna de hospedagens. Mais informações sobre essa pesquisa podem ser encontradas na plataforma disponível online no link <http://insideairbnb.com/>.

3.1.1.3 Restaurantes em Las Vegas

Por fim, como mais um exemplo, existe outro conjunto de datasets espaciais referentes a restaurantes disponíveis em várias regiões pelo mundo. Esse conjunto é mantido pela empresa Yelp [1] que oferece um serviço de pesquisa de restaurantes de forma personalizada que pode levar em consideração vários argumentos de acordo com sua necessidade, por exemplo: o local em que você queira encontrar os restaurantes próximos, o tipo de restaurante, a faixa de preço do seu interesse, enfim, uma porção de possibilidades para sua consulta.

Com essa estrutura é possível realizar diversas análises focadas em soluções específicas e que possam contribuir para contextos mais complexos. Essas possibilidades são tão frequentes que existe um desafio da própria empresa Yelp que convida estudantes de cursos superiores para submeterem propostas de análises desses conjuntos de dados voltadas para um tema específico e que acabem resolvendo problemas da própria empresa. As informações mais detalhadas desse desafio podem ser encontradas nesse site da Yelp <https://www.yelp.com/dataset/challenge>.

Incentivos como esse mostram o quão importante é o investimento nas áreas de análises de dados espaciais e o quanto ainda se pode aprimorar nessas análises para trazer informações mais úteis sobre os datasets e o que se pode fazer com esses determinados conjuntos de dados. Isso é uma das provas que essas análises podem ser tarefas complexas e que precisa-se de conhecimento mais qualificado para se conseguir melhores resultados,

3.1.2 Outliers no GeoGuide

Como dito anteriormente, o GeoGuide visa melhorar a experiência do analista frente à grandes montantes de dados espaciais que possam o deixar perdido durante sua análise, e isso, na maioria dos casos, prejudicará sua análise fazendo com que ele não leve em informações que podem ser relevantes para sua pesquisa. Além disso, a existência de Outliers num dataset pode ser uma fonte de novas informações sobre aquele conjunto de dados. Por exemplo, em um conjunto de dados referente aos gastos do cartão de crédito de

[1] https://www.yelp.com/

uma pessoa, caso tenha outliers, é um indicador muito forte de uma fraude nesse cartão, com isso, as empresas financeiras podem tomar medidas de segurança para evitar esses problemas.

No caso de datasets espaciais, as aplicações podem ser ainda maiores, pois além dos atributos referentes aos próprios pontos do conjunto, os próprios atributos referentes a localização do ponto (latitude e longitude) podem, por si só, caracterizar determinado ponto como um outlier. Dessa forma, a identificação desses pontos durante a análise do pesquisador pode trazer novas informações que estimule a criação de novas abordagens e decisões sobre aquele conjunto.

A título de exemplo, se pegarmos o dataset referente aos táxis de Nova Iorque e selecionarmos um conjunto de corridas de uma parte de Manhattan do mesmo dia e mesmo horário e com uma margem próxima de preço entre elas, podemos buscar por pontos considerados outliers em relação ao seu preço e o do conjunto, com isso, podemos detectar possíveis fraudes no serviço de táxi (caso esse ponto seja um outlier por um preço "abaixo do normal") ou então abuso de cobrança por parte do motorista (caso esse ponto seja um outlier por um preço "acima do normal").

Sabendo disso, podemos aplicar essa detecção de outliers no GeoGuide enquanto o analista está no processo de descobrimento do dataset. Dessa forma, enquanto ele vai percorrendo pelo conjunto de dados e selecionando quais os pontos são de seu interesse, a ferramenta vai procurando, de forma assíncrona e implícita, pontos que podem ser outliers em relação aos pontos já previamentes definidos como interessantes para o usuário e assim, de forma totalmente transparente e sem precisar de nenhuma nova ação do analista, novas informações sobre o dataset podem ser destacadas para quem está analisando e abrir um leque maior de opções e ideias sobre aquele conjunto de dados, ampliando sua análise mais facilmente e lhe mostrando o quê mais pode ser importante naquele ambiente.

3.1.3 Algoritmo

TODO: Mostrar o algoritmo de detecção selecionado. Usar uma imagem dele e explicar passo a passo, além de falar sobre sua complexidade, mostrando os prós e contras do seu uso.

4 IDRs

4.1 IDRs

Nesta seção nós iremos explanar mais profundamente o conceito de Regiões Densamente Interessantes e qual a sua utilidade em uma análise de dados espaciais. Também detalharemos o passo a passo para a criação de uma Região desse tipo, qual a função do feedback do usuário nessa construção, qual o algoritmo usado para esse processo, qual sua utilização no GeoGuide e quais as vantagens que podemos obter quando utilizamos Regiões Densamente Interessantes para fazer análise de dados espaciais com uma abordagem de orientação do usuário através da ferramenta.

4.1.1 Feedback do usuário

Para poder explicar melhor o conceito de IDR, precisamos começar falando sobre o *feedback* do usuário, como funciona e como isso pode ser utilizado na criação das Regiões. Durante a utilização de um sistema, o usuário vai interagindo com suas funcionalidades e o sistema vai respondendo aos seus comandos e ações, isso faz com que um sistema seja interativo e dinâmico, podendo ser atualizado conforme o usuário e as características de cada um. Em sistemas de análise de dados espaciais, é muito comum que seja utilizado um mapa, um gráfico ou qualquer recuso visual que facilite a interpretação do usuário e dê uma noção sobre o que se trata o dataset em si.

A partir disso, o usuário pode ir "caminhando" pelo mapa (ou figura) para ir conhecendo mais afundo os dados e as particularidades de cada ponto. Esse "caminhar" pode se tornar interessante para o sistema de forma que isso ajude a conhecer os interesses do usuário. Essa resposta que o usuário concede ao sistema enquanto o estar utlizando, é o que caracteriza o *feedback* e isso pode trazer muitas utilizações para diversos tipos de sistemas.

O feedback pode ser dividido em duas categorias sobre a abordagem utilizada para

sua coleta: o *explícito* e o *implícito*. O primeiro se refere a quando o usuário define como interessante de forma consciente utilizando meios que o próprio sistema indica como funciona e para que serve. Por exemplo: quando o usuário indica se gostou de determinada indicação, quando ele vota de 0 a 5 estrelas num filme de sua preferência, quando ele indica algum restaurante para alguém utilizando o sistema, quando ele clica num ponto no mapa para obter mais informações e de várias outras formas pode se obter um feedback explícito.

Já no caso da segunda categoria de feedback, o usuário não precisa dizer diretamente qual o seu interesse no sistema, do contrário, o sistema vai detectando progressivamente o que o usuário vai fazendo na plataforma e vai registrando isso para, a partir de uma determinada quantidade de informação, conseguir caracterizar algo como interessante para o usuário. Por exemplo, o sistema pode rastrear o movimento do mouse ou dos olhos para conseguir definir onde o usuário mais foca no sistema durante sua utilização, para isso ele precisa registrar os pontos rastreados do usuário e ir salvando isso. Então, com uma grande coleção de pontos, o sistema pode utilizar algoritmos de clusterização para encontrar as áreas mais marcantes desse conjunto e classificar elas como de interesse do usuário.

Esse tipo de feedback implícito é utilizado no GeoGuide para a captura de interesse do usuário através do rastreamento do mouse e utilizando algoritmos de clusterização para construir essas áreas que consideramos de interesse do usuário durante sua exploração pelo mapa da plataforma. Tudo isso de forma que o usuário não saiba e não precisa gastar nenhum esforço para indicar o que lhe é interessante, somente usando a plataforma já podemos detectar essa informação.

4.1.2 Regiões Densamente Interessantes

Como dito anteriormente, o feedback que o usuário entrega ao sistema é de extremo valor para o aprimoramento da ferramenta e das análises que ela performar. Isso faz com que a ferramenta tenha a característica de se adequar a cada usuário e trazer o melhor resultado para cada um em específico. Também foi abordado que no GeoGuide utilizamos o feedback implícito do rastreamento do mouse para detectar as áreas de interesse do usuário.

Entretanto, adicionamos o conceito de IDR para ir mais além nessa detecção e aprimorar a análise do usuário reforçando alguns aspectos que somente com áreas isoladas não iríamos conseguir. Esse conceito de Regiões Densamente Interessantes é baseado na detecção de áreas de interesse do usuário construídas a partir do rastreamento do seu

mouse.

No GeoGuide, a partir da coleta dos pontos de rastreamento do mouse, registramos isso periodicamente e clusterizamos esses conjuntos para a formação das áreas de interesse. Cada área é representada por um conjunto de pontos agrupados por sua proximidade em relação de um aos outros. A partir desse conjunto, construímos um polígono convexo utilizando os pontos mais distantes do centro. Cada polígono é a representação da área de interesse do usuário. Isto posto, para a formação de uma IDR, nós dividimos uma quantidade de momentos de coleta dos polígonos e definimos como IDR a interseção desses polígonos entre si. Ou seja, na nossa ferramenta, decidimos que a cada 20 segundos de rastreamento do mouse, o sistema vai construir uma pequena série de polígonos formados a partir do cluster desses pontos de rastreio e registrar para o próximo passo. A cada 3 momentos de construção de clusters, nós selecionamos todos os polígonos resultantes e calculamos novos polígonos formados pela interseção das áreas de cada momento.

O resultado dessas interseções são o que chamamos de Regiões Densamente Interessantes, e isso torna a formação da área de interesse do usuário mais reforçada e precisa, pois cada IDR vai demonstrar uma região que o usuário focou em mais de um momento em tempos diferentes, tornando aquela área interessante para o usuário e podendo ser utilizada tanto para um foco mais preciso naquela região, caso queira restringir mais o escopo da análise, quanto para diminuir as sugestões naquela região e assim abrir mais o leque de opções do analista e expandir sua área de pesquisa, incrementando novas possibilidades para sua análise.

4.1.3 Algoritmo

TODO: Mostrar o algoritmo utilizado para detecção do IDR e explicar seu funcionamento. Apresentar sua complexidade com seus prós e contras.

5 Considerações finais

As considerações finais formam a parte final (fechamento) do texto, sendo dito de forma resumida (1) o que foi desenvolvido no presente trabalho e quais os resultados do mesmo, (2) o que se pôde concluir após o desenvolvimento bem como as principais contribuições do trabalho, e (3) perspectivas para o desenvolvimento de trabalhos futuros, como listado nos exemplos de seção abaixo. O texto referente às considerações finais do autor deve salientar a extensão e os resultados da contribuição do trabalho e os argumentos utilizados estar baseados em dados comprovados e fundamentados nos resultados e na discussão do texto, contendo deduções lógicas correspondentes aos objetivos do trabalho, propostos inicialmente.

5.1 Principais contribuições

Texto.

5.2 Limitações

Texto.

5.3 Trabalhos futuros

Texto.

Referências

ANGIULLI, F.; PIZZUTI, C. Fast outlier detection in high dimensional spaces. In: ELOMAA, T.; MANNILA, H.; TOIVONEN, H. (Ed.). *Principles of Data Mining and Knowledge Discovery*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002. p. 15–27. ISBN 978-3-540-45681-0.

BAY, S. D.; SCHWABACHER, M. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In: *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2003. (KDD '03), p. 29–38. ISBN 1-58113-737-0. Disponível em: <http://doi.acm.org/10.1145/956750.956758>.

BREUNIG, M. M. et al. Lof: Identifying density-based local outliers. *SIGMOD Rec.*, ACM, New York, NY, USA, v. 29, n. 2, p. 93–104, maio 2000. ISSN 0163-5808. Disponível em: <http://doi.acm.org/10.1145/335191.335388>.

ESTER, M. et al. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. AAAI Press, 1996. (KDD'96), p. 226–231. Disponível em: <http://dl.acm.org/citation.cfm?id=3001460.3001507>.

FREIRE, J. et al. Exploring what not to clean in urban data: A study using new york city taxi trips. *IEEE Data Eng. Bull.*, v. 39, n. 2, p. 63–77, 2016. Disponível em: <http://sites.computer.org/debull/A16june/p63.pdf>.

GHOTING, A.; PARTHASARATHY, S.; OTEY, M. E. Fast mining of distance-based outliers in high-dimensional datasets. In: *Proceedings of the 2006 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, 2006. p. 609–613. Disponível em: <https://doi.org/10.1137/1.9781611972764.70>.

JOHNSON, T.; KWOK, I.; NG, R. Fast computation of 2-dimensional depth contours. In: *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*. AAAI Press, 1998. (KDD'98), p. 224–228. Disponível em: <http://dl.acm.org/citation.cfm?id=3000292.3000332>.

KRIEGEL, H.-P.; HUBERT, M. S.; ZIMEK, A. Angle-based outlier detection in high-dimensional data. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2008. (KDD '08), p. 444–452. ISBN 978-1-60558-193-4. Disponível em: <http://doi.acm.org/10.1145/1401890.1401946>.

LIU, F. T.; TING, K. M.; ZHOU, Z. Isolation forest. In: *2008 Eighth IEEE International Conference on Data Mining*. [S.l.: s.n.], 2008. p. 413–422. ISSN 1550-4786.

OMIDVAR-TEHRANI, B. et al. Geoguide: An interactive guidance approach for spatial data. In: *2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), Exeter, United Kingdom, June 21-23, 2017*. [s.n.], 2017. p. 1112–1117. Disponível em: <https://doi.org/10.1109/iThings-GreenCom-CPSCom-SmartData.2017.170>.

XU, H. et al. Index based hidden outlier detection in metric space. *Scientific Programming*, Hindawi Limited, v. 2016, p. 1–14, 2016. Disponível em: <https://doi.org/10.1155/2016/8048246>.

APÊNDICE A – Primeiro apêndice

Os apêndices são textos ou documentos elaborados pelo autor, a fim de complementar sua argumentação, sem prejuízo da unidade nuclear do trabalho.

ANEXO A – Primeiro anexo

Os anexos são textos ou documentos não elaborado pelo autor, que servem de fundamentação, comprovação e ilustração.