

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 de Denso Interesse e Outliers em um ambiente de dados espaço-temporal

Francisco Bento da Silva Júnior

Natal-RN
Dezembro, 2018

Francisco Bento da Silva Júnior

Explorando Regiões de Denso Interesse e Outliers em um ambiente de dados espaço-temporal

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 Tecnólogo 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 de Denso Interesse e Outliers em um ambiente de dados espaço-temporal

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 de Denso Interesse e Outliers em um ambiente de dados espaço-temporal

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

- 1 Figura tirada do (FREIRE et al., 2016) mostrando a relação entre o número
de corridas de táxis e a velocidade do vento p. 14
- 2 Imagem do GeoGuide no dataset do Airbnb - Cidade de Pais p. 18

Lista de tabelas

1	Comparação dos Algoritmos de Detecção de Outlier apresentados . . .	p. 22
---	---	-------

Lista de abreviaturas e siglas

Sumário

1	Introdução	p. 13
1.1	Contextualização	p. 14
1.2	Objetivos	p. 15
1.2.1	Objetivos Gerais	p. 15
1.2.2	Objetivos Específicos	p. 16
1.3	Organização do Trabalho	p. 16
2	Background	p. 17
2.1	Trabalhos Relacionados	p. 17
2.1.1	GeoGuide	p. 17
2.1.2	Outliers	p. 18
2.2	Algorithms	p. 19
2.2.1	Z-Score	p. 19
2.2.2	DBSCAN	p. 19
2.2.3	Isolation Forests	p. 19
2.2.4	FDC	p. 20
2.2.5	HOD	p. 20
2.2.6	ORCA	p. 20
2.2.7	Linearization	p. 20
2.2.8	RBRP	p. 21
2.2.9	LOF	p. 21
2.2.10	ABOD	p. 21

3	IDRs and Outliers	p. 23
3.1	IDRs	p. 23
3.1.1	User feedback	p. 23
3.1.2	Interesting Dense Regions	p. 23
3.1.3	Algorithm	p. 23
3.2	Outliers	p. 23
3.2.1	Atypical spatial data	p. 23
3.2.2	Outliers in GeoGuide	p. 24
3.2.3	Algorithm	p. 24
4	Considerações finais	p. 25
4.1	Principais contribuições	p. 25
4.2	Limitações	p. 25
4.3	Trabalhos futuros	p. 25
	Referências	p. 26
	Apêndice A – Primeiro apêndice	p. 28
	Anexo A – Primeiro anexo	p. 29

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 dêem 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 fluxos de taxis 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

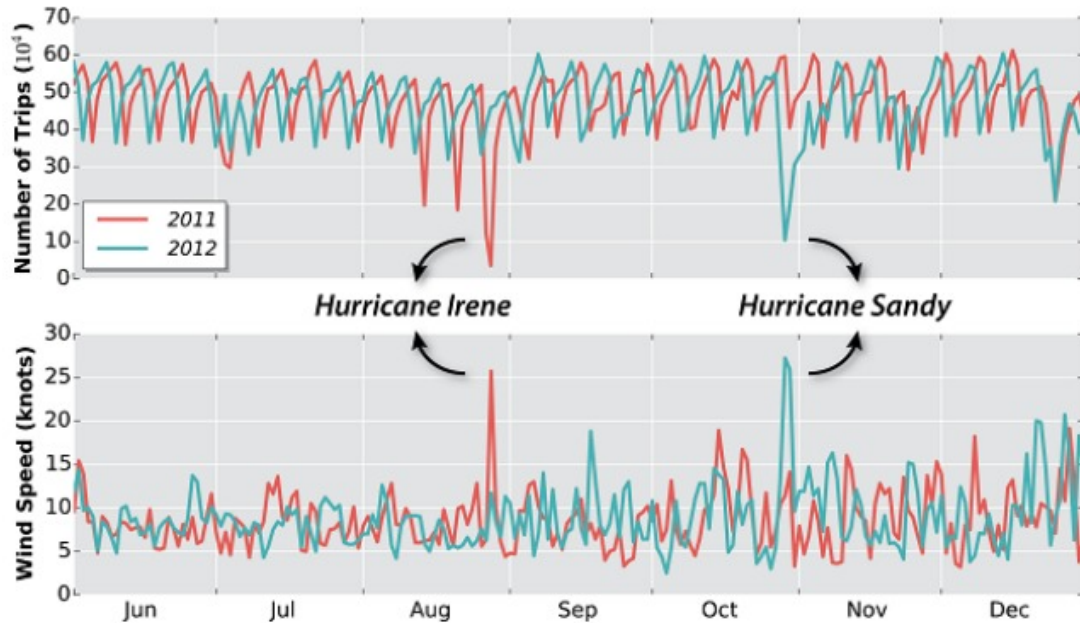


Figura 1: Figura tirada do (FREIRE et al., 2016) mostrando a relação entre o número de corridas de táxis e a velocidade do vento

conjunto de dados isolado sobre a velocidade do vento na região de Nova Iorque e no mesmo intervalo de tempo, nós iremos perceber alguns 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 esses 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, taxis, 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 baseadas 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

Nesse 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 Airbnb¹ 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>

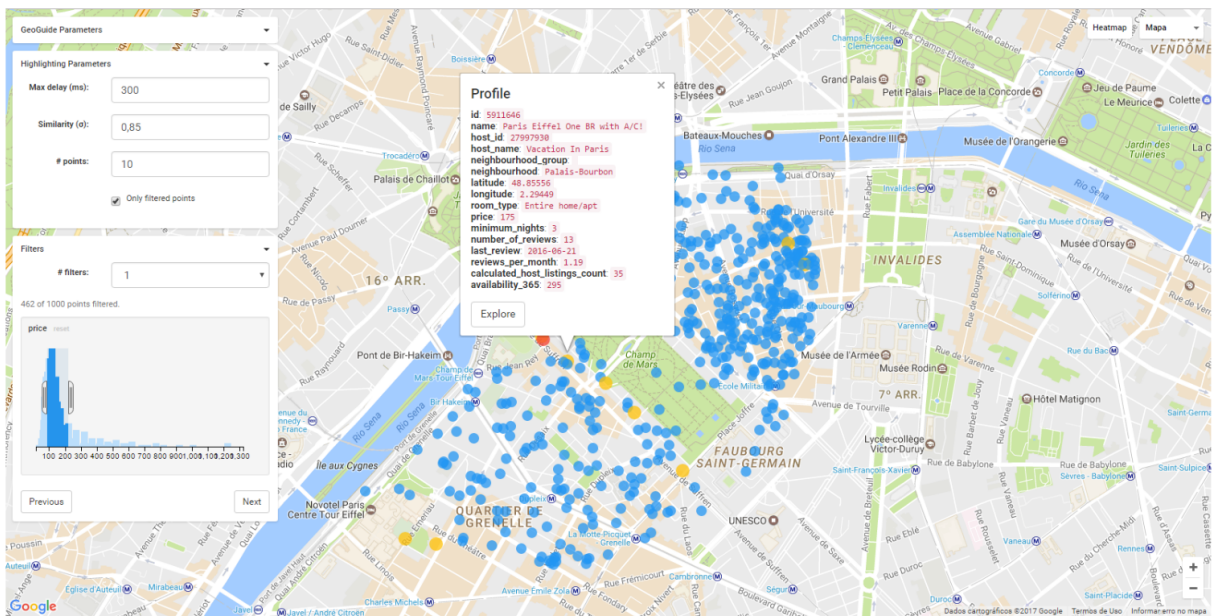


Figura 2: Imagem do GeoGuide no dataset do Airbnb - Cidade de Paris

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 pequenas 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 Algorithms

2.2.1 Z-Score

Z-Score is one of the simplest methods for detecting outliers in a dataset. It is a parametric method and takes into account only one attribute per execution. It is also necessary the input of a threshold (usually is a value between 2.5 to 3.5) to be able to define if a given data can be considered an outlier or not. This method is suitable for small datasets that follow the Gaussian distribution.

2.2.2 DBSCAN

It is a density-based spatial clustering algorithm (ESTER et al., 1996) that can be applied in datasets that cannot be presumed what their distribution. It accepts multidimensional datasets (with 3 or more dimensions). However, you need a parameter (MinPts) that defines how many minimum points are needed to form a cluster. Thus, if the size of the set change, this parameter will need to be updated, otherwise the DBSCAN can become inefficient.

2.2.3 Isolation Forests

It is an algorithm of detection of outliers (LIU; TING; ZHOU, 2008) that uses a concept of machine learning that is the decision tree. It is one-dimensional (only takes one attribute at a time) and is required few parameters (this facilitates the configuration and use of the algorithm). No need to climb your values and a very robust algorithm for large datasets.

2.2.4 FDC

It is a depth-based algorithm (JOHNSON; KWOK; NG, 1998) approach for detection of outliers in 2D datasets based on the concept of ISODEPTH algorithm. The FDC computes the first k 2D depth contours (the points that can be considered outliers) by restricting to a small part of the complete dataset. In this way, it is more efficient by not having to calculate in the complete dataset and thus scaling more easily for large datasets of two dimensions.

2.2.5 HOD

It is a distance-based outlier detection method (XU et al., 2016) that emerges to overcome the statistics-based concept because in the vast majority of datasets the probability distribution is not known. In this way, the method search for outliers based on their distance from their neighbors and if that point has a distance greater than a predefined parameter, then that point is considered an outlier. However, if there is a cluster of outliers in the dataset, this can affect its detection by distance-based algorithms, with this comes the concept of HOD (Hidden Outlier Detection) algorithms that aim to find outliers even when they are grouped and in enough quantity to form a cluster.

2.2.6 ORCA

It is a distance-based algorithm (BAY; SCHWABACHER, 2003) that optimizes a simple nested loop algorithm (which are logarithmic algorithms and extremely inefficient when dealing with large datasets) by removing possible non-outliers during their execution. This way, instead of processing the complete dataset by calculating all possible distances, it removes unnecessary calculations that would be executed if a non-outlier point were taken to the end. From him, new researches have emerged further refining this concept.

2.2.7 Linearization

It is a distance-based algorithm (ANGIULLI; PIZZUTI, 2002) that detects outliers by calculating the sum of the distances of a point in relation to its neighbor, calling it weight, and setting an outlier as the points with the greatest weights in the dataset. In this way, it is an efficient algorithm and it is linearly scaled both in the number of points and in the number of dimensions. To calculate these outliers more efficiently the algorithm uses

the concept of the Hilbert space-filling curve.

2.2.8 RBRP

It is an algorithm for high-performance multidimensional datasets that is based on distances between the points to be able to define what the outliers are (GHOTING; PARTHASARATHY; OTEY, 2006). Its difference to the other distance-based algorithms is that it is more efficient for datasets with multiple dimensions and in comparisons with others, its scalability is approximately linear for the number of dimensions and logarithmic for the number of points in the dataset.

2.2.9 LOF

It is a density-based algorithm that adds a new concept in the search for outliers: the Local Outlier Factor (LOF) (BREUNIG et al., 2000), which is a degree of propensity to be an outlier so that the process of outlier definition is not more binary, but something gradual. With this, the approach is not to define whether a point is an outlier or not, but rather the “how outlier” that point is in that dataset. The outlier factor is local in the sense that only a neighborhood of that point is taken into account to define its factor.

2.2.10 ABOD

It is an angle-based algorithm (KRIEGEL; HUBERT; ZIMEK, 2008) for detection of outliers that is focused on high-dimensional datasets, different from other distance-based algorithms that end up damaged when one has many dimensions. Your approach is based on the calculation of a degree of angle between the different vectors of a point with its neighbors. With this, more centralized points within the cluster will have this degree calculated with a high value, the points more on the edge of the clusters will have this degree a little smaller and the possible outliers will have that degree with a very small value, since they will generally be far from the cluster in a particular direction.

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*

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

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 parametros. 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: *Baseado em Modelo*, *Baseado em Densidade*, *Baseado em Profundidade*, *Baseado em Distância* e *Altas Dimensões*. A questão **III** é relativa a performace 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 perfomance. A questão **IV** é sobre a perfomance 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 IDRs and Outliers

3.1 IDRs

TODO: brief description about the next subsections

3.1.1 User feedback

TODO: describe about user feedback and its benefit, *implicit and explicit* feedback with examples: gaze track, mouse click, mouse tracking, mouse track polygons, polygons intersections. use some image to demonstrate the idea

3.1.2 Interesting Dense Regions

TODO: describe about Interesting Dense Regions and its usability and advantages based on mouse track polygons. Use the figure from the IDR paper with description.

3.1.3 Algorithm

TODO: Show the used algorithm for IDR detection and explain how it works. present its complexity with pros and cons

3.2 Outliers

TODO: brief description about the next subsections

3.2.1 Atypical spatial data

TODO: explain about the existing process to capture spatial data and its different sources. show the existing spatial datasets with brief descriptions about each one with

references. summaries the problems of cleaning datasets and the bad analysis with *noise points*

3.2.2 Outliers in GeoGuide

TODO: explain the utility of detect outliers in GeoGuide and its advantages. describe a practical usage with real datasets examples

3.2.3 Algorithm

TODO: show the selected outlier detection algorithm. use image and explain step by step. explain about its complexity. present its pros and cons.

4 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.

4.1 Principais contribuições

Texto.

4.2 Limitações

Texto.

4.3 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.