UNIVERSIDADE ESTADUAL DO CEARÁ

CENTRO DE CIÊNCIAS E TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

MESTRADO PROFISSIONAL EM COMPUTAÇÃO APLICADA

BRUNO BEZERRA CHAVES

MÉTODOS COMBINATORIAIS PARA PROBLEMAS EM REDES DINÂMICAS: ALGORITMOS DE AGRUPAMENTO E PREVISÃO DINÂMICOS

FORTALEZA – CEARÁ

2018

BRUNO BEZERRA CHAVES

MÉTODOS COMBINATORIAIS PARA PROBLEMAS EM REDES DINÂMICAS: ALGORITMOS DE AGRUPAMENTO E PREVISÃO DINÂMICOS

Dissertação apresentada ao Curso de Mestrado Profissional em Computação Aplicada do Programa de Pós-Graduação em Ciência da Computação do Centro de Ciências e Tecnologia da Universidade Estadual do Ceará, como requisito parcial à obtenção do título de mestre em Ciência da Computação. Área de Concentração: Ciência da Computação

Orientador: Prof. Dr. Marcos José Negreiros Gomes

FORTALEZA – CEARÁ

2018

RESUMO

Arboviroses são doenças causadas pelos chamados arbovírus, que incluem o vírus da dengue, Zika vírus, febre chikungunya e febre amarela. Essas doenças estão cada vez mais voltando a atenção da OMS-TDR e das autoridades de saúde brasileiras para um esforço maior na prevenção e combate a endemias. Esta pesquisa fez importantes esforços para desenvolver, projetar e implementar uma estrutura computacional baseada na web, para ajudar a rastrear e gerenciar os recursos e pessoas no processo de prevenção e combate à arbovírus. Além disso, apresenta uma abordagem para solucionar o problema da previsão de agrupamentos dinâmicos espaço-temporal e combater as arboviroses com o esforço coordenado de uma estrutura de Sistemas de Apoio à Decisão para rastrear simultaneamente o mosquito e os casos em humanos, para prevenir e combater os territórios afetados. Foram implementados 2 métodos para visualização dos grupos formados. O primeiro é uma biblioteca que cria e gerencia grupos de acordo com o nível de zoom. O segundo, chamado algoritmo Convex Hull, consiste em gerar o menor polígono que englobe um determinado conjunto de pontos. O algoritmo ST-DBSCAN foi implementado como base para alcançar o método proposto. Por último, foram desenvolvidos 2 aprimoramentos no software DYNAGRAPH para possibilitar a avaliação dos métodos. Um deles foi a integração com um Editor de Características, que permite alterar os atributos visuais dos vértices e arestas de um grafo dinâmico. Outro novo recurso permitiu a visulalização da formação de novos grupos dinâmicos. TODO - predição

Palavras-chave: Agrupamento Dinâmicos, Grafos dinâmicos.

ABSTRACT

abstract **Keywords:**

LISTA DE ALGORITMOS

SUMÁRIO

1 INTRODUÇÃO

Grandes quantidades de dados estão disponíveis para análise em organizações hoje em dia. Estas enfrentam vários desafios quando se tenta analisar dados gerados com o objetivo de extrair informações úteis. Esta capacidade analítica precisa ser reforçada com ferramentas capazes de lidar com grandes conjuntos de dados sem tornar o processo de análise uma tarefa árdua. Agrupamento de dados normalmente são usados no processo de análise de dados, pois esta técnica não exige qualquer conhecimento prévio dos dados. Contudo, os algoritmos de agrupamento geralmente requerem um ou mais parâmetros de entrada que influenciam o processo de agrupamento e os resultados que podem ser obtidos.

Nos últimos anos, o problema de agrupamento dinâmico tem atraído o interesse de pesquisas, impulsionado pelo aumento da disponibilidade de grandes conjuntos de dados contendo elementos espaciais e temporais. Este problema pode ser analisado como um problema de otimização. Seu objetivo principal é maximizar as diferenças das características dos indivíduos de grupos distintos, e minimizar as diferenças das características dos indivíduos de um mesmo grupo.

Agrupamento de dados ganhou uso muito difundido, especialmente para dados estáticos. No entanto, o rápido crescimento de dados espaço-temporais de inúmeros instrumentos, como os satélites em órbita terrestre, criou uma necessidade de métodos de agrupamento espaço-temporais para extrair e monitorar clusters dinâmicos. O agrupamento espaço-temporal dinâmico enfrenta dois grandes desafios: primeiro, os clusters são dinâmicos e podem mudar de tamanho, forma e propriedades estatísticas ao longo do tempo. Em segundo lugar, vários dados espaço-temporais são incompletos, ruidosos, heterogêneos e altamente variáveis sobre espaço e tempo.

O problema de agrupamento dinâmico com a componente de previsão divide-se em passos. A primeira etapa é obtenção das informações espaço-temporais mapeáveis e características do indivíduo. Neste passo, segue-se três estratégias para resolução do problema: os dados são analisados como um só grupo (Agrupamento Estático); trata-se os dados por intervalos pré-definidos; e mapeamento das evoluções entre intervalos observados. Sendo assim, pretende-se indicar o conjunto de grupos espacialmente correlacionados também no tempo.

Já o problema de previsão de grupos dinâmicos introduz o conceito de indicar os possíveis grupos que serão formados no tempo após um conjunto de eventos serem observados previamente.

O algoritmo proposto para previsão de agrupamentos é uma importante contribuição

deste trabalho, uma vez que poderá ser usado na obtenção de informações, na previsão de movimentação dos grupos e recomendação para o combate a endemias. O serviço proposto se baseia na localização passada dos casos de Dengue e Chikungunya.

1.1 OBJETIVOS

A seguir, são expostos os objetivos desta dissertação, definindo o produto final a ser obtido.

1.1.1 Objetivo Geral

Estudo e aplicação de métodos existentes e proposta de um método para resolver o Problema de Agrupamento em Grafos Dinâmicos e previsão de evolução destes agrupamentos.

1.1.2 Objetivos Específicos

Para que se alcance o objetivo geral, as seguintes metas foram estabelecidas:

a) Utilizar o software Dynagraph como ambiente de suporte à visualização e interação com os resultados dos métodos de agrupamento espaço-temporal utilizados.
b) Extração de características de previsão espaço-temporal sobre a evolução dos agrupamentos dinâmicos.
c) Avaliação dos resultados sobre bases de dados reais ligadas a evolução de casos de Dengue e Chikungunya e outras bases dinâmicas.

1.2 HIPÓTESES

As hipóteses a seguir conduziram a elaboração desta dissertação:

a) É possível a criação de um algoritmo capaz de sugerir novos agrupamentos geolocalizados baseados no tempo.
b) É exequível a integração de um editor de características ao DYNAGRAPH, que é um software extensível.
c) É realizável a utilização do modelo proposto de agrupamentos em grafos dinâmicos em um ambiente Web.

1.3 JUSTIFICATIVA

Esta pesquisa justifica-se por perceber-se a necessidade de ferramentas e estudos relacionando os assuntos abordados: agrupamento, previsão em dados dinâmicos espaço-temporais, grafos dinâmicos e sistemas web de forma integrada. E também, acelerar técnicas de agrupamento em grafos dinâmicos para tomada de decisão.

A relevância da pesquisa está em permitir uma análise dos dados extraídos para apoio à tomada de decisão, onde concentra-se na avaliação dos resultados sobre bases de dados dinâmicas relativas a casos de Dengue e Chikungunya. A pesquisa toma como base as características de evolução dos casos da doença observados entre 2015 e 2018 em Fortaleza. Os dados foram tomados a partir de (SIMDA, 2018), onde um estado é definido como o período de uma semana.

1.4 METODOLOGIA

A seguir, são descritas as etapas metodológicas para o desenvolvimento da dissertação.

1.4.1 Etapas metodológicas do projeto

A presente pesquisa pode ser caracterizada quanto ao procedimento quantitativa e comparativa. A pesquisa quantitativa prioriza apontar numericamente a frequência e a intensidade dos comportamentos dos indivíduos de um determinado grupo, ou população. O método comparativo constitui-se em investigar coisas ou fatos e explicá-los de acordo com suas semelhanças e suas diferenças. Possibilita a análise de dados concretos e a dedução de semelhanças e divergências de elementos contínuos, abstratos e gerais, facilitando investigações de caráter indireto (FACHIN, 2001).

1.4.1.1 Revisão da literatura e soluções existentes

As fontes principais de pesquisa foram sites especializados em pesquisas científicas, por exemplo, o portal de periódicos da CAPES, IEEE e outros sites de referências em que possuem livros, periódicos e dissertações disponíveis. Os temas essenciais abordados na pesquisa foram:

- Estrutura de dados em grafos dinâmicos

- Modelos de previsão espaço-temporais
- Algoritmos de agrupamentos dinâmicos

1.4.1.2 Análise de requisitos

Detectou-se as necessidades de informações baseadas na previsão de agrupamentos dinâmicos em grafos. Assim sendo, e após obter os dados a partir de (SIMDA, 2018), houve a necessidade de um software para representação e tratamentos de grafos dinâmicos. Com isso foi escolhido o DYNAGRAPH, que tem como característica a extensibilidade.

1.4.1.3 Arquitetura do software

A arquitetura do software desenvolvido é apresentada, assim como o diagrama de caso de uso (UML).

1.4.1.4 Modelo e desenvolvimento de software

Foi desenvolvido um modelo capaz de representar agrupamentos e previsão dinâmicos. Em seguida, foi desenvolvido a partir do software DYNAGRAPH executar o modelo proposto e apresentar os resultados para validar a aplicação da ferramenta.

1.4.1.5 Ferramentas e Materiais

O desenvolvimento do trabalho foi realizado a partir dos seguintes equipamentos e materiais:

- Macbook Pro 13"modelo 2015 / macOS High Sierra 10.13.2:
  Processador Intel Core i5 2.7 GHz;
  Memória de 8GB 1867 MHzs DDR3;
  HD SSD 128GB;
- Ambiente de desenvolvimento Webstorm;
- Navegador de internet Google Chrome 63.0.3239.132.
- Controle de versão Git (Software e dissertação);
- LaTeXpara produção da dissertação;
- Portal de periódicos CAPES;

1.5 ORGANIZAÇÃO DO TEXTO

Esta dissertação está organizada em 5 capítulos. O capítulo 1 apresenta uma introdução à necessidade da representação e manipulação do agrupamento espaço-temporal dinâmico, assim como a previsão da formação de novos grupos dinâmicos. Em seguida são apresentados os objetivos, hipóteses, a metodologia utilizada e contribuições. O capítulo 2 constitui a revisão bibliográfica em modelagem com grafos dinâmicos, métodos de agrupamento por densidade, redes dinâmicas, o DYNAGRAPH, um editor de características e um conjunto de trabalhos relacionados a esta área do conhecimento. O capítulo 3 apresenta o modelo de agrupamento e previsão em redes dinâmicas. O capítulo 4 destaca os resultados e comparação dos algoritmos apresentados. O capítulo 5 apresenta as considerações finais e propostas de trabalhos futuros.

2 CONCEITOS E REVISÃO BIBLIOGRÁFICA

2.1 AGRUPAMENTOS

A técnica de agrupamento, também chamada de clustering, é uma das técnicas de mineração de dados mais comuns e é usada para descobrir padrões de distribuição nos dados. O agrupamento é feito com base na similaridade das características e na posição dos objetos. Dessa maneira, o objetivo é que objetos de mesmo grupo sejam muito similares entre si e muito diferentes dos objetos de outros grupos.

Essa técnica é muito utilizada para dados estáticos. No entanto, há pouco trabalho no âmbito espaço-temporal onde os dados estão na forma de campos espaço-temporais contínuos e os agrupamentos são dinâmicos. Além disso, os dados espaço-temporais originados por satélites em órbita terrestre, telefones celulares e outros sensores tendem a ser ruidosos, incompletos e heterogêneos, tornando sua análise especialmente desafiadora (FAGHMOUS; KUMAR, 2013).

Agrupamentos dinâmicos podem mudar seu tamanho, forma, localização e propriedades estatísticas de um único passo para o próximo. Embora os agrupamentos possam se mover ou mudar de forma, existem vários pontos que não alteram as associações de grupos por um período de tempo. Tendo isso em vista é possível extrair de forma autônoma agrupamentos dinâmicos em dados espaço-temporais contínuos que podem conter valores, ruídos ou características muito variáveis.

Agrupamento é o processo de encontrar grupos, também chamados de clusters, em um conjunto de dados a partir de suas características, de forma que elementos que pertencem ao mesmo grupo são mais similares entre si do que entre elementos de outros grupos. As medidas de similaridades podem variar de acordo com o algoritmo ou de acordo com os dados. Ao contrário do processo de classificação, onde existem rótulos predefinidos para categorizar os dados, no agrupamento não se sabe a priori quantos e quais clusters serão encontrados. Os principais métodos de agrupamento existentes na literatura podem ser categorizados, como mostra as próximas sessões.

2.1.1 Métodos baseados em particionamento

Esses métodos consistem na maneira mais fundamental de agrupamento. Eles recebem como entrada, além de um conjunto de dados D, um inteiro k que representa o número de clusters que se deseja encontrar. Inicialmente, o centro de cada grupo, que não necessariamente

é um ponto de D, é determinado. A cada passo, os elementos de D são associados ao cluster definido pelo centro mais próximo de cada ponto. A distância entre um elemento de D e um centro é calculada de acordo com algum critério de similaridade estabelecido pelo algoritmo. Após a etapa de associação, os centros dos grupos são atualizados. Esse processo é repetido até que um ótimo local seja atingido. Dentre os algoritmos mais conhecidos nessa categoria destacam-se o k-means e o k-medoids. No algoritmo k-means os centros dos grupos são conhecidos como centróides. A escolha dos centróides é comumente feita através de amostras aleatórias de D. Uma vez finalizada a etapa de associação de cada elemento do conjunto D ao seu centróide mais próximo, os valores dos centróides são atualizados de acordo com a média dos elementos que estão associados àquele centróide. Esses passos são repetidos até que os centróides não mudem. A dificuldade da escolha dos valores iniciais para os centróides, além do número de centróides são as principais desvantagens do algoritmo k-means. Ainda, não há garantia que o algoritmo convirja para um ótimo global, tornando o resultado do agrupamento bastante sensível à inicialização dos centróides. Por outro lado, a baixa complexidade computacional do algoritmo descrito o torna atrativo para aplicações onde existe um conhecimento prévio sobre o domínio do conjunto de dados

2.1.2 Agrupamento hierárquicos

Os métodos hierárquicos têm como resultado um aninhamento de grupos e o grau de similaridade entre esses grupos. Os algoritmos hierárquicos podem ser subcategorizados em aglomerativos e divisivos. Os aglomerativos funcionam de uma maneira botton-up, assumindo inicialmente que cada elemento do conjunto de dados representa um cluster. A partir daí, de acordo com algum critério similaridade, os clusters se unem. Já os divisivos trabalham de maneira top-down, considerando todo o conjunto de dados como um só cluster, que é dividido de maneira recursiva de acordo com a medida de similaridade estabelecida. O algoritmo BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies), considerado um dos métodos hierárquicos mais utilizados na literatura, faz a integração de outros métodos, tais como particionamento ou densidade, com organização hierárquica de clusters. A agrupamento acontece essencialmente em duas fases. Na primeira delas, uma árvore balanceada é construída através de estruturas chamadas clustering feature (CF), que sumarizam informações estatísticas sobre os clusters. Essa árvore, chamada de CF-tree, é utilizada para representar a hierarquia dos clusters. Na segunda fase, um algoritmo de agrupamento selecionado é aplicado às folhas da CF-tree, removendo

clusters esparsos e agrupando os mais densos em clusters maiores. Uma das maiores vantagens do algoritmo BIRCH é sua complexidade linear em relação ao número de objetos a ser agrupado.

2.1.3 Métodos baseados em grids

Enquanto que os outros métodos de agrupamento discutidos são orientados aos dados, os métodos baseados em grade são orientados ao espaço. Essencialmente, o espaço é dividido em células de uma estrutura de grade independente da distribuição dos dados. Essa estrutura é capaz de representar os objetos do conjunto de entrada através das células da grade. Todas as operações de agrupamento são realizadas na estrutura de grade criada. Essa é uma das principais vantagens dessa classe de algoritmos, uma vez que seu desempenho não depende diretamente do número de entradas do conjunto de dados, mas sim do número de células em cada dimensão da grade. O algoritmo STING (STatistical INformation Grid) se baseia na construção de diversas camadas de grade, onde células de uma camada mais alta são subdivididas para a criação de células nas camadas mais baixas. A partir dessa organização, informações estatísticas sobre as células são pré-computadas e armazenadas para a identificação dos clusters. Os resultados produzidos pelo STING se aproximam do agrupamento produzido pelo DBSCAN a medida que a granularidade da estrutura de grade se aproxima de 0, podendo também ser considerado como um método baseado em densidade. O STING também apresenta como vantagem complexidade linear de tempo em relação ao número de objetos a serem agrupados.

2.1.4 Agrupamento por densidade

Os métodos baseados em densidade assumem que os elementos que pertencem a um determinado cluster seguem uma mesma distribuição. O objetivo dessas técnicas é identificar os clusters de acordo com os parâmetros que descrevem essa distribuição. Em outras palavras, os clusters são modelados como regiões densas do conjunto de dados, divididos por áreas de regiões esparsas. Dentre as vantagens desses métodos destacase a capacidade de identificar clusters de formatos arbitrários, enquanto que os métodos baseados em particionamento, por exemplo, apresentam melhores resultados com clusters de formato circular. Dentre os algoritmos baseados em densidade existentes na literatura, o DBSCAN se destaca como um dos mais utilizados. As características e detalhes do algoritmo DBSCAN são descritos na Seção 1.

Os agrupamentos baseados em densidade analisam a quantidade de elementos dentro de uma vizinhança de acordo com determinados parâmetros. A idéia-chave é que, para cada

instância de um cluster, a vizinhança de um determinado raio deve conter pelo menos um número mínimo de instâncias. A possibilidade de encontrar agrupamentos de forma eventual e o fato de não precisar da definição do número de agrupamentos (YIP; DING; CHAN, 2005) como parâmetro inicial são as principais vantagens dos métodos baseados em densidade. Entretanto, alguns algoritmos podem exigir a definição de outros parâmetros, como o caso do algoritmo DBSCAN (ESTER *et al.*, 1996) abordado na próxima seção.

2.1.5 Método DBSCAN

Este algoritmo calcula a densidade de uma região contando quantos pontos existem em uma determinada área seguindo uma determinada métrica, geralmente uma medida de distância, como a euclidiana ou manhattan. O método DBSCAN separa os pontos de dados em três classes: • Pontos principais. Estes são pontos que estão no interior de um cluster. Um ponto é um ponto interior se houver pontos suficientes em sua vizinhança. • Pontos de fronteira. Um ponto de fronteira é um ponto que não é um ponto central, ou seja, não há pontos suficientes em sua vizinhança, mas ele está dentro da vizinhança de um ponto central. • Pontos de ruído. Um ponto de ruído é qualquer ponto que não é um ponto central ou um ponto de fronteira.

Para encontrar um cluster, o DBSCAN começa com uma instância arbitrária (p) no conjunto de dados (D) e recupera todas as instâncias de D em relação a Eps e MinPts Density-Based Algorithms for Discovering Clusters in Large Spatial Databases with Noise (DBSCAN) DBSCAN [1] é um algoritmo baseado em densidade que descobre clusters com forma arbitrária e com um número mínimo de parâmetros de entrada. Os parâmetros de entrada necessários para este algoritmo são o raio do cluster (Eps) e os pontos mínimos necessários dentro do cluster (Minpts).

Para agrupar os pontos levando em conta o fator tempo é necessário uma alteração no algoritmo DBScan, e com isso detectar os grupos em relação ao tempo. Logo, o algoritmo determinada para esta implementação foi o ST-DBScan (BIRANT; KUT, 2007), abordado a seguir.

2.2. Descrição do Algoritmo Nesta seção, o algoritmo DBSCAN [7] Clustering espacial baseado em densidade de aplicativos com ruído é projetado para descobrir os clusters de dados espaciais com ruído. As etapas envolvidas neste algoritmo são as seguintes, ... (i) Selecione um ponto arbitrário p (ii) Recuperar todos os pontos de densidade-reachable de p w.r.t. Eps e Minpts. (iii) Se p é um ponto central, um cluster é formado. (iv) Se p é um ponto de borda,

nenhum ponto é densidade acessível de p e DBSCAN visita o próximo ponto do banco de dados. (v) Continue o processo até que todos os pontos tenham sido processados.

2.3 Impacto do Algoritmo DBSCAN requer dois parâmetros de entrada (pontos mínimos e raio) e suporta o usuário ao encontrar um valor aproximado para ele usando o gráfico k-dist [7]. Ele descobre grupos de forma arbitrária. Ele é válido para grandes bancos de dados espaciais. . . .

2.4 Trabalho futuro O algoritmo DBSCAN aqui considera [1] apenas objetos de ponto, mas pode ser estendido para outros objetos espaciais, como polígonos. As aplicações do DBSCAN para espaços de recursos de alta dimensão devem ser investigadas e a geração de raio para esses dados de alta dimensão também precisa ser explorada. Também não consegue detectar agrupamentos com densidade variada.

DBSCAN (Density-based Spatial Clustering of Applications with Noise) é um algoritmo de agrupamento baseado em densidade vastamente utilizado pela comunidade científica. Seu objetivo principal consiste em encontrar concentrações de elementos que estão espacialmente próximos. Em outras palavras, o algoritmo busca por pontos que possuem mais que um limiar de vizinhos dentro de um certo raio. Caso um elemento p satisfaça essa propriedade, os vizinhos de p pertencerão ao mesmo cluster que p e o mesmo processo é aplicado a todos os seus vizinhos. Além do conjunto de dados a ser agrupado, o DBSCAN recebe também dois parâmetros de entrada: minPoints e eps. O primeiro deles se refere à quantidade mínima de pontos em um certo raio de vizinhança para a formação de um cluster. Já o segundo parâmetro se refere ao raio no qual a verificação de vizinhança é realizada. A função de distância utilizada para determinar a vizinhança de um certo ponto é definida de acordo com o tipo de dado a ser agrupado e deve obedecer às restrições de uma função de distância, tais como simetria e a desigualdade triangular, além de assumir que a distância entre dois elementos x e y só é igual a 0 se x = y. Dentre suas vantagens, o algoritmo DBSCAN se destaca por ser capaz de encontrar clusters com formatos arbitrários, além de ser capaz de lidar com ruídos nos dados, característica que não está presente na maioria dos algoritmos de agrupamento, como mostrado na Figura 2.1. As definições básicas utilizadas no DBSCAN são apresentadas a seguir: • |A|: cardinalidade do conjunto A. • Neps(p): é o conjunto de pontos q que estão a uma distância menor que eps do ponto p. Também é chamado de conjunto dos vizinhos de p. • Diretamente alcançável por densidade (DDR): um ponto p é DDR a partir de um ponto q se p E Neps(q) e |Neps(q)| >= minPoints.

Figura 2.1: Clusters de formatos arbitrários encontrados pelo algoritmo DBSCAN Fonte: Data Mining - The Hypertextbook • Alcançável por densidade (DR): um ponto p é DR a

partir de um ponto q, se existe uma sequência de pontos p1,..., pn onde p1 = p e pn = q, tal que pi+1 é DDR a partir de pi . • Conectado por densidade (DC): Um ponto p está conectado por densidade a um ponto q se existe um ponto o tal que p e q são DR a partir de o. • Core point: um ponto p é classificado como core point se |Neps(o)| >= minPoints. • Border point: um ponto p é classificado como border point se |Neps(p)| < minPoints e p é DDR a partir de um core point. • Noise: Um ponto p é classificado como noise se |Neps(q)| < minPoints e p não é DDR a partir de nenhum core point. No contexto do algoritmo DBSCAN, um cluster C é definido como um subconjunto não vazio dos dados que satisfaz as seguintes propriedades: • Maximalidade: Para quaisquer dois pontos p e q, se p E C e q é alcançável por densidade (DR) a partir de p, então q E C. • Conectividade: Para quaisquer dois pontos p,q E C, p é q são conectados por densidade (DC). O Algoritmo 1 mostra o pseudocódigo do DBSCAN. Para cada elemento p ainda não visitado o conjunto dos seus vizinhos Neps(p) é encontrado, como podemos ver entre as linhas 2 e 5. Caso a cardinalidade desse conjunto de vizinhos seja maior que o valor de minPoints (Linha 6 do Algoritmo), um novo cluster C é criado e p e seus vizinhos serão atribuídos a C. Ainda, os pontos não visitados de C serão expandidos em um processo similar. A agrupamento acaba quando todos os elementos do conjunto foram visitados. O Algoritmo 2 implementa a função de expansão de um cluster C. A expansão de um cluster a partir de um ponto p encontra todos os elementos que são conectados por densidade (DC) a p. Essa função recebe como entrada p, seu conjunto de vizinhos NeighborPts, o identificador C do cluster a ser expandido, e os parâmetros eps e minPoints. Para cada ponto p' do conjunto de vizinhos de p, caso esse ponto ainda não tenha sido visitado, sua vizinhança é recuperada e adicionada ao conjunto NeighborPts de vizinhos que serão verificados (linha 8). Após sua verificação, caso p' não pertença a nenhum cluster, ele é adicionado ao cluster C (linhas 11 a 13). O processo finaliza quando o conjunto NeighborPts está vazio.

Alg 1 e alg 2

Como podemos ver nos Algoritmos 1 e 2, a complexidade do DBSCAN depende diretamente do custo computacional para recuperação da vizinhança de um ponto (linhas 7 e 6 dos Algoritmos 1 e 2, respectivamente). Em uma solução ingênua essa operação poderia ser executada em tempo linear, onde uma simples busca exaustiva em todo o conjunto de dados retornaria apenas os elementos a uma certa distância do ponto de consulta. Tal solução faria com que a complexidade do algoritmo DBSCAN fosse O(n2). Por outro lado, com o auxílio de estruturas de índices, como k-d-Trees ou R-Trees, a complexidade do DBSCAN pode ser significantemente reduzida. Recentemente foi provado em (GAN; TAO, 2015) que para

dimensões maiores que 2 o algoritmo DBSCAN executa em uma complexidade Omega(n4/3). No entanto, consideraremos nesse trabalho conjuntos de dados de apenas duas dimensões.

2.1.6 Método ST-DBSCAN

6.1. Introdução

O algoritmo ST-DBSCAN é construído modificando o algoritmo DBSCAN [7]. Em contraste com o algoritmo de agrupamento baseado em densidade existente, o algoritmo ST-DBSCAN [12] tem a capacidade de descobrir clusters em relação aos valores não espaciais, espaciais e temporais dos objetos. As três modificações feitas no algoritmo DBSCAN são as seguintes,

(i) O algoritmo ST-DBSCAN pode agrupar dados espaciais-temporais de acordo com atributos não espaciais, espaciais e temporais. (ii) DBSCAN não detecta pontos de ruído quando é de densidade variada, mas isso o algoritmo supera esse problema ao atribuir o fator de densidade a cada cluster. (iii) Para resolver os conflitos em objetos de borda, ele compara o valor médio de um cluster com o novo valor que vem.

6.2. Descrição do Algoritmo O algoritmo começa com o primeiro ponto p no banco de dados D. (i) Este ponto p é processado de acordo com o algoritmo DBSCAN e o próximo ponto é tomado. (ii) A função RetrieveNeighbors (objeto, Ep1, Ep2) recupera todos os objetos densidade-acessível do objeto selecionado em relação a Eps1, Eps2 e Minpts. Se os pontos devolvidos no Eps-neighborhood são menores do que Minpts, o objeto é atribuído como ruído. (iii) Os pontos marcados como ruído podem ser alterados posteriormente, e os pontos não são diretamente acessíveis, mas serão densidade-acessível. ... (iv) Se o ponto selecionado for um objeto central, um novo cluster será construído. Então, todos os vizinhos de densidade direta de este núcleo de objetos também estão incluídos. (v) Então, o algoritmo coleta de forma iterativa objetos atingidos pela densidade do objeto do núcleo usando a pilha. (vi) Se o objeto não estiver marcado como ruído ou não estiver em um cluster e a diferença entre o valor médio do cluster e o novo valor é menor do que DeltaE, ele é colocado no cluster atual.

2.2 REDES DINÂMICAS

2.2.1 O modelo Dynagraph

2.2.2 Editor de características

2.3 TRABALHOS RELACIONADOS

Como há uma carência de estudos relacionando os assuntos abordados: agrupamento, previsão em dados dinâmicos espaço-temporais, grafos dinâmicos e sistemas web de forma integrada, foi necessário dividir o problema de agrupamentos e previsões dinâmicos em três etapas:

- Estrutura de dados em grafos dinâmicos
- Modelos de previsão espaço-temporais
- Algoritmos de agrupamentos dinâmicos

A pesquisa aborda estrutura de dados em grafos dinâmicos usando passos já descritos na literatura, principalmente o modelo Dynagraph (CALIXTO; NEGREIROS, 2013), que é baseado na primeira proposta em (CALIXTO; NEGREIROS, 2012), onde o Dynagraph usa sequências temporais para vértices, arestas, características modificáveis dos vértices e arestas e o relacionamento entre suas características. Com isso, é formado um grafo com as informações necessárias para qualquer instante no tempo. O Dynagraph é capaz de visualizar o comportamento do grafo ao longo de um período de tempo, e editá-lo.

A ideia central de (KIM; ANDERSON, 2012) é modelar uma rede dinâmica como digrafos orientados ao tempo (*time-ordered graph*), que é gerada através da ligação de instantes temporais com arestas direcionadas que unem cada nó ao seu sucessor no tempo. Com isso, transformar uma rede dinâmica em um grafo maior, mas facilmente analisável. Isto permite não só a utilização dos algoritmos desenvolvidos para grafos estáticos, mas também para melhor definir métricas para grafos dinâmicos. Segundo (KIM; ANDERSON, 2012) um sistema de grafos dinâmicos é um objeto de representação visual que pode descrever melhor o comportamento dinâmico de objetos relacionados a eventos dinâmicos e introduzir novas formas de enxergar ou descrever a evolução de eventos dinâmicos na natureza.

(KOSTAKOS, 2009) considera a estrutura de grafos temporais como grafos estáticos, no entanto avança sobre as métricas introduzindo conceitos como disponibilidade temporal, proximidade temporal e geodésica, e estuda os seus grafos sobre redes reais.

Segundo (ESTER *et al.*, 1996), o algoritmo DBScan(*Density-Based Spatial Clustering of Applications With Noise*) calcula a densidade de uma região contando quantos pontos existem em uma determinada área seguindo uma determinada métrica. Ele permite a redução de pontos não pertencentes a nenhum padrão, assim como possibilita a formação de grupos de diferentes formas. Seu objetivo principal é dividir os pontos em grupos através da densidade de cada região.

(LAHIRI; BERGER-WOLF, 2007) apresentam um algoritmo de predição em redes temporais, e que usa a ideia de que certas interações sinalizam a ocorrência de outros em algum momento no futuro. Através de análises estatísticas o algoritmo mede o atraso entre as interações, e com isso pode-se prever quando certas interações vão ocorrer com base em observações passadas e atuais. Propõe-se a utilização de subgrafos frequentes e discute como identificar subgrafos que são persistidos em redes temporais. (LAHIRI; BERGER-WOLF, 2008) em seguida propõe um novo problema de mineração de dados para redes dinâmicas: detecção de todos os padrões de interação que ocorrem em intervalos de tempo regulares.

3 MODELO DE AGRUPAMENTO E PREVISÃO EM REDES DINÂMICAS

4 AVALIAÇÃO DO MÉTODO PROPOSTO

5 CONCLUSÕES E TRABALHOS FUTUROS

5.1 CONSIDERAÇÕES FINAIS

Criação de um método para resolver o Problema de Agrupamento em Grafos Dinâmicos e previsão de evolução destes agrupamentos. A expectativa é de que ao final desta pesquisa, tenha-se uma extensão do Dynagraph para visualização de agrupamentos dinâmicos e previsão dinâmica. Utilizar a ferramenta para outros tipos de doenças como: Chikungunya e Zika Vírus. Finalmente, espera-se que o produto final e os resultados obtidos possibilitem a previsão e prevenção de novos casos de dengue para um combate efetivo à doença.

5.2 LIMITAÇÕES

Dentre as dificuldades que podem interferir na execução deste projeto de pesquisa, as seguintes podem ser citadas: 1. A escassez de estudos relacionando os assuntos abordados: agrupamento, previsão em dados dinâmicos espaço-temporais, grafos dinâmicos e sistemas web de forma integrada; 2. Obtenção das informações dos focos e casos de dengue geolocalizadas e tempos das ocorrências. A extração dos dados semanais requer um processo manual em (SIMDA, 2018), pois é necessário o usuário selecionar o ano e a semana correspondente; 3. Visualização dos agrupamentos dinâmicos. Para contornar as dificuldades apresentadas pretende-se: 1. Automatizar a forma de obtenção dos dados; 2. Utilizar o software Dynagraph como ambiente de suporte à validação e interação com os resultados dos métodos de agrupamento espaço-temporal utilizados.

5.3 TRABALHOS FUTUROS

REFERÊNCIAS

BIRANT, D.; KUT, A. St-dbscan: An algorithm for clustering spatial-temporal data. **Data Knowl. Eng.**, v. 60, p. 208–221, 2007.

CALIXTO, A.; NEGREIROS, M. DYNAGRAPH: Um Modelo de Edição e Representação de Grafos Dinâmicos. 1 ed. CLAIO/SBPO, p. 8, 2012.

CALIXTO, A.; NEGREIROS, M. **DYNAGRAPH: Um Modelo de Edição e Representação de Grafos Dinâmicos**. Dissertação (Mestrado) — Mestrado Profissional em Computação Aplicada (MPCOMP), Universidade Estadual do Ceará, 2013.

ESTER, M.; KRIEGEL, H.-P.; SANDER, J.; XU, X. A density-based algorithm for discovering clusters in large spatial databases with noise. KDD, v. 96, n. 34, p. 226–231, 1996.

FACHIN, O. **Fundamentos de metodologia**. [S.l.]: 3a Edição, Editora Saraiva, 2001.

FAGHMOUS, J. H.; KUMAR, V. **Spatio-Temporal Data Mining for Climate Data: Advances, Challenges, and Opportunities**. [S.l.: s.n.], 2013. 83-116 p.

KIM, H.; ANDERSON, R. Temporal node centrality in complex networks. 1 ed. PHYSICAL REVIEW, p. 8, 2012.

KOSTAKOS, V. Temporal graphs. **Physica A**, p. 1007–1023, 2009.

LAHIRI, M.; BERGER-WOLF, T. Structure prediction in temporal networks using frequent subgraphs. IEEE Symposium on Computational Intelligence and Data Mining, p. 35–42, 2007.

LAHIRI, M.; BERGER-WOLF, T. Mining periodic behavior in dynamic social networks. Eighth IEEE International Conference on Data Mining, 2008.

SIMDA. **Sistema de Monitoramento Diário de Agravos**. 2018. Disponível em: <http://tc1.sms.fortaleza.ce.gov.br/simda/index>.

YIP, A. M.; DING, C.; CHAN, T. F. Dynamic cluster formation using level set methods. In: **Proceedings of the 9th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining**. Berlin, Heidelberg: Springer-Verlag, 2005. (PAKDD'05), p. 388–398. ISBN 3-540-26076-5, 978-3-540-26076-9. Disponível em: <http://dx.doi.org/10.1007/11430919_46>.