Projeto 500 Cities: Detecção de Comunidades utilizando Algoritmos de Clusterização

500 Cities Project: Detection of Communities Using Clustering Algorithms

Anderson Vinícius Alves Ferreira[1] orcid.org/0000-0001-8598-6574
Lizandra Raflesia Monteiro de Lira[1] orcid.org/0000-0002-2379-5868
Thiago José da Silva[1] orcid.org/0000-0002-1710-2148
Carmelo José Albanez Bastos Filho[1] orcid.org/0000-0002-0924-5341

[1] Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil,

E-mail do autor principal: avaf@ecomp.poli.br

Resumo

A saúde pública é uma extensa área com problemas complexos e que constantemente necessita de bastantes investimentos. Frequentemente os órgãos governamentais enfrentam desafios para entender como oferecer melhores serviços de saúde e prevenir epidemias futuras. Métodos preventivos tem sido até o momento a melhor opção para controlar doenças e epidemias ou mesmo extingui-las. Este trabalho utilizou dados epidemiológicos provenientes do projeto 500 Cities e técnicas de agrupamento (clusterização) de dados para identificar comunidades com características relevantes para dar suporte na prevenção de epidemias e doenças.

Palavras-Chave: Saúde Pública; Agrupamento; Clusterização; Comunidades;

Abstract

Public health is a large area with complex problems and constantly needs huge investments. Frequently, government agencies face challenges in understanding how to deliver effective and targeted health services and prevent future epidemics. Prevention has so far been the best option to control diseases and epidemics or even extinguish them. This work used epidemiological data provided by the 500 Cities project along with data clustering techniques to identify communities with relevant characteristics to support epidemics and diseases prevention.

Key-words:*Public health; Clustering; Communities;*

 http://dx.doi.org/10.25286/repa.v3i3.966

1 Introdução

A saúde pública é uma extensa área com problemas complexos e que constantemente necessita de bastantes investimentos. Frequentemente, os órgãos governamentais enfrentam desafios para entender como oferecer melhores serviços de saúde e prevenir epidemias futuras. Como controlar as epidemias? Como identificar as epidemias? Como o governo pode interferir na prevenção de epidemias e doenças? Existem tendências regionais de saúde? Existem regiões com comportamento incomum de piora na saúde? Alguma região se destaca por altos níveis de bem-estar? Será que é possível predizer as condições de saúde de uma cidade baseado na situação de cidades vizinhas?

A fim de criar soluções e intervenções plausíveis e efetivas para as atuais necessidades, os métodos preventivos têm sido até o momento a melhor opção, fazendo com que em um futuro próximo algumas doenças sejam controladas ou mesmo extintas. Torna-se importante, portanto, saber quais os hábitos de saúde das populações das cidades e quais impactos tais hábitos poderiam ter sobre as condições gerais de saúde das populações.

Em 2015, nos Estados Unidos, a Fundação Robert Wood Johnson lançou o projeto 500 Cities **[1]** (em tradução livre, 500 Cidades), que tem como objetivo reportar dados epidemiológicos das 500 maiores cidades americanas. O projeto prevê que os dados sejam utilizados por órgãos governamentais para ajudar a desenvolver e implementar atividades de prevenção efetivas e direcionadas, identificar problemas de saúde pública emergentes, e estabelecer e monitorar objetivos relevantes para a saúde pública.

O conjunto de medidas disponibilizado pelo projeto *500 Cities* baseia-se em doenças crônicas prioritárias e de maior impacto na saúde pública. As medidas incluem os comportamentos de maior risco que causam doenças, sofrimento e morte precoce relacionados a doenças e condições crônicas, bem como as condições e doenças mais comuns, dispendiosas e previníveis entre todos os problemas de saúde. O total de 27 medidas inclui 5 comportamentos não saudáveis, 13 indicadores de saúde e 9 práticas de prevenção.

O projeto *500 Cities* representa o primeiro projeto a fornecer informações em larga escala para cidades e pequenas áreas dentro das cidades. E, apesar dos dados reportados serem provenientes de questionários respondidos pelos residentes das cidades, pode ser possível obter análises interessantes a respeito das condições de saúde de uma determinada população.

Este trabalho visa a identificar características relevantes para dar suporte na prevenção de epidemias e doenças e agrupar as diferentes cidades de acordo com seus hábitos e condições de saúde.

2 Fundamentação teórica

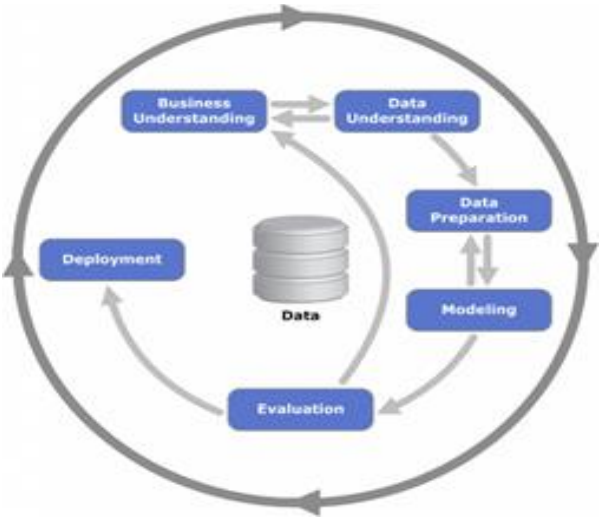
2.1 Área de conhecimento

A quantidade de dados que são gerados e armazenados cresce exponencialmente a cada ano. Dessa forma, é fundamental utilizar essas informações e verificar se existe alguma informação ou padrão dentro delas. Mineração de dados, ou *Data Mining* em inglês, refere-se a extração de informação implícita, previamente desconhecida e potencialmente útil em grandes conjuntos de dados.

O processo de mineração de dados comumente utilizado é chamado de CRISP-DM.

2.2 Mineração de Dados

A técnica de trabalho CRISP-DM é dividida em 6 principais etapas:



2.2.1 Entendimento do Negócio

Esta é a etapa de compreender de forma adequada o problema que necessita ser resolvido. É preciso buscar detalhes sobre como a questão afeta a organização e quais são os principais objetivos e expectativas em relação ao trabalho como um todo.

2.2.2 Compreensão dos dados

Após a primeira etapa, o objetivo torna-se inspecionar, organizar e descrever todos os dados disponíveis. É fundamental a avaliação em busca de quais dados podem ser relevantes para decifrar o problema.

2.2.3 Preparação dos dados

Definidos, organizados e bem inspecionados, nesta etapa é preciso preparar todas as databases, definir o formato que será necessário para a análise e ajustar demais questões técnicas.

2.2.4 Modelagem

Neste quarto momento, são selecionadas e aplicadas as técnicas de mineração de dados mais apropriadas, dependendo dos objetivos identificados na primeira fase.

2.2.5 Avaliação

Considerada uma etapa de *after-work*, mas ainda assim extremamente importante para a vitalidade do ciclo, a quinta fase pede o acompanhamento dos resultados objetivos e a avaliação da aplicabilidade confiável dos insights e conhecimentos obtidos.

2.2.6 Desenvolvimento

Todo o conhecimento que for obtido por meio do trabalho de mineração e modelagem agora poderá ser aplicado de forma prática.

2.3 Agrupamento (Clusterização)

Agrupamento (ou clusterização) é a tarefa de agrupar um conjunto de objetos de forma que os objetos do mesmo grupo (chamados de cluster) sejam mais semelhantes (em algum sentido) uns aos outros do que àqueles de outros grupos (clusters). É uma tarefa principal de mineração exploratória de dados e uma técnica comum para análise de dados estatísticos, usada em muitos campos, incluindo aprendizado de máquina, reconhecimento de padrões, análise de imagens, recuperação de informações, bioinformática, compactação de dados e computação gráfica.

Em Clusterização, não existe o conceito de um algoritmo "correto", mas sim o algoritmo mais apropriado para um problema em particular. Normalmente, a escolha do algoritmo precisa ser feita experimentalmente, a menos que haja uma razão matemática para preferir um modelo de cluster a outro. A visão geral a seguir listará apenas os exemplos mais proeminentes de algoritmos de agrupamento.

2.3.1 *K-Means*

O algoritmo *K-Means* **[2]** agrupa dados tentando separar amostras em *n* grupos de igual variância, minimizando um critério conhecido como inércia ou soma de quadrados intra-*cluster*. Esse algoritmo requer que o número de *clusters* seja especificado. Ele se adapta bem a um grande número de amostras e foi usado em uma grande variedade de áreas de aplicação em muitos campos diferentes.

O algoritmo *K-Means* divide um conjunto de **N** amostras **X** em **K** clusters disjuntos **C**, cada um descrito pela média μj das amostras no cluster. As médias são comumente chamadas de "centróides" do *cluster*; note que eles não são, em geral, pontos de **X**, apesar de estarem no mesmo espaço. O algoritmo *K-Means* visa escolher centróides que minimizem a inércia, ou a soma dos quadrados *intra-cluster*. A inércia pode ser considerada como uma medida de quão internamente coerentes são os clusters.

$$\sum_{i=0}^{n} \min_{\mu_j \in C}(||x_j - \mu_i||^2) \quad \textbf{(1)}$$

http://dx.doi.org/10.25286/repa.v3i3.966

2.3.2 Fuzzy C-Means

O Fuzzy C-Means (FCM) **[3]** é uma técnica de agrupamento em que cada objeto pode pertencer a mais de um *cluster*.

No algoritmos não-fuzzy, os dados são divididos em *clusters* distintos, em que cada objeto pode pertencer apenas a um *cluster*. No FCM, os objetos podem pertencer a vários *clusters*.

Os graus de pertinência são atribuídos a cada um dos objetos. A pertinência indica o grau em que os objetos pertencem a cada *cluster*. Assim, os pontos na borda de um *cluster*, com graus de pertinência inferiores, podem estar no *cluster* em um grau menor que os pontos no centro do *cluster*.

O FCM visa minimizar a seguinte função objetivo:

$$\arg\min_{C} \sum_{i=1}^{n} \sum_{j=1}^{c} w_{ij}^{m} \|\mathbf{x}_i - \mathbf{c}_j\|^2, \tag{2}$$

onde:

$$w_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|\mathbf{x}_i - \mathbf{c}_j\|}{\|\mathbf{x}_i - \mathbf{c}_k\|} \right)^{\frac{2}{m-1}}}. \tag{3}$$

2.3.3 *Affinity Propagation*

O algoritmo *Affinity Propagation* **[4]** cria *clusters* enviando mensagens entre pares de amostras até a convergência. Um conjunto de dados é então descrito usando um pequeno número de exemplares, que são identificados como os mais representativos de outras amostras. As mensagens enviadas entre pares representam a adequação para uma amostra ser o exemplar da outra, que é atualizada em resposta aos valores de outros pares. Essa atualização ocorre iterativamente até a convergência, momento em que os exemplares finais são escolhidos e, portanto, o agrupamento final é dado.

As mensagens enviadas entre pontos pertencem a uma das duas categorias. A primeira é a responsabilidade $r(i,k)$, que é a evidência acumulada de que a amostra *'k'* deve ser o exemplar da amostra *'i'*. A segunda é a disponibilidade $a(i,k)$, que é a evidência acumulada de que a amostra *'i'* deve escolher a amostra *'k'* para ser seu exemplar, e considera os valores para todas as outras amostras que *'k'* devem ser exemplares. Deste modo, os exemplares são escolhidos por amostras se forem semelhantes o suficiente para muitas amostras e escolhidas por muitas amostras para serem representativas de si mesmas.

Mais formalmente, a responsabilidade de uma amostra *'k'* ser o exemplar da amostra *'i'* é dada por:

$$r(i,k) \leftarrow s(i,k) - max[a(i,k') + s(i,k') \forall k' \neq k] \tag{4}$$

Onde $s(i,k)$ é a semelhança entre amostras *'i'* e *'k'*. A disponibilidade de amostra *'k'* para ser o exemplar da amostra *'i'* é dada por:

$$a(i,k) \leftarrow min[0, r(k,k) + \sum_{i' \ s.t. \ i' \notin \{i,k\}} r(i',k)] \tag{5}$$

2.3.4 *Mean Shift*

O algoritmo *Mean Shift* **[5]** visa descobrir 'bolhas' em uma densidade suave de amostras. É um algoritmo baseado em centróide, que funciona através da atualização de candidatos para centróides para ser a média dos pontos dentro de uma determinada região. Esses candidatos são então filtrados em um estágio de pós-processamento para eliminar quasi-duplicatas para formar o conjunto final de centróides.

Dado um centróide candidato, a atualização do centróide é realizada de acordo com a seguinte equação:

$$x_i^{t+1} = x_i^t + m(x_i^t) \tag{6}$$

Onde:

$$m(x_i) = \frac{\sum_{x_j \in N(x_i)} K(x_j - x_i) x_j}{\sum_{x_j \in N(x_i)} K(x_j - x_i)} \tag{7}$$

$N(x_i)$ é a vizinhança de amostras dentro de uma determinada distância ao redor de x_i e m é o vetor de deslocamento médio que é calculado para cada centróide e que aponta para uma região do aumento máximo na densidade de pontos.

O algoritmo Mean Shift define automaticamente o número de clusters, em vez de depender de um parâmetro.

2.3.5 *Spectral Clustering*

O algoritmo *Spectral Clustering* **[6]** faz uma incorporação de baixa dimensão da matriz de afinidade entre as amostras, seguida por um K-Means no espaço de baixa dimensão. O *Spectral Clustering* requer o número de clusters a serem especificados. Ele funciona bem para um pequeno número de clusters, mas não é aconselhável ao usar muitos clusters.

Para dois clusters, ele resolve um relaxamento convexo do problema de cortes normalizados no grafo de similaridade: cortar o gráfico em dois, de modo que o peso do corte das bordas seja pequeno em comparação com os pesos das bordas dentro de cada cluster.

2.3.6 *Agglomerative Clustering*

O algoritmo *Agglomerative Clustering* **[7]** executa um agrupamento hierárquico usando uma abordagem de baixo para cima: cada objeto inicia em seu próprio cluster e os clusters são mesclados sucessivamente. Os critérios de ligação determinam a métrica usada para a estratégia de mesclagem:

- 'Ward' minimiza a soma das diferenças quadradas dentro de todos os clusters. É uma abordagem de minimização de variações e, nesse sentido, é semelhante à função objetivo do K-Means, mas com uma abordagem hierárquica aglomerativa;
- A ligação máxima ou completa minimiza a distância máxima entre observações de pares de clusters;
- A ligação média minimiza a média das distâncias entre todas as observações de pares de clusters.

2.3.7 *Birch*

O algoritmo *Birch* **[8]** constrói uma árvore chamada *Characteristic Feature Tree* (CFT) para os dados fornecidos. Os dados são essencialmente compactados para um conjunto de nós *Characteristic Feature* (CF Nodes). Os nós CF têm um número de *subclusters* chamados *subclusters Characteristic Feature* (CF Subclusters) e estes *subclusters* CF localizados nos nós CF-não-terminais podem ter CF Nodes como filhos.

Os *subclusters* CF armazenam as informações necessárias para o agrupamento em cluster, o que elimina a necessidade de manter todos os dados de entrada na memória.

O algoritmo *Birch* possui dois parâmetros, o limiar e o fator de ramificação. O fator de ramificação limita o número de *subclusters* em um nó e o limiar limita a distância entre a amostra inserida e os *subclusters* existentes.

2.3.8 *Particle Swarm Optimization for Clustering* (PSOC)

A otimização de enxame de partículas (PSO) é um processo de busca estocástica, modelado a partir do comportamento social de um bando de aves **[9]**. O algoritmo mantém uma população de partículas, onde cada partícula representa uma potencial solução para um problema de otimização. No contexto do PSO, um enxame refere-se a um número de soluções potenciais para o problema de otimização, onde cada solução potencial é referida como uma partícula. O objetivo do PSO é encontrar a posição das partículas que resulta na melhor avaliação de uma determinada função de aptidão (objetivo).

No contexto do algoritmo PSOC **[10]**, uma única partícula representa os vetores dos centróides dos *clusters*. Ou seja, cada partícula é construída do seguinte modo:

$$\mathbf{x}_i = \left(\mathbf{m}_{i1}, \cdots, \mathbf{m}_{ij}, \cdots, \mathbf{m}_{iN_c}\right) \quad \textbf{(8)}$$

onde mij se refere ao vetor centróide do j-ésimo cluster da i-ésima partícula no cluster Cij. Portanto, um enxame representa um número de agrupamentos candidatos para os vetores de dados atuais. A aptidão das partículas é medida como o erro quantização:

$$J_e = \frac{\sum_{j=1}^{N_c} [\sum_{\forall \mathbf{z}_p \in C_{ij}} d(\mathbf{z}_p, \mathbf{m}_j)/|C_{ij}|]}{N_c}$$

http://dx.doi.org/10.25286/repa.v3i3.966

(9)

onde d é a distância euclidiana e |Cij| é o número de objetos pertencentes ao cluster Cij, ou seja, a freqüência desse cluster.

2.3.9 *Particle Swarm Clustering* (PSC)

O algoritmo *Particle Swarm* for *Clustering* (PSC) **[11]** apresenta uma codificação diferente das partículas em relação ao PSOC para o problema de agrupamento. Nessa abordagem, cada partícula pi representa um único centróideck, que navega pelo espaço de buscas até encontrar a posição ótima correspondente aos centróides das regiões de alta densidade na base de dados. Como cada partícula é um único centróide, a solução para o problema, neste caso, é todo o enxame.

2.3.10 Métricas para Avaliação de *Clusters*

A avaliação de desempenho de um algoritmo de clusterização não é trivial quanto contar o número de erros ou a precisão e a recuperação de um algoritmo de classificação supervisionado. Em particular, qualquer métrica de avaliação não deve levar em consideração os valores absolutos dos rótulos dos clusters, mas sim se esse agrupamento define uma separação dos dados semelhante a algum conjunto verdadeiro preestabelecido de classes ou então se satisfaz alguma suposição de que os membros que pertencem à mesma classe são mais semelhantes que membros de classes diferentes de acordo com alguma métrica de similaridade.

2.3.10.1 Coeficiente de Silhueta

O Coeficiente de Silhueta **[12]** é calculado usando a distância intra-cluster média (a) e a distância média do *cluster* mais próximo (b) para cada amostra. O coeficiente de silhueta para uma amostra é (b - a) / max (a, b). Para esclarecer, b é a distância entre uma amostra e o *cluster* mais próximo do qual a amostra não faz parte.
O melhor valor do Coeficiente de Silhueta é 1 e o pior valor é -1. Valores próximos a 0 indicam *clusters* sobrepostos. Valores negativos geralmente indicam que uma amostra foi atribuída ao *cluster* errado, pois um *cluster* diferente é mais semelhante.

2.3.10.2 Índice de Calinski-Harabasz

Se os rótulos para as classes dos dados não forem conhecidos a priori, o índice de Calinski-Harabasz **[13]** pode ser usado para avaliar o modelo, onde uma pontuação mais elevada de Calinski-Harabasz se relaciona a um modelo com clusters melhor definidos.

Para clusters, o índice de Calinski-Harabasz é dado como a razão entre a média da dispersão entre os clusters e a dispersão dentro do cluster:

$$s(k) = \frac{\mathrm{Tr}(B_k)}{\mathrm{Tr}(W_k)} \times \frac{N-k}{k-1}$$

(10)

onde B_K é a matriz de dispersão entre grupos e W_K é a matriz de dispersão dentro do cluster definida por:

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T$$
$$B_k = \sum_{q} n_q (c_q - c)(c_q - c)^T$$

(11)

com Nigual ao número de objetos, C_q o conjunto de objetos no cluster q, c_q o centro do cluster q, c o centro de E, n_q o número de objetos no cluster q.

3 Materiais e Métodos

3.1 Descrição da base de dados

Os dados utilizados neste trabalho são provenientes de uma base pública fornecida pelo projeto 500 Cities. O projeto 500 Cities reporta dados de cidades e dados censitários, obtidos usando métodos de estimação em pequenas áreas, para 27 medidas de doenças crônicas nas 500 maiores cidades americanas. O total de 27 medidas inclui 5 comportamentos não-saudáveis, 13 indicadores de saúde e 9 práticas de prevenção.

Indicadores de Saúde	Práticas de Prevenção	Comportamentos Não-Saudáveis
Artrite	Triagem de colesterol	Beber compulsivamente
Câncer (excluindo câncer de pele)	Falta de seguro de saúde entre adultos de 18 a 64 anos	Tabagismo
Doença renal crônica	Tomar remédio para o controle da pressão alta entre aqueles com pressão alta	Nenhuma atividade física de lazer
Doença de obstrução pulmonar crônica	Visitas ao médico para check-up de rotina no último ano	Dormir menos de 7 horas
Doença cardíaca coronariana	Visitas ao dentista ou clínica odontológica	Obesidade
Asma	Uso de mamografia entre mulheres de 50 a 74 anos	

Indicadores de Saúde	Práticas de Prevenção	Comportamentos Não-Saudáveis
Diabetes	Papanicolaou entre mulheres adultas com idade entre 21-65 anos	
Pressão alta	Exame de sangue oculto nas fezes, sigmoidoscopia ou colonoscopia entre adultos de 50 a 75 anos	
Colesterol alto entre adultos com idade ≥ 18 anos que foram rastreados nos últimos 5 anos	Idosos ≥ 65 anos que estejam atualizados em um conjunto de serviços clínicos preventivos (Homens: vacina contra Polissacarídeos Pneumocócicos - PPV, Rastreio do câncer colorretal; Mulheres: Igual ao anterior e Mamografia nos últimos 2 anosanos)	
Saúde mental instável por ≥14 dias		
Saúde física instável por ≥14 dias		
Acidente vascular encefálico		
Todos os dentes perdidos entre adultos com idade ≥65 anos		

O número de cidades por estado varia de 1 a 121. E as cidades variam de 42.417 pessoas em Burlington (Vermont) a 8.175.133 em Nova York (Nova York). Entre a 500 cidades, existem aproximadamente 28.000 setores censitários, para os quais foram fornecidos dados. Os intervalos para os setores variam em população de menos de 50 a 28.960, e em tamanho de menos de 1 milha quadrada a mais de 642 milhas quadradas. O número de setores por cidade varia de 8 a 2.140.

O projeto 500 Cities inclui uma população total de 103.020.808, o que representa 33,4% da população total dos Estados Unidos de 308.745.538.

A base de dados tem um total de 117 dimensões para um total de 500 entradas (cidades).

Existem 3 categorias principais:

- Indicadores de Saúde;
- Prevenção;
- Comportamentos não saudáveis.

http://dx.doi.org/10.25286/repa.v3i3.966

Cada uma dessas categorias é dividida em subcategorias para as quais são fornecidas 4 medições:

- Dados brutos;
- Dados ajustados por idade;
- Nível de confiança de 95% bruto;
- Nível de confiança de 95% ajustado por idade.

A maioria das porcentagens é definida como a proporção de entrevistados com idade superior ou igual a 18 anos que responderam "sim" a uma determinada pergunta de um questionário, sobre entrevistados com idade superior ou igual a 18 anos que relataram "sim" ou "não" (excluindo aqueles que se recusaram a responder, ou que responderam "não sei / não tenho certeza").

Neste trabalho, serão considerados apenas os dados ajustados por idade.

3.1.1 Dicionário de Dados

Como especificado no tópico anterior (3.1), a base possui 117 atributos, dos quais 112 estão relacionados às categorias e suas subcategorias. Para cada atributo são fornecidas 4 medições, que apresentam a seguinte configuração:

- Dados brutos: seu tipo é Numeric, com tamanho máximo igual a 4. Aceita o seguinte formato de valores: 00.0 ;
- Dados ajustados por idade: segue o mesmo padrão de Dados brutos;
- Nível de confiança de 95% bruto: seu tipo é String, com tamanho máximo igual a 12. Aceita o seguinte formato de valores: "(ponto mais inferior, ponto mais superior)";
- Nível de confiança de 95% ajustado por idade: segue o mesmo padrão do nível de confiança bruto

Os demais 5 atributos estão relacionados a:

- Estado: do tipo String, relacionado à abreviação do Estado, com apenas 2 caracteres;
- Nome da cidade: do tipo String do lugar ou subdivisão do município;
- Código FIPS: do tipo Numeric, com no máximo 7 caracteres. Relacionado Código FIPS do estado + código FIPS de subdivisão do local ou do condado;
- População: do tipo Numeric, com no máximo 8 caracteres. Dados retirados do Censo 2010 dos EUA;
- Geolocalização: do tipo String, informando latitude e longitude do local.

O Dicionário de Dados completo pode ser visualizado em Anexos.

3.1.2 Distribuição de Frequência

Abaixo seguem as tabela que contém um resumo dos dados obtido em uma amostra. A distribuição é organizada em formato de tabela, e cada entrada da tabela contém a frequência dos dados em um determinado intervalo, ou em um grupo.

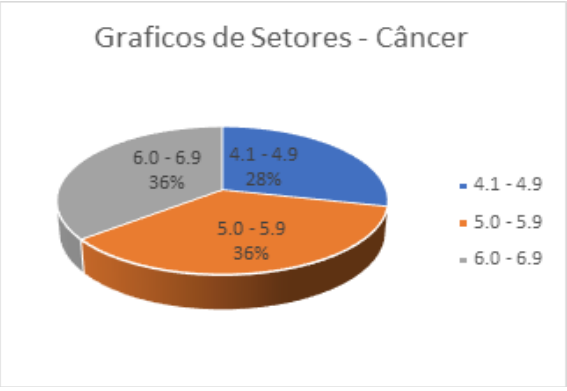
Câncer (excluindo câncer de pele) entre adultos com idade> = 18 anos - 2015					
Classe	Limite Inferior	Ponto Médio	Limite Superior	Frequência Absoluta	Amplitude de classe
1	4,1	4,5	4,9	8	0,8
2	5	5,45	5,9	10	0,9
3	6	6,45	6,9	10	0,9

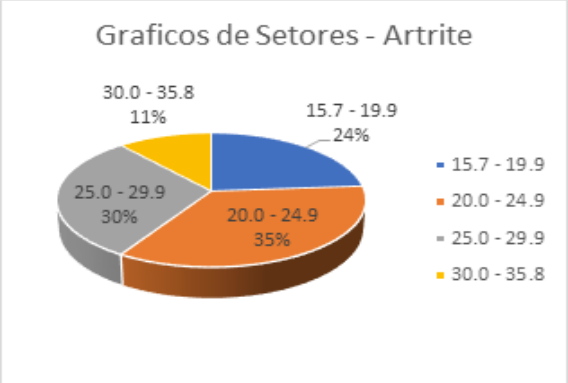
Artrite entre adultos com idade> = 18 anos - 2015					
Classe	Limite Inferior	Ponto Médio	Limite Superior	Frequência Absoluta	Amplitude de classe
1	15,7	17,8	19,9	34	4,2
2	20	22,45	24,9	50	4,9
3	25	27,45	29,9	43	4,9
4	30	32,9	35,8	16	5,8

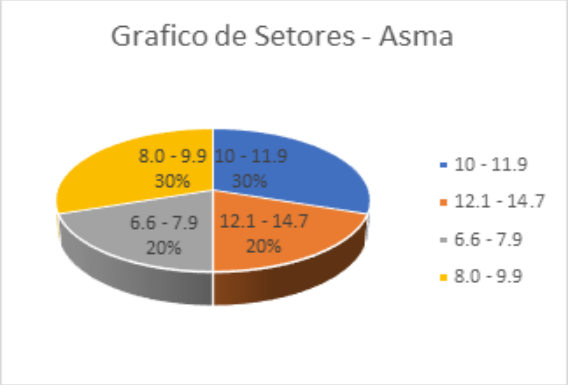
Asma atual entre adultos com idade> = 18 anos - 2015					
Classe	Limite Inferior	Ponto Médio	Limite Superior	Frequência Absoluta	Amplitude de classe
1	6,6	7,25	7,9	13	1,3
2	8	8,95	9,9	20	1,9
3	10	10,95	11,9	20	1,9
4	12,1	13,4	14,7	13	2,6

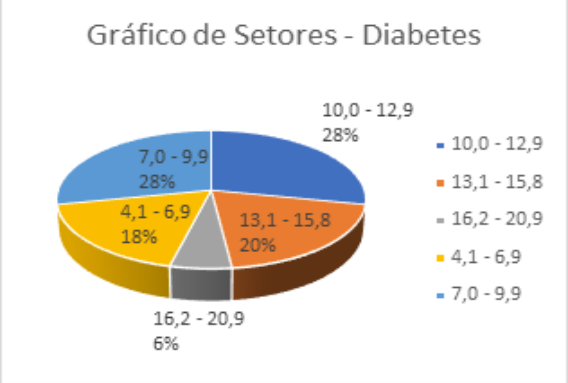
Diabetes diagnosticado entre adultos com idade> = 18 anos - 2015					
Classe	Limite Inferior	Ponto Médio	Limite Superior	Frequência Absoluta	Amplitude de classe
1	4,1	5,5	6,9	19	2,8
2	7	8,45	9,9	30	2,9
3	10	11,45	12,9	30	2,9
4	13,1	14,45	15,8	21	2,7
5	16,2	18,55	20,9	6	4,7

3.1.3 Visualização dos dados









3.1.4 Medidas de Resumo

Porcentagem - Diabetes	
Moda	8,9
Mediana	10,3
Média	10,5

Porcentagem - Artrite	
Moda	22,4
Mediana	23,65
Média	23,9

Porcentagem - Câncer	
Moda	6
Mediana	5.6
Média	5,55

Porcentagem - Asma	
Moda	9
Mediana	14,65
Média	10,1

3.2 Parametrização das técnicas

1. K-Means:
a. K variando de 2 a 13
b. Inicialização aleatória dos centróides
2. *Fuzzy C-Means*:
a. K variando de 2 a 13
b. *Fuzzificador* m igual a 2
3. Affinity Propagation
a. *Preference* igual a -50
4. Agglomerative Clustering
a. K variando de 2 a 13
b. Tipo de ligação: média
5. Birch
a. K variando de 2 a 13
6. Mean Shift
a. *Bandwidth* igual a 0,81
7. Spectral Clustering
a. K variando de 2 a 13
b. Eigen solver: arpack
c. Afinidade: *nearest_neighbors*
8. PSC
a. Quantidade de partículas (K) variando de 2 a 13
b. Número de iterações igual a 1000
c. Inércia igual a 0,95
d. c1=c2=2,05
e. c3=c4=1,0
f. Velocidade máxima das partículas: 0,001
9. PSOC
a. K variando de 2 a 13
b. Quantidade de partículas igual a 30
c. Número de iterações igual a 1000
d. Inércia igual a 0,72,
e. c1=c2=1,49

4 Experimentos realizados

Cada algoritmo de clusterização foi executado 30 vezes para cada conjunto de parâmetros para observação das métricas de Silhueta e Calinski-Harabasz. As duas métricas foram utilizadas para a escolha do melhor número de clusters para o problema. A Tabela 1 mostra as estatísticas para os algoritmos com o melhor número de clusters baseado nas duas métricas.

Tabela 1 - Estatísticas das métricas

Algoritmo	Silhueta	Calinski-

http://dx.doi.org/10.25286/repa.v3i3.966

(n_clusters)	(desvio-padrão)	Harabasz(desvio-padrão)
Fuzzy C-Means **(2)**	0.27399 **(4.53246e-17)**	247.8647 **(3.476e-17)**
PSOC **(2)**	0.27398 **(7.3687e-18)**	247.845 **(0.007212)**
PSC **(2)**	0.27228 **(0.00046)**	229.572 **(14.57811)**
K-Means **(2)**	0.27106 **(2.67750e-05)**	247.85433 **(0.00572)**
Spectral **(2)**	0.25755 **(5.55111e-17)**	231.66074 **(8.52651e-14)**
Birch **(2)**	0.25136 **(0.0)**	196.50644 **(0.0)**
Affinity **(3)**	0.16760 **(5.55111e-17)**	163.81091 **(2.84217e-14)**
Agglomerative **(3)**	0.27787 **(0.0)**	43.16529 **(0.0)**
Mean Shift **(5)**	0.14607 **(2.77555e-17)**	45.52593 **(1.42108e-14)**

Os algoritmos com os melhores valores para as métricas distribuíram os dados em dois clusters. O *Affinity Propagation* e o *Agglomerative Clustering* indicaram a divisão em três clusters, porém com valores menores para as métricas. O *Agglomerative Clustering* teve um resultado bom comparando a Silhueta, mas um resultado ruim quando analisando o índice de *Calinski-Harabasz*. O *Mean Shift* por sua vez em seu melhor resultado agrupou os dados em 5 clusters e teve o pior resultado dentre os algoritmos testados nessa base de dados, considerando as duas métricas utilizadas nesse trabalho.
A visualização geográfica dos clusters para os resultados do K-Means e do AffinityPropagation pode ser observada nas figuras a seguir.

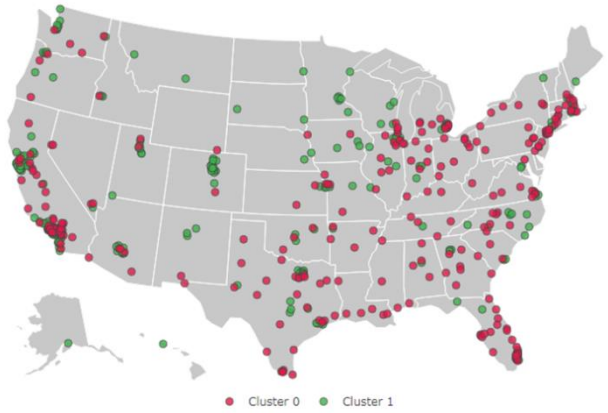


Figura - K-Means

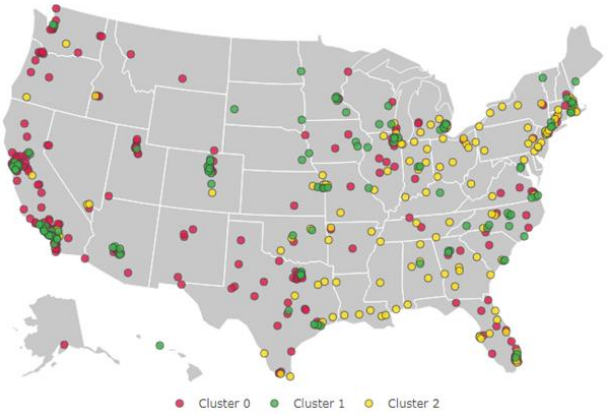


Figura – *Affinity Propagation*

Para analisar mais profundamente o que cada cluster significa a matriz de correlação para cada cluster pode ser útil para inferir as relações entre as diferentes dimensões do problema. Por exemplo, para o K-Means, temos as seguintes matrizes de correlação:

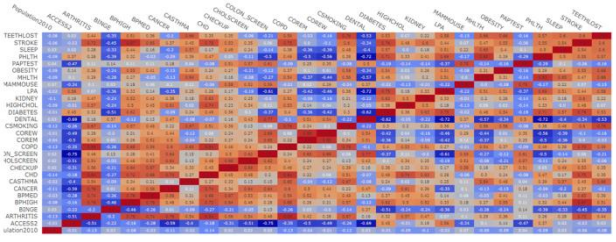


Figura - Matriz de correlação Cluster 0 (K-Means)

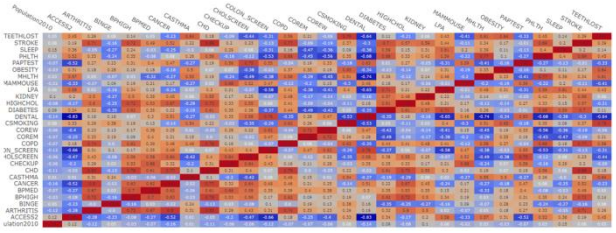


Figura - Matriz de correlação Cluster 1 (K-Means)

Juntamente com as matrizes de correlação, os centróides de cada cluster indicam as principais características dos clusters.

O K-Means, por exemplo, agrupou os clusters em cidades saudáveis (Cluster 1) e não-saudáveis (Cluster 0). As principais variáveis para considerar uma cidade saudável foram:

- Menores índices de doenças (artrite, asma, doenças cardíacas, doenças pulmonares, diabetes, colesterol alto e doenças renais);
- Menores índices de maus hábitos (alcoolismo, tabagismo, sedentarismo);
- Melhores índices de prevenção (controle de pressão alta, colonoscopia, exames de rotina para homens e mulheres, visitas ao dentista).

Dentre as 10 cidades mais populosas presentes na base de dados, temos o seguinte agrupamento:

- Cidades saudáveis: San Diego, Honolulu;
- Cidades não-saudáveis: Nova York, Los Angeles, Chicago, Houston, Philadelphia, Phoenix, San Antonio e Dallas.

As principais características das cidades de San Diego e Honolulu em relação às outras é o alto índice de acesso a planos de saúde, visitas periódicas ao dentista e a prática de atividades físicas de lazer regulares.

Por sua vez, o *Affinity Propagation* agrupou os dados em três clusters: cidades saudáveis (Cluster 1), cidades não-saudáveis (Cluster 0) e cidades críticas (Cluster 2). As diferenças principais entre as cidades não-saudáveis e as cidades críticas são que as cidades não-saudáveis tem menores índices de prevenção, porém as cidades críticas tem índices elevados na maioria das doenças (exceto câncer) e em todos os maus hábitos (alcoolismo, tabagismo, sedentarismo, obesidade e poucas horas diárias de sono).

Dentre as 10 cidades mais populosas presentes na base de dados, temos o seguinte agrupamento:

- Cidades saudáveis: Honolulu
- Cidades não-saudáveis: Nova York, Los Angeles, Chicago, Houston, Phoenix, San Antonio, Dallas e San Diego
- Cidades críticas: Philadelphia

Em relação ao resultado do K-Means, a cidade de Philadelphia passou a ser classificada como crítica e a cidade de San Diego passou a ser considerada como não saudável. Entretanto, ao realizar uma análise de similaridade entre as cidades de Honolulu e San Diego, as duas são mais similares do que San Diego e Nova York, por exemplo. Isso pode indicar que San Diego foi erroneamente agrupada pelo AffinityPropagation. Tendo em vista que o AffinityPropagation teve resultados de métrica piores que o K-Means, este pode ser um bom indício.

5 Conclusões e trabalhos futuros

A saúde pública é uma área com problemas bastante complexos e que os órgãos governamentais frequentemente enfrentam desafios para entender como oferecer melhores serviços de saúde e prevenir epidemias futuras. Os métodos preventivos normalmente são a melhor para controlar ou extinguir doenças, tornado importante saber quais os hábitos de saúde das populações das cidades e quais impactos tais hábitos poderiam ter sobre as condições gerais de saúde das populações.

O projeto 500 Cities reporta dados epidemiológicos das 500 maiores cidades americanas e baseia-se em doenças crônicas prioritárias e de maior impacto na saúde pública. As medidas reportadas incluem os índices de doenças, comportamentos de maior risco que causam doenças e práticas de prevenção.

Este trabalho procurou identificar características relevantes para dar suporte na prevenção de epidemias e doenças e agrupar as diferentes cidades de acordo com seus hábitos e condições de saúde. Essa detecção de comunidades foi feita utilizando vários algoritmos de clusterização para identificar os principais agrupamentos e suas principais características.

A maioria dos algoritmos agrupou as cidades em dois clusters, sendo um com bons indicadores

http://dx.doi.org/10.25286/repa.v3i3.966

de saúde e outro com indicando cidades não-saudáveis. Outros algoritmos subdividiram o grupo de cidades não-saudáveis para incluir o grupo de cidades críticas (com péssimo indicadores). Porém, essa divisão em mais agrupamentos levou a uma diminuição da avaliação dos clusters, o que pode indicar um aumento de cidades sendo erroneamente agrupadas.

Trabalhos futuros que possam agregar informações relevantes ao presente trabalho incluem:

- Estudar como utilizar efetivamente a informação de geolocalização no processo de clusterização. Métricas euclidianas tendem a não ser efetivas com dados de latitude e longitude;
- Analisar o comportamento de outras métricas de avaliação de agrupamentos (principalmente a estatística Gap, mas possivelmente outras como índice de Xu, índice de Hartigan);
- Utilizar outros algoritmos de agrupamento que utilizem outras abordagens para cálculo de similaridade, como os algoritmos que utilizam conceitos de Teoria da Informação e que não dependam da distância euclidiana.

Referências

[1] CENTERS FOR DISEASE CONTROL AND PREVENTION. **500 Cities**: local data for better health. National Center for Chronic Disease Prevention and Health, Promotion, Division of Population Health, 2016. Disponível em: <https://www.cdc.gov/500cities> Acesso em: 23 out. 2017.

[2] MACQUEEN, James B. et al. Some Methods for classification and Analysis of Multivariate Observations. In: Berkeley Symposium on Mathematical Statistics and Probability, 5., 1967, California. **Proceedings**…Berkeley, CA: University of California, Press. p. 281–297.

[3] BEZDEK, James C. **Pattern Recognition with Fuzzy Objective Function Algorithms**. New York, London: Plenum Press, 1981.

[4] FREY, Brendan J.; DUECK, Delbert. Clustering by passing messages between data points, **Science**, v. 315, n. 5814, p. 972-976, 2007.

[5]COMANICIU, Dorin; MEAN, Peter. Shift: A robust approach toward feature space analysis. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 24, n.5, p. 603-619, 2002.

[6] NG, Andrew Y.; JORDAN, Michael I.; WEISS, Yair. On spectral clustering: Analysis and an algorithm. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 15., 2002. **Proceedings**... Press,2002. p. 849-856.

[7] ZHANG, et al. Graph degree linkage: Agglomerative clustering on a directed graph. In: EUROPEAN CONFERENCE ON COMPUTER VISION, 12., 2012, Florence. **Proceedings**… Florence, IT, 2012.

[8] ZHANG, Tian; RAMAKRISHNAN, Raghu; LIVNY, Miron. BIRCH: an efficient data clustering method for very large databases. In: **ACM Sigmod Record**. ACM, 1996. p. 103-114.

[9] KENNEDY, James. Particle swarm optimization. In: Encyclopedia of machine learning, p. 760-766. Springer, 2011.

[10] MERWE, D. W. van der; ENGELBRCHT, A. P. Data clustering using particle swarm optimization. In: Congress on Evolutionary Computation, 2003. **Proceedings**… 2003.

[11] COHEN, Sandra C. M.; CASTRO Leandro N. de. Data Clustering with Particle Swarms. In: IEEE Congress on Evolutionary Computations. 2006.

[12] ROUSSEEUW Peter J. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. **Computational and Applied Mathematics**, v. 20, p. 53-65, 1987.

[13] CALIŃSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. **Communications in Statistics-theory and Methods**, v. 3, p. 1-27, 1974.

http://dx.doi.org/10.25286/repa.v3i3.966