UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Estudo de um caso de uso dos algoritmos de estimação de distribuição

Celso Oviedo da Silva Lopes

Trabalho de Conclusão de Curso
MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Estudo de um caso de uso dos algoritmos de estimação de distribuição

Celso Oviedo da Silva Lopes

USP - São Carlos

2023

Celso Oviedo da Silva Lopes

Estudo de um caso de uso dos algoritmos de estimação de distribuição

Trabalho de conclusão de curso apresentado ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Prof. Dr. Alexandre C. B. Delbem

USP - São Carlos

2023

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

O96e

Oviedo da Silva Lopes, Celso
Estudo de um caso de uso dos algoritmos de estimação de distribuição / Celso Oviedo da Silva Lopes; orientador Alexandre Claudio Botazzo Delbem. -- São Carlos, 2023.
57 p.

Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. Algoritmos de estimação de distribuição. 2. DAMiCORE. I. Claudio Botazzo Delbem, Alexandre, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

AGRADECIMENTOS

À minha filha pelo apoio e seu companheirismo.

Ao meu orientador Profº Alexandre Cláudio Botazzo Delbem, pelo apoio, amizade e enorme dedicação ao ensino.

Aos amigos do laboratório (LCR): Renata e Erick pela amizade e por sempre pararem para uma boa discussão.

Ao programa de MBA do ICMC, pela oportunidade de participar e em especial à Profª Solange O. Rezende pelo empenho em me incentivar a entrega desse TCC.

RESUMO

Lopes, C. O. S. **Estudo de um caso de uso dos algoritmos de estimação de distribuição**. 2023. 56f. Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

Resumo:

Os desafios inerentes à Segurança Alimentar e Nutricional se apresentam como obstáculos substanciais na busca por compreensão e solução, especialmente quando a complexidade das variáveis envolvidas e sua relação com os fenômenos em análise carecem de conhecimento prévio. Um exemplo paradigmático desse contexto desafiador é a intrincada questão da garantia de alimentação adequada em megacidades, em que os preços dos alimentos em alta, a inflação crescente e os impactos da pandemia de Covid-19 contribuem para agravar o problema. Desertos alimentares são áreas geográficas onde o acesso a alimentos saudáveis e de qualidade é limitado. Eles são frequentemente encontrados em áreas de baixa renda e com alta população, onde os moradores têm menos acesso a transporte e recursos financeiros. Assim, para auxiliar na análise e identificação de desertos alimentares, esse trabalho tem como objetivo utilizar técnicas de algoritmos evolutivos, mais especificamente algoritmos de estimação de distribuição em dados complexos, partindo de dados obtidos de plataformas governamentais, que fornecem informações confiáveis relacionadas ao problema, para obtermos uma amostra rápida da situação.

Palavras-chave: Desertos alimentares; algoritmos genéticos; algoritmos de estimação de distribuição; DAMICORE.

ABSTRACT

Lopes, C. O. S. **A Study of a Use Case for Estimation Distribution Algorithms**. 2023. 57 f. Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

The challenges inherent to Food and Nutritional Security present themselves as substantial obstacles in the quest for understanding and solutions, especially when the complexity of the variables involved and their relation to the phenomena under analysis lack prior knowledge. A paradigmatic example of this challenging context is the intricate issue of ensuring adequate food supply in megacities, where high food prices, increasing inflation, and the impacts of the Covid-19 pandemic contribute to exacerbating the problem. Food deserts are geographical areas where access to healthy and quality food is limited. They are often found in low-income and densely populated areas, where residents have less access to transportation and financial resources. Therefore, to assist in the analysis and identification of food deserts, this work aims to use evolutionary algorithm techniques, specifically distribution estimation algorithms in complex data, based on data obtained from government platforms that provide reliable information related to the problem, to obtain a quick sample of the situation.

Keywords: Food deserts; genetic algorithms; distribution estimation algorithms; DAMICORE.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

Sumário

1- INTRODUÇÃO

A fome é um problema global que afeta milhões de pessoas em todo o mundo. De acordo com a Organização das Nações Unidas para Alimentação e Agricultura (FAO), estima-se que entre 691 e 783 milhões de pessoas no mundo enfrentaram a fome em 2022. Considerando a faixa média (cerca de 735 milhões), 122 milhões de pessoas a mais enfrentaram a fome em 2022 do que em 2019, antes da pandemia (The State of the World , 2023).

Ela é mais prevalente em países em desenvolvimento, onde os sistemas agrícolas são menos eficientes e as desigualdades sociais são maiores. Nos países desenvolvidos, a fome também é um problema, mas é geralmente associada a fatores como pobreza, insegurança alimentar e doenças crônicas.O estado de São Paulo é um dos mais desenvolvidos do Brasil, mas ainda enfrenta problemas de segurança alimentar e nutricional. De acordo com dados da Pesquisa Nacional de Saúde (PNS), realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE) em 2019, 2,9% das pessoas com 18 anos ou mais no estado estavam em situação de insegurança alimentar grave (IBGE, 2019)(SAN, 2019), ou seja, não tinham acesso aos alimentos em quantidade suficiente para uma alimentação saudável.

De acordo com um estudo da Rede Brasileira de Pesquisa em Soberania e Segurança Alimentar e Nutricional (Rede PENSSAN), os níveis de Insegurança Alimentar (IA) são graves em 14,6% dos domicílios do estado de São Paulo, e 1,2 milhão de pessoas vivem em áreas de São Paulo onde o acesso a alimentos saudáveis é limitado. As estatísticas foram coletadas entre novembro de 2021 e abril de 2022, a partir da realização de entrevistas em 12.745 domicílios, em áreas urbanas e rurais de 577 municípios, distribuídos nos 26 estados e no Distrito Federal. A Segurança Alimentar e a Insegurança Alimentar foram medidas, mais uma vez, pela Escala Brasileira de Insegurança Alimentar (Ebia), que também é utilizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE).

A fome tem um impacto devastador na saúde e no bem-estar das pessoas. Ela pode levar a uma série de problemas de saúde, incluindo desnutrição, anemia, retardo do crescimento e desenvolvimento, e doenças infecciosas. A fome também pode causar problemas psicológicos, como ansiedade, depressão e irritabilidade (The State of the World , 2022).

Existem várias causas da fome. Entre as principais causas podemos incluir:

- Pobreza: A pobreza é a principal causa da fome. As pessoas que vivem na pobreza geralmente não têm recursos suficientes para comprar alimentos suficientes ou nutritivos.
- Insegurança alimentar: A insegurança alimentar é a falta de acesso confiável a alimentos nutritivos. Ela pode ser causada por fatores como conflitos, desastres naturais, mudanças climáticas e políticas governamentais inadequadas.
- Doenças crônicas: As doenças crônicas, como diabetes e hipertensão, podem dificultar a absorção de nutrientes e aumentar a necessidade de alimentos.

O combate à fome é um desafio global que requer esforços coordenados de governos, organizações não governamentais e indivíduos. Com ações concertadas, é possível erradicar a fome e garantir que todas as pessoas tenham acesso a alimentos suficientes e nutritivos.

Aqui estão alguns exemplos específicos de como a fome pode afetar a saúde e o bem-estar das pessoas:

Desnutrição: A desnutrição é uma condição em que o corpo não recebe os nutrientes de que precisa para funcionar adequadamente. Ela pode causar uma série de problemas de saúde, incluindo retardo do crescimento e desenvolvimento, problemas de aprendizagem, problemas de visão, problemas de pele e aumento da suscetibilidade a infecções.

Anemia: A anemia é uma condição em que o corpo não tem glóbulos vermelhos suficientes para transportar oxigênio para os tecidos. Ela pode causar fadiga, falta de ar, tontura e irritabilidade.

Retardo do crescimento e desenvolvimento: O retardo do crescimento e desenvolvimento é um atraso no crescimento físico e mental. Ele pode ser causado por uma série de fatores, incluindo desnutrição, doenças crônicas e exposição a toxinas.

Doenças infecciosas: As pessoas que estão desnutridas são mais suscetíveis a doenças infecciosas, como pneumonia, diarreia e malária.

A fome também pode causar problemas psicológicos, como ansiedade, depressão e irritabilidade. Isso ocorre porque a fome pode causar estresse e medo.

A complexidade desse problema envolve múltiplos domínios de conhecimento, cada um estudando um conjunto específico de fatores relacionados à avaliação da situação atual, simulação de cenários futuros e formulação de políticas públicas para reduzir seus impactos. Contudo, na maioria dos casos, os especialistas nos diferentes domínios de conhecimento não possuem uma linguagem comum e uma metodologia compartilhada para interpretar dados e propor políticas potenciais.

Existem várias coisas que podem ser feitas para combater a fome. Algumas das principais ações incluem:

- Desenvolvimento agrícola: O desenvolvimento agrícola é essencial para aumentar a produção de alimentos e reduzir o custo dos alimentos.
- Redução da pobreza: A redução da pobreza é fundamental para garantir que as pessoas tenham recursos suficientes para comprar alimentos.
- Melhoria do acesso a alimentos: O acesso a alimentos pode ser melhorado por meio de programas de transferência de renda, programas de alimentação escolar e programas de segurança alimentar.

1.1– Ações governamentais

Nesta seção, serão abordadas as ações governamentais destinadas a implementar um conjunto de estratégias, políticas e medidas com o propósito de proporcionar alimentos saudáveis às populações de baixa renda.

1.1.1– Alimentos in natura, RAIS e CAISAN

Em São Paulo, a Secretaria de Agricultura e Abastecimento do Estado (SAA) é responsável pela gestão do CAISAN (Câmara Interministerial de Segurança Alimentar e Nutricional)(Portal SAN, 2023)(SAGE, 2023). A SAA divulga anualmente um catálogo de produtos da agricultura familiar que podem ser adquiridos por meio do CAISAN.

O CAISAN é um programa de compras públicas que destina recursos para a aquisição de alimentos da agricultura familiar. Isso fortalece a agricultura familiar e garante o acesso a alimentos de qualidade e a preços acessíveis para a população. Os recursos do CAISAN são utilizados por entidades públicas, como prefeituras, governos estaduais e escolas públicas.

Há um outro programa do governo federal chamado RAIS (Relação Anual de Informações Sociais). A RAIS é obrigatória para todas as empresas com mais de 20 funcionários. A empresa deve informar ao governo quais alimentos comprou da agricultura familiar no ano anterior.

Os estabelecimentos com alimentos in natura e os programas como RAIS e CAISAN são importantes para promover a segurança alimentar e nutricional.

Há diversas iniciativas que podem ser feitas pelo setor público para intervir no sistema alimentar, que vão desde o fomento à produção de alimentos agroecológicos e orgânicos, até a utilização de instrumentos econômicos e medidas fiscais, como a taxação de produtos alimentícios com alto teor de gorduras saturadas, açúcar e sal. Tais iniciativas passam também pelos incentivos à criação de ambientes saudáveis nos comércios varejistas e serviços de alimentação. Os varejistas e os pontos de venda de alimentos, como supermercados, mercearias, feiras, restaurantes, lojas de conveniência, dentre outros, são os meios pelos quais a maior parte dos consumidores acessam os alimentos. Assim, tais estabelecimentos são, potencialmente, pontos de oferta de uma dieta mais saudável e é neles que os responsáveis pelas políticas podem intervir (Painel Global, 2017).

Dessa forma, na tentativa de mapear e entender melhor o contexto do comércio varejista de alimentos no Brasil e sua distribuição geográfica, a CAISAN desenvolveu uma metodologia para mapear desertos alimentares no país.

Os desertos alimentares são áreas onde há pouca ou nenhuma oferta de alimentos saudáveis e acessíveis. Eles são um problema sério de segurança alimentar e nutricional, pois dificultam o acesso a uma alimentação saudável para a população.

A metodologia desenvolvida pela CAISAN baseia-se na análise de dados de censos demográficos, dados de comércio varejista de alimentos e dados de infraestrutura urbana. Esses dados são utilizados para identificar as áreas onde há pouca oferta de alimentos saudáveis e acessíveis.

Ela pode ajudar a identificar os desertos alimentares e orientar o desenvolvimento de políticas públicas para mitigar esse problema.

Neste trabalho, vamos utilizar parte dos dados gerados por esta base em nossos estudos.

1.1.2 - IPVS - Índice Paulista de Vulnerabilidade Social

O IPVS (IPVS, 2023) é um indicador sintético que classifica todos os setores censitários do Estado de São Paulo em 6 grupos, segundo dimensões socioeconômicas e demográficas.

O IPVS é calculado com base em dados do Censo Demográfico do Instituto Brasileiro de Geografia e Estatística (IBGE) e do Sistema Integrado de Informações sobre Comércio Exterior (SISCOMEX).

As dimensões socioeconômicas e demográficas utilizadas para o cálculo do IPVS são:

- Condições de vida: renda média, escolaridade, saneamento básico, habitação, transporte, saúde e segurança pública.
- Características demográficas: estrutura etária, composição familiar e migração.

O IPVS é um importante instrumento para a identificação de áreas com maior vulnerabilidade social. Ele pode ser utilizado para orientar o desenvolvimento de políticas públicas e ações de intervenção social.

O IPVS é um indicador sintético que combina diferentes dimensões socioeconômicas e demográficas (Tabela 1). Isso permite que ele seja mais sensível à realidade da vulnerabilidade social.

Tabela 1: IPVS Índice Paulista de Vulnerabilidade Social

	IPVS
Grupo 1	Baixíssima vulnerabilidade
Grupo 2	Vulnerabilidade muito baixa
Grupo 3	Vulnerabilidade baixa
Grupo 4	Vulnerabilidade média - setores urbanos
Grupo 5	Vulnerabilidade alta - setores urbanos
Grupo 6	Vulnerabilidade muito alta - aglomerados subnormais
Grupo 7	Vulnerabilidade alta - setores rurais

Fonte: autoria própria

É um instrumento importante para a promoção da equidade social no Estado de São Paulo.

1.1.3 – Feiras livres

As feiras livres são uma importante fonte de alimentos saudáveis e acessíveis para a população, especialmente para as pessoas que vivem em áreas com desertos alimentares.

As feiras livres oferecem uma variedade de alimentos, incluindo frutas, legumes, verduras, carnes, aves, peixes, laticínios e produtos orgânicos. Os alimentos vendidos nas feiras livres são geralmente mais frescos e de melhor qualidade do que os alimentos encontrados em supermercados e outros estabelecimentos comerciais. Além disso, os preços dos alimentos nas feiras livres são geralmente mais acessíveis.

A presença de feiras livres em áreas com desertos alimentares pode contribuir para a redução do problema. As feiras livres podem oferecer à população uma alternativa para comprar alimentos saudáveis e acessíveis.

Além disso, as feiras livres podem contribuir para a promoção da agricultura familiar e da economia local. As feiras livres são um importante canal de comercialização para os agricultores familiares, que vendem seus produtos diretamente aos consumidores. Isso contribui para o fortalecimento da agricultura familiar e da economia local.

A seguir, são apresentados alguns benefícios das feiras livres para as áreas com desertos alimentares:

- Oferta de alimentos saudáveis e acessíveis: as feiras livres oferecem uma variedade de alimentos saudáveis e acessíveis, o que pode contribuir para a melhoria da alimentação da população.
- Promoção da agricultura familiar: as feiras livres são um importante canal de comercialização para os agricultores familiares, o que pode contribuir para o fortalecimento da agricultura familiar.
- Promoção da economia local: as feiras livres movimentam a economia local, gerando emprego e renda para a população.

De acordo com o Relatório do Plano Estadual de Segurança Alimentar e Nutricional Sustentável (CONSEA, 2023), as principais ações para mitigar os problemas de segurança alimentar e nutricional no estado de São Paulo incluem:

- Integrar políticas humanitárias, de desenvolvimento e de construção da paz em áreas afetadas por conflitos.
- Aumentar a adaptação climática em todos os sistemas alimentares.
- Fortalecimento da adaptação dos mais vulneráveis à adversidade econômica.

- Intervir ao longo das cadeias de abastecimento alimentar para reduzir o custo dos alimentos nutritivos.
- Combater a pobreza e as desigualdades estruturais, garantindo que as intervenções sejam inclusivas e em favor dos menos favorecidos economicamente.

1.2 - Computação Evolutiva

A Computação Evolutiva engloba um conjunto de Algoritmos Evolutivos (EAs), uma classe de algoritmos de otimização, fundamentados na Teoria da Evolução das Espécies proposta por Darwin (Darwin, 1859), que afirma que as espécies evoluíram ao longo do tempo por meio de um processo de seleção natural. Pesquisadores na área de Biologia têm demonstrado interesse nos EAs para a compreensão dos processos evolutivos e a simulação de modelos de evolução natural. Ao mesmo tempo, cientistas da computação e engenheiros têm investigado o desenvolvimento desses algoritmos e explorado seu potencial para resolver diversos problemas considerados complexos (DeJong, 2008).

Os EAs podem ser úteis para auxiliar neste estudo, porque são capazes de explorar grandes espaços de busca de forma eficiente, o que pode ser importante para identificar áreas com acesso limitado a alimentos saudáveis.

Os EAs funcionam da seguinte forma:

1. Criação de uma população inicial: Uma população inicial é composta por um conjunto de soluções aleatórias ou baseadas em conhecimento prévio sobre o problema.
2. Avaliação das soluções: Cada solução da população é avaliada por uma função de aptidão, que atribui a cada solução uma pontuação que indica sua qualidade.
3. Seleção das soluções: As soluções com maior aptidão são selecionadas para participarem do processo de reprodução.
4. Reprodução: Novas soluções são criadas a partir das soluções selecionadas.
5. Evolução: Os passos 2 a 4 são repetidos até que seja encontrada uma solução aceitável ou até que seja atingida uma condição de parada.

Os operadores de reprodução mais comuns são a recombinação e a mutação.

- Recombinação: A recombinação consiste na combinação de duas ou mais soluções para criar uma nova solução.
- Mutação: A mutação consiste na alteração de uma ou mais características de uma solução.

Os EAs podem ser classificados em dois tipos principais:

- EAs canônicos: Os EAs canônicos, como os Algoritmos Genéticos (GAs), utilizam operadores de recombinação e mutação simples.
- EAs baseados em modelos probabilísticos: Os EAs baseados em modelos probabilísticos, como os Algoritmos de Estimação de Distribuição (EDAs), utilizam modelos probabilísticos para gerar novas soluções.

1.3 - As metaheurísticas baseadas em processos evolutivos

Segundo (Schwefel, 1996), as metaheurísticas são técnicas de otimização que não garantem a solução ótima, mas podem encontrar soluções de boa qualidade para problemas complexos, geralmente utilizadas para resolver problemas difíceis, onde métodos tradicionais podem falhar.

Processos evolutivos referem-se a abordagens inspiradas na evolução natural, como algoritmos genéticos, algoritmos evolutivos, programação evolutiva, entre outros.

O EDA refere-se a uma classe de algoritmos evolutivos que utiliza modelos probabilísticos para estabelecer e atualizar as distribuições de probabilidade das variáveis de decisão. Em vez de representar soluções individuais, como cromossomos em algoritmos genéticos, os EDAs modelam e atualizam as distribuições de probabilidade das variáveis relevantes para o problema.

As metaheurísticas evolutivas muitas vezes incluem técnicas de estimação de distribuição para guiar a busca no espaço de solução de maneira mais eficiente. São técnicas usadas para encontrar soluções aproximadas para problemas de otimização global, explorando um espaço de busca de soluções.

1.4 - Dados complexos e heterogêneos multifontes

Dados complexos e heterogêneos multifontes são dados que são coletados de diferentes fontes, têm diferentes formatos, estruturas e tipos de dados, e são difíceis de entender e analisar. Eles podem ser coletados de uma variedade de fontes, incluindo sensores, sistemas de informação, mídias sociais e pessoas.

Esses tipos de dados podem ser um desafio para lidar, pois podem ser difíceis de integrar, limpar, normalizar e analisar. No entanto, eles também podem oferecer uma riqueza de informações que não podem ser obtidas de dados de uma única fonte.

1.5 - Modelos probabilísticos de soluções candidatas

Os EDAs trabalham com modelos probabilísticos de soluções candidatas. Esses modelos representam a distribuição de probabilidade das soluções candidatas, ou seja, a probabilidade de uma solução candidata ser encontrada. Os modelos probabilísticos podem ser usados para gerar novas soluções candidatas, para avaliar a qualidade das soluções candidatas e para guiar a busca por soluções ótimas.

Os modelos probabilísticos possibilitam uma descrição mais precisa do comportamento dos EDAs. Isso é importante para análises formais dos EDAs, como estimativas de convergência e de complexidade de tempo.

Os EDAs têm sido bem sucedidos em vários problemas complexos e particularmente adequados para problemas com grande espaço de busca, como problemas de otimização combinatória.

A Clusterização baseada em Análise de Filograma (CPA) e o DAMICORE (DAta MIning of Code REpositories), que serão vistos com mais detalhes na Seção 2, são técnicas de clusterização que utilizam a análise de filogramas para agrupar objetos com características semelhantes.

Segundo (Raven et al, 2007), um filograma é uma representação gráfica de uma árvore filogenética, que é uma representação das relações evolutivas entre organismos. Os nós de um filograma representam os organismos e as arestas representam as relações de parentesco entre eles.

A CPA utiliza a estrutura de um filograma para identificar grupos de objetos que são mais semelhantes entre si do que a outros objetos. Para isso, a CPA calcula uma medida de similaridade entre os objetos, que pode ser baseada em diferentes características, como o conteúdo, a estrutura ou a função dos objetos.

o DAMICORE é um conjunto de algoritmos agnósticos, isto é, independentemente do tipo de dados que é fornecido (texto, imagens, números, por exemplo) graças ao NCD (Distância de Compressão Normalizada), uma métrica baseada em compressor que vê qualquer pedaço de dados simplesmente como uma sequência de bits. Isso é particularmente adequado para conjuntos de dados heterogêneos, conjuntos de dados com características difíceis de extrair e conjuntos de dados de texto.

O objetivo principal deste trabalho é tentar fornecer informações que identifiquem possíveis desertos alimentares que comprometam a segurança alimentar, sem informações a priori, considerando diversas variáveis provenientes de 3 diferentes bancos de dados, na cidade de São Paulo, bancos de dados esses que são multifontes complexas de dados heterogêneos.

2 - FUNDAMENTAÇÃO TEÓRICA

No mundo atual, impulsionado por dados, a capacidade de extrair informações significativas de grandes quantidades é essencial para enfrentar desafios globais complexos. Em particular, a capacidade de analisar dados de forma rápida e eficaz é crucial para tomar decisões informadas em áreas como alívio da fome, mitigação das mudanças climáticas e gestão da migração.

Um dos principais desafios na análise de dados é a identificação de padrões e relacionamentos em grandes conjuntos de dados. É aqui que as técnicas de modelagem estatística, como aprendizado de máquina e mineração de dados, desempenham um papel fundamental. Essas técnicas podem ser usadas para identificar variáveis potencialmente relacionadas, relacionamentos entre variáveis e outliers, fornecendo insights valiosos sobre a estrutura subjacente dos dados.

Por exemplo, no contexto do alívio da fome, algoritmos de aprendizado de máquina podem ser usados para analisar dados sobre produção de alimentos, distribuição e padrões de consumo para identificar regiões ou populações que são mais vulneráveis à insegurança alimentar. Essas informações podem então ser usadas para direcionar intervenções e alocar recursos de forma mais eficaz.

A Computação Evolutiva é uma área de pesquisa que estuda o desenvolvimento de Algoritmos Evolutivos (EAs), que são inspirados na teoria da evolução natural das espécies. Esses algoritmos são capazes de encontrar soluções para problemas complexos, como sistemas de redes elétricas, robótica, projetos de redes, jogos, predição de estruturas de proteínas, entre outros (Goldberg, 2002).

Cientistas de computação e engenheiros investigam o desenvolvimento desses algoritmos e seu potencial para resolver problemas complexos.

Os principais componentes de um sistema evolutivo são:

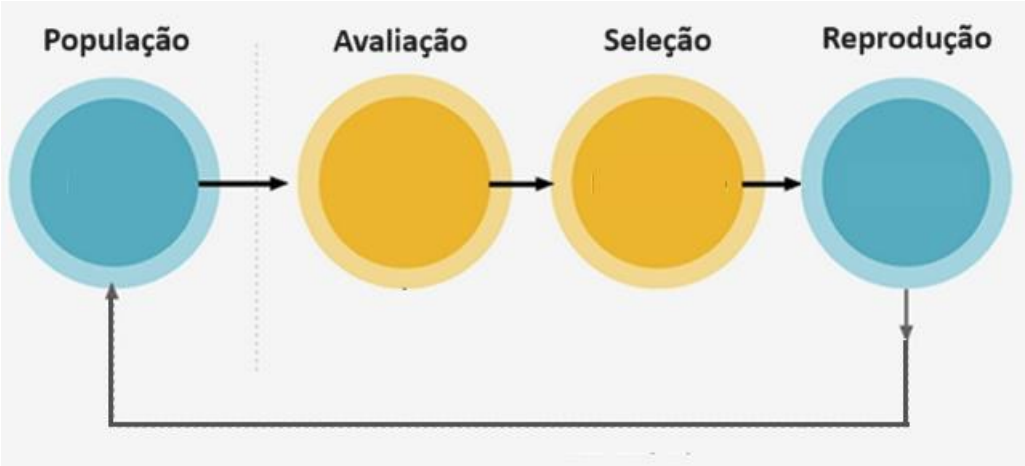
- Uma ou mais populações de indivíduos que competem por recursos limitados.
- A noção de mudanças dinâmicas nas populações devido ao nascimento e à morte dos indivíduos.
- O conceito de aptidão, que reflete a habilidade do indivíduo para sobreviver e se reproduzir.
- A variação na hereditariedade, ou seja, os novos indivíduos preservam características dos seus pais, embora não sejam iguais.

Um algoritmo evolutivo típico, conforme Figura 1, funciona da seguinte maneira:

- Inicialmente, uma população é formada gerando um conjunto inicial de vetores aleatórios que descrevem as características de um indivíduo.

- Em seguida, os vetores são avaliados de acordo com algum critério, atribuindo-se um nível de aptidão para cada indivíduo.
- Com base na aptidão, um subconjunto dos vetores é selecionado. A partir desses vetores, são gerados novos vetores (mimetizando o processo de reprodução) que compõem uma nova população.

Figura 1: Etapas de uma geração de um EA típico.



Fonte: autoria própria

Os EAs mais conhecidos são os Algoritmos Genéticos (GAs), que combinam soluções existentes por meio de operadores reprodutivos, como crossover e mutação.

Os EAs têm sido utilizados com sucesso em uma ampla variedade de problemas, incluindo problemas de otimização combinatória, problemas de roteamento, problemas de planejamento e problemas de controle.

A seguir, são apresentados alguns conceitos importantes, segundo (Gaspar-Cunha, 2012), que são abordados no texto:

- Cromossomo: Uma solução de um EA é representada por um cromossomo. O cromossomo pode ser representado por um array de símbolos, como um array binário ou um array de números reais.
- Gene: Cada elemento de um cromossomo é chamado de gene. O gene representa uma característica da solução.
- Aptidão: A função de aptidão é utilizada para avaliar a qualidade de uma solução. A função de aptidão deve ser específica do problema que está sendo otimizado.

- Seleção proporcional: A seleção proporcional é um mecanismo de seleção que seleciona as soluções com maior aptidão com maior probabilidade.
- Torneio de j indivíduos: O torneio de j indivíduos é um mecanismo de seleção que seleciona as soluções com maior aptidão entre um conjunto de j indivíduos sorteados.
- Seleção por truncamento: A seleção por truncamento seleciona as w soluções com maior aptidão, sendo w o número de indivíduos necessários para a próxima geração.
- Crossover: O crossover é um operador de recombinação que combina duas soluções para criar uma nova solução.
- Mutação: A mutação é um operador de recombinação que altera uma ou mais características de uma solução.
- Algoritmo Genético: O Algoritmo Genético é um EA canônico que utiliza operadores de recombinação e mutação simples.
- Estratégias Evolutivas: As Estratégias Evolutivas são um EA canônico que utilizam operadores de recombinação e mutação mais complexos.
- Programação Evolutiva: A Programação Evolutiva é um EA canônico que utiliza operadores de recombinação e mutação específicos para problemas de programação.
- Algoritmos de Estimação de Distribuição: Os EDAs são EAs baseados em modelos probabilísticos que utilizam modelos probabilísticos para gerar novas soluções.

2.1 Algoritmo Genético

O GA é uma técnica de otimização inspirada na evolução biológica. Foi inicialmente proposto por John Henry Holland em 1975. Além dos elementos comuns a todos os algoritmos evolutivos, como a população, o GA utiliza os conceitos de cromossomos, genes e alelos.

No GA, um cromossomo representa uma solução para o problema que está sendo resolvido. Ele é formado por uma sequência de genes, que representam as variáveis do problema.

Cada gene pode assumir um valor, chamado de alelo. Por exemplo, em problemas envolvendo variáveis binárias, os alelos podem ser 0 e 1. A avaliação de cada indivíduo de uma população é realizada pela função objetivo do problema. A função objetivo retorna um valor denominado aptidão (ou fitness) do indivíduo.

Com base nas aptidões, os indivíduos são selecionados para reprodução. Os critérios de seleção mais comuns são: roleta, torneio e truncamento.

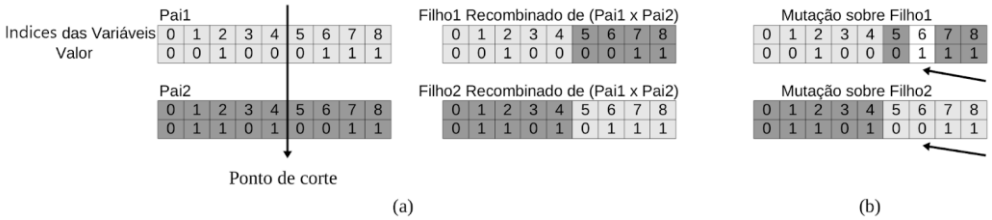
Os indivíduos são ordenados de acordo com sua aptidão, sendo que indivíduos com maior aptidão têm maior probabilidade de serem selecionados. Uma segunda estratégia de seleção é

o torneio, no qual dois indivíduos são selecionados aleatoriamente e o indivíduo com maior aptidão é escolhido. Por fim, a seleção por truncamento consiste em selecionar os "Ns" melhores indivíduos de uma população, descartando o restante (Gaspar-Cunha, 2012).

Após a seleção, os operadores de reprodução são aplicados aos indivíduos selecionados. Os operadores de reprodução do AG são a recombinação (também chamada de crossover) e a mutação.

A recombinação mimetiza o processo de crossing-over da genética. Ela separa os cromossomos de dois indivíduos selecionados para reproduzir (também chamados de pais) em dois pedaços, com base em um ponto de corte. Todas as possíveis combinações desses pedaços são geradas (sem repetição de posição), gerando dois novos cromossomos (indivíduos). O operador de mutação sorteia uma posição de um vetor e altera o seu valor (em geral a troca do valor de um bit). A Figura 2 mostra a reprodução de um GA a partir de dois indivíduos selecionados (Pai1 e Pai2) em que os operadores de recombinação e mutação são utilizados em sequência. É importante notar que a escolha do ponto de corte dos progenitores não leva em conta as correlações entre as variáveis. A determinação desse tipo de relação pode ser utilizada para aumentar o desempenho de EAs para problemas de maior complexidade.

Figura 2: Recombinação (a) e Mutação (b) em um GA.



Fonte: Soares , 2014.

No entanto, GAs podem apresentar limitações para problemas complexos (Soares , 2014). Por exemplo, problemas deceptivos são problemas de otimização em que soluções ruins podem parecer boas e vice-versa. Isso ocorre porque a função objetivo que define o problema pode ser enganadora, dando a impressão de que uma solução é melhor do que realmente é.

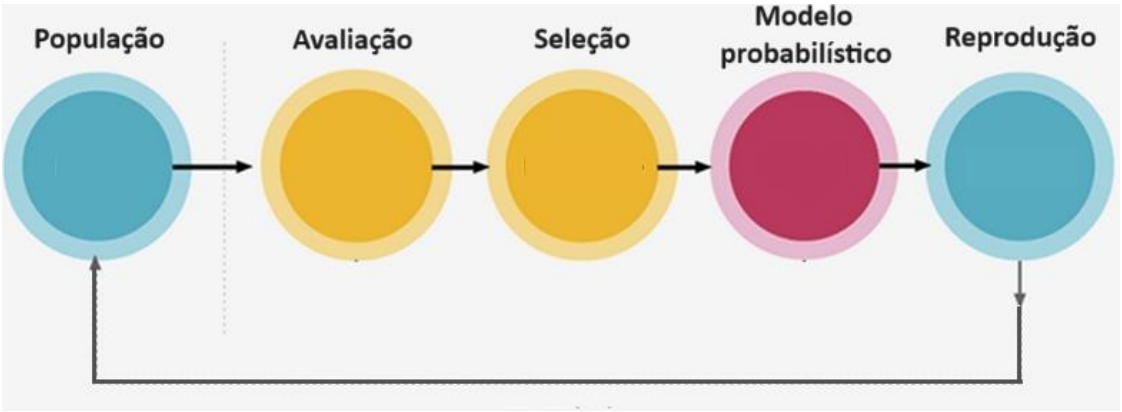
Os GAs em problemas deceptivos podem apresentar um desempenho inferior a outros algoritmos evolutivos, como os Algoritmos de Estimação de Distribuição (EDAs) (Pelikan et. al, 2003). Os GAs que utilizam modelos probabilísticos de distribuição dos valores das variáveis (para orientar o processo de busca da solução de problemas) são chamados Algoritmos de Estimação de Distribuição.

2.2 Algoritmos de Estimação de Distribuição (EDAs)

Os EDAs, introduzidos por Mühlenbein (Mühlenbein, 1997), utilizam uma distribuição de probabilidades para melhorar o procedimento de recombinação. Posteriormente, outros trabalhos incorporaram métodos para identificar correlações entre variáveis (Larrañaga, 2001). Esses modelos probabilísticos, de forma direta ou indireta, geram aproximações dos pontos ótimos de um problema.

A Figura 3 apresenta as etapas de um EDA. A diferença essencial em relação a um Algoritmo Evolutivo (EA) típico (Figura 1) é a construção de um modelo probabilístico com base nos indivíduos selecionados e o processo de reprodução baseado no modelo.

Figura 3: Etapas de uma geração de um EDA típico.



Fonte: autoria própria

Os EDAs constroem um modelo probabilístico baseado nas soluções existentes na população e, então, geram novas soluções a partir das informações captadas pelo modelo.

A modelagem probabilística é uma ferramenta poderosa para algoritmos evolutivos (EAs). Ela permite que os EAs construam modelos de regiões promissoras no espaço de busca. Esses modelos podem ser usados para gerar novas soluções que são mais propensas a serem ótimas.

A exatidão de um modelo probabilístico afeta diretamente a taxa de convergência do EDA (Goldberg, 2002) e a capacidade de encontrar a solução ótima do problema. No entanto, um modelo muito preciso pode ser lento de construir, o que pode prejudicar a eficiência do EA.

Uma estratégia para equilibrar a precisão e a eficiência é decompor o problema original em subproblemas mais simples. A partir dos modelos construídos para esses subproblemas, os modelos para o problema original podem ser obtidos de forma mais eficiente, mantendo sua capacidade de representar regiões promissoras.

A decomposição do problema original pode ser feita identificando grupos de variáveis altamente correlacionadas, que são chamadas de blocos de construção (BBs). Os EAs mais eficientes e eficazes são aqueles que constroem modelos probabilísticos capazes de mapear corretamente mais BBs.

Modelos probabilísticos que representem bem as melhores instâncias de cada ponto ótimo global podem evitar a convergência prematura para um ótimo local (armadilha) e exigir menos avaliações de soluções candidatas até convergir para o ótimo global.

A representatividade do modelo pode ser melhorada com o aumento do número de amostras (indivíduos selecionados da população). Portanto, um EDA pode necessitar de uma população muitas vezes maior que um EA típico para que seu modelo seja vantajoso. O custo de uma população maior pode ser compensado por uma quantidade de gerações do EDA menor que a do EA.

Por outro lado, a construção de um modelo probabilístico representativo por meio de um algoritmo computacionalmente eficiente é um dos principais desafios na área de EDAs. Em geral, tal construção requer alto custo computacional, com o custo aumentando com o nível da representatividade do modelo.

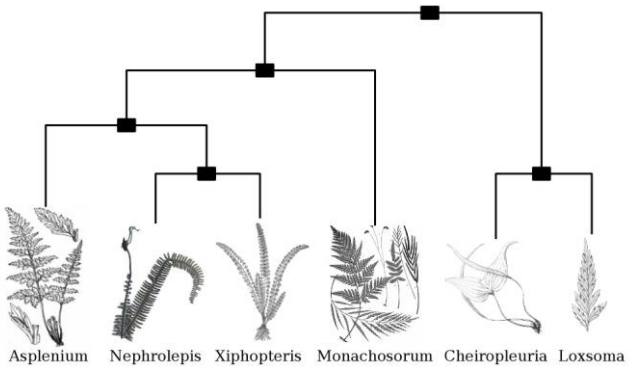
Aqui cabe uma introdução do conceito de filogenia como um modelo probabilístico para EDAs.

2.3 Filogenia

Uma filogenia é uma árvore que representa as relações entre espécies de origem comum. O termo 'árvore filogenética' é empregado para descrever filogenias derivadas tanto de dados morfológicos quanto de sequências genéticas. Neste estudo, são desenvolvidas filogenias a partir de diversas fontes de dados, com o objetivo de identificar grupos de objetos (filos), estabelecendo correlações entre esses dados.

Uma filogenia é uma representação gráfica das relações evolutivas entre um conjunto de espécies. Ela é geralmente representada por uma árvore, que é um tipo de gráfico conexo e acíclico. Elas são compostas por ramos que representam a evolução de uma espécie ao longo do tempo, e as folhas representam as espécies atuais. As árvores filogenéticas são geralmente árvores binárias, em que cada nó interno representa um ancestral comum e cada ramo representa uma divergência evolutiva. A Figura 4 mostra uma filogenia que destaca as relações evolutivas entre as espécies de plantas.

Figura 4 - Uma filogenia para um grupo de plantas usando sequências de nucleotídeos do conjunto teste rbcL_55

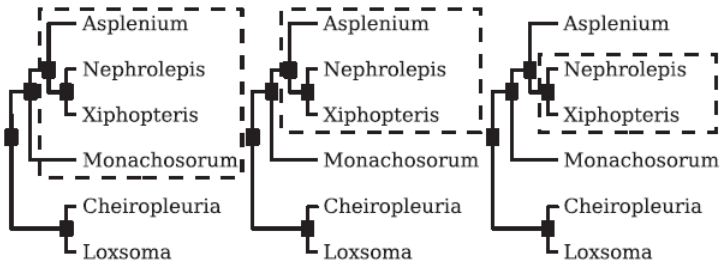


Fonte: Cancino, 2007.

Em uma filogenia, as folhas representam espécies existentes, enquanto os nós internos representam ancestrais ou espécies extintas. Clados são grupos que incluem um ancestral e todos os seus descendentes. Eles são a informação mais útil em uma filogenia porque as relações dentro de um clado são fortes.

Clados são grupos de espécies que compartilham um ancestral comum. Eles são a informação mais útil em uma filogenia porque as relações dentro de um clado são fortes. A Figura 5 ilustra possíveis clados contendo as espécies Nephrolepis e Xiphopteris. Essas espécies são irmãs, o que significa que compartilham um ancestral comum mais recente do que qualquer outra espécie.

Figura 5 - Possíveis clados (em tracejado) obtidos da filogenia da Figura 5.



Fonte: Cancino, 2007.

Um clado corresponde a um subconjunto de variáveis correlacionadas, que podem ser vistas como BBs. Uma variável pode estar presente em mais de um BB, e esses podem se sobrepor se tiverem variáveis em comum

Em biologia, sequências de DNA se sobrepõem se tiverem material genético em comum, como geralmente ocorre em um clado. Um processo de reamostragem (Felsenstein, 2003) pode ser usado para gerar vários filogramas (filogenias quando sequências de DNA são usadas) do mesmo conjunto de dados para fornecer significância estatística às relações encontradas.

As variáveis (características) podem ser organizadas num método de clusterização, conhecido como Clusterização baseada em Análise de Filograma (CPA).

Por meio de filogramas, o CPA pode encontrar um consenso dessas relações e compor uma rede baseada nelas. Nessa rede, variáveis altamente correlacionadas (indicando um BB) são modeladas por um clique de grafo. Em seguida, a técnica CPA usa um algoritmo de detecção de comunidade de redes complexas (Donetti, 2004) para detectar automaticamente comunidades (BBs do ponto de vista de otimização) a partir da rede.

A clusterização ou clustering é uma técnica de aprendizado não supervisionado que divide um conjunto de dados em grupos de pontos de dados que são semelhantes entre si.

Portanto, o CPA usa a detecção de comunidade para identificar grupos de variáveis que estão fortemente correlacionadas entre si. Esses grupos são chamados de blocos de busca.

Os métodos de clustering podem ser divididos em dois grupos principais: clustering hierárquico e clustering não hierárquico.

- Clustering hierárquico: Os métodos de agrupamento hierárquico criam uma hierarquia de clusters, começando com cada ponto de dados como um cluster individual e, em seguida, combinando os clusters semelhantes em clusters maiores até que apenas um cluster permaneça.
- Clustering não hierárquico: Os métodos de agrupamento não hierárquico dividem os dados em um número fixo de clusters. Os métodos mais comuns de agrupamento não hierárquico são o k-means, k-medoids e agglomerative.

Vantagens do agrupamento hierárquico:

O clustering hierárquico oferece algumas vantagens sobre o clustering não hierárquico, incluindo:

- Não requer a especificação do número de clusters: O clustering hierárquico pode gerar uma hierarquia de clusters, permitindo que os usuários escolham o número de clusters que melhor atende às suas necessidades.
- É mais robusto a outliers: O clustering hierárquico é menos afetado por outliers do que o agrupamento não hierárquico.

Segundo (Han et al, 2012), o clustering pode ser usado para uma variedade de propósitos além da identificação de grupos de variáveis ou pontos de dados. Por exemplo:

- Segmentar clientes: O agrupamento pode ser usado para segmentar clientes em grupos com características semelhantes, o que pode ajudar as empresas a direcionar suas ofertas de produtos e serviços de forma mais eficaz.
- Descobrir padrões: O agrupamento pode ser usado para descobrir padrões nos dados, o que pode ajudar os pesquisadores a entender melhor o mundo ao seu redor.
- Realizar previsões: O agrupamento pode ser usado para realizar previsões sobre o comportamento futuro dos dados, o que pode ajudar as empresas a tomar decisões mais informadas.

2.4 Clusterização baseada em Análise de Filograma (CPA)

O CPA é um método que pode construir modelos de relações entre variáveis, considerando problemas com diferentes níveis de dificuldade (Cancino, 2007). As variáveis podem ser divididas em três tipos: BBs sem sobreposição, BBs com alguma sobreposição e BBs com sobreposição total. O CPA possui as seguintes etapas principais, que são executadas em sequência:

Passo 1: Bootstrapping

Bootstrap é um método estatístico que usa um único conjunto de dados para criar inúmeras amostras "simuladas". Esse processo permite calcular erros padrão, construir intervalos de confiança e realizar testes de hipóteses para vários tipos de estatísticas de amostra.

O primeiro passo é criar inúmeras amostras do mesmo tamanho da amostra original, com reposição. Isso significa que cada observação da amostra original pode ser selecionada mais de uma vez em uma reamostragem.

É importante destacar que os processos de reamostragem podem ser repetidos várias vezes para obter várias populações selecionadas, melhorando as estimativas produzidas pelo CPA.

Após criar as reamostras, podemos usar as estatísticas de cada reamostra para estimar a distribuição da estatística de amostra original. Por exemplo, podemos usar as médias das reamostras para estimar a distribuição da média populacional.

Essa distribuição pode ser usada para calcular erros padrão, construir intervalos de confiança e realizar testes de hipóteses.

Passo 2: Matriz de distância

Para cada população selecionada amostrada no Passo 1, uma matriz de distância é calculada usando uma métrica aplicada a cada par de variáveis do problema. Todos os valores atribuídos a uma variável □□ na população selecionada compõem um vetor v□, e a distância entre as variáveis □□ e □□, é calculada usando seus respectivos vetores v□ e v□,.

O CPA utiliza duas métricas para construir a matriz de distância: Informação Mútua (IM) (Kraskov et. al 2003) para problemas com variáveis binárias, e NCD (Cilibrasi, 2005) para variáveis reais ou mistas. Outras métricas podem ser aplicadas de acordo com o conhecimento prévio sobre o domínio do problema.

Passo 3: Reamostragem de filogramas

No terceiro passo, o CPA gera um conjunto de filogramas possíveis a partir da matriz de distância determinada no passo anterior. Para isso, o algoritmo Neighbor Joining (NJ) é usado para construir cada filograma. O NJ é um algoritmo eficiente quando comparado a outros métodos usados para construir árvores filogenéticas.

No entanto, o NJ não pode garantir um filograma ótimo, pois é um algoritmo guloso e gera apenas um filograma possível. Para superar esse obstáculo, o CPA permuta as colunas e linhas da matriz de distância. Quando as colunas e linhas da matriz de distância são permutadas, o NJ normalmente constrói diferentes filogramas sub-ótimos para cada permutação.

Esses filogramas diferem entre si, mas, em geral, preservam alguns clados (possíveis BBs). Isso ocorre porque uma sub-árvore associada a um clado é calculada pelo NJ usando menos indireções (distâncias relativas envolvendo nós internos) para objetos.

Passo 4: Conversão para formato de rede

No quarto passo, o CPA converte a saída do NJ do formato Newick (Felsenstein, 2003) para uma matriz de adjacência. O formato de rede é mais adequado para a aplicação de algoritmos de detecção de comunidades em redes complexas.

Passo 5: Determinação de particionamentos

No quinto passo, o CPA determina um conjunto de possíveis particionamentos de variáveis do problema. Ele aplica o algoritmo de detecção de comunidade Fast Newman (FN), que apresenta um trade-off adequado entre a qualidade das comunidades obtidas e o tempo de execução necessário (Crocomo, 2013).

O FN é um algoritmo guloso, o que significa que ele constrói uma comunidade a cada passo, adicionando a ela a variável com a menor entropia. As permutações aleatórias de linhas e colunas da matriz de adjacência podem levar a diferentes particionamentos de variáveis. Para garantir a robustez do processo, o CPA aplica o FN a várias permutações, obtendo um conjunto de particionamentos. Os particionamentos menos frequentes são então eliminados.

Passo 6: Conversão para formato de rede de folhas

No sexto passo, o CPA converte cada particionamento de nós folha (obtido no Passo 5) para um formato de rede de folhas. Para isso, assume-se que todos os nós em uma mesma partição (um possível clado ou BB) formam um clique do grafo.

O formato de rede de folhas é necessário no próximo passo, que também representa os índices de linhas e colunas em uma ordem padronizada (a ordem ascendente de índices é considerada).

Passo 7: Construção de uma Rede Equivalente Relativa a uma População Específica

Neste sétimo passo, todas as matrizes de adjacência que representam os particionamentos identificados em cada rede de folhas, provenientes da mesma população selecionada (ou seja, a partir da mesma reamostragem), são combinadas através da soma lógica ou do operador lógico "OR" (Lipschutz, 2004). O resultado dessa operação é uma matriz de adjacência equivalente, também conhecida como rede equivalente.

Passo 8: Derivação do Particionamento Final a Partir de Diversas Populações Selecionadas

No oitavo passo, o algoritmo Fast Newman (FN) é empregado, desta vez, em cada matriz de adjacência equivalente (rede equivalente), gerando um particionamento para cada população selecionada. Posteriormente, todas as matrizes de adjacência equivalentes são combinadas através da soma lógica ou do operador lógico "OR" (Lipschutz, 2004), seguindo o mesmo procedimento descrito no Passo 7. Esse processo resulta em uma única matriz de adjacência, denominada rede final. Em seguida, o FN é aplicado à rede final para obter o particionamento final.

Durante este passo, a CPA demonstra sua capacidade de extrair qualidade (relacionamentos relevantes) a partir da quantidade. Importante notar que as relações presentes em uma rede de baixa qualidade (por exemplo, aquelas com conexões incorretas entre partições) não devem impactar significativamente o resultado final (comunidades obtidas ou Ramos de Bifurcação) da rede final. Isso se deve à importância relativamente pequena de conexões espúrias entre nós em comparação com as arestas corretamente incluídas, que são mais frequentes.

Além disso, a análise da rede final, resultante da combinação dessas matrizes, revela que uma pequena partição em uma matriz de adjacência pode se expandir (formando um clado) por meio da verificação de consenso entre elas. Adicionalmente, a inclusão de arestas entre nós pode tornar uma partição mais representativa ou confiável.

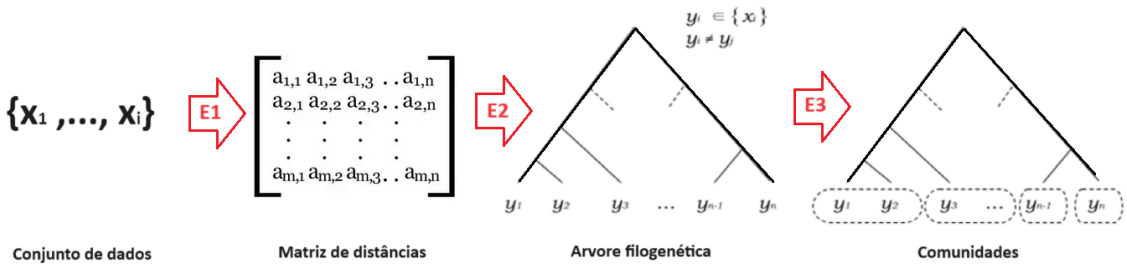
No entanto, poucas implementações consideram uma mistura de dados numéricos e categóricos. Isso pode ser uma limitação em aplicações que envolvam dados mistos, como dados de clientes, que incluem informações demográficas, preferências de compra e histórico de compras.

2.5 DAMICORE (DAta MIning of Code Repositories)

O modelo DAMICORE (Sanches, 2011), aborda essa limitação diretamente, fornecendo um algoritmo que lida com qualquer tipo de dado (números inteiros, reais e complexos, dados categóricos, imagens, entre outros).

Baseia-se em conceitos extraídos da Teoria da Informação, Redes Complexas e Inferência Filogenética, e visa revelar relações hierárquicas entre objetos de dados não estruturados. Seus princípios de funcionamento são implementados em três etapas:

Figura 6 - Passos do modelo DAMICORE



Fonte: Cancino, 2007.

Na Figura 6, os elementos a□□ da matriz de distância correspondem a uma medida de dissimilaridade entre os elementos □□ e □□, de acordo com alguma métrica fornecida. A matriz é então decomposta em uma árvore onde a distância entre quaisquer dois itens (nós folha) corresponde à soma dos comprimentos dos ramos que conectam esses dois itens. Finalmente, a terceira etapa identifica grupos de itens que estão significativamente conectados em clusters de similaridade distintos.

Detalhando as etapas do modelo DAMICORE, temos:

E1 - A primeira etapa consiste na construção de uma matriz de distância entre todos os pares de objetos de dados. A matriz de distância é construída usando uma métrica de similaridade, que é uma função que mede a similaridade entre dois objetos. No DAMICORE, a métrica de similaridade utilizada é a NCD.

A NCD é uma medida de dissimilaridade que explora o fato de que, para objetos semelhantes, deve ser relativamente fácil descrever um em termos do outro. Formalmente, a NCD é definida como:

D_z(ab) = C_z(ab) - min{C_z(a), C_z(b)} / max{C_z(a), C_z(b)}

onde:

- 'a' e 'b' são os dois objetos de dados a serem comparados;
- 'ab' é a concatenação de 'a' e 'b';
- 'C_z(x)' é o tamanho da versão compactada do objeto 'x', obtida pela aplicação de um algoritmo de compressão 'z'.

Para um compressor ideal e dois arquivos idênticos, C_z(ab) = C_z(a) = C_z(b), resultando em D_z(a,b) = 0. Já para dois arquivos sem nenhuma similaridade, C_z(ab) = C_z(a) + C_z(b), resultando em D_z(a,b) = 1. Para compressores não ideais, D_z(a,b) varia de 0 a 1.

E2 - A segunda etapa consiste na construção de uma árvore filogenética a partir da matriz de distância. A árvore filogenética é uma estrutura que representa as relações evolutivas entre os objetos de dados.

No DAMICORE, a árvore filogenética é construída usando um algoritmo de construção de árvores filogenéticas. Esse algoritmo utilizado no DAMICORE é o Algoritmo de UPGMA.

O UPGMA consiste em um algoritmo de aglomeração que agrupa os objetos de dados de acordo com sua similaridade. O algoritmo inicia com cada objeto de dados como um nó separado. Em seguida, os nós são agrupados em pares, de acordo com sua distância mínima. O processo é repetido até que todos os objetos de dados sejam agrupados em um único nó.

E3 - A terceira etapa consiste na detecção de comunidades na árvore filogenética. Uma comunidade é um conjunto de objetos de dados que estão significativamente conectados entre si.

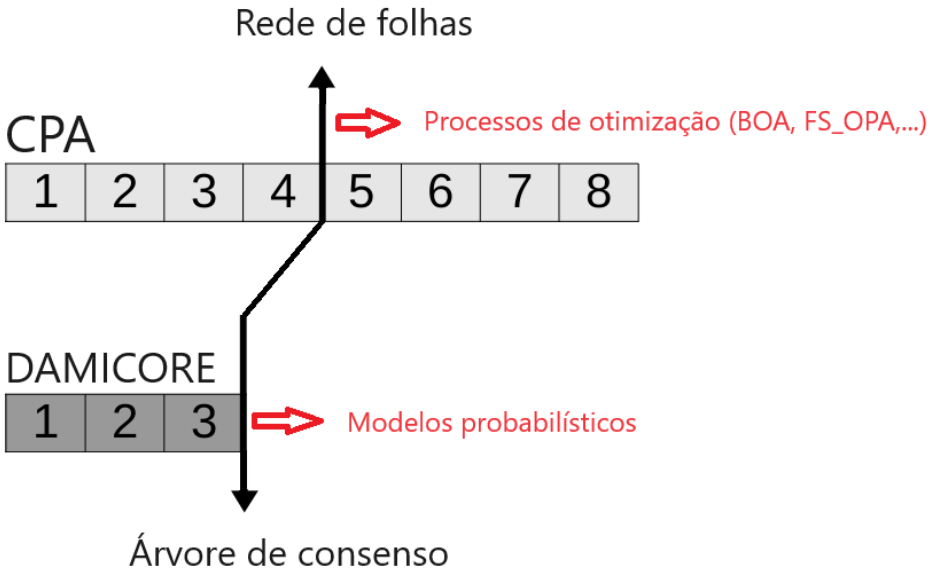
O modelo DAMICORE é um método de detecção de comunidades em grafos que é agnóstico em relação ao tipo de dados, oferecendo várias vantagens sobre os métodos de agrupamento existentes que não são capazes de lidar com dados mistos. Essas vantagens incluem:

- A capacidade de lidar com qualquer tipo de dado: O modelo DAMICORE é capaz de lidar com uma ampla gama de tipos de dados, incluindo números inteiros, reais e complexos, dados categóricos, imagens, entre outros. Isso o torna uma opção versátil para aplicações que envolvem dados mistos.
- A robustez a outliers: O modelo DAMICORE é menos afetado por outliers do que os métodos de agrupamento existentes. Isso ocorre porque o DAMICORE usa a distância euclidiana para calcular a similaridade entre pontos de dados. A distância euclidiana é menos sensível a outliers do que outras medidas de similaridade, como a correlação.

No DAMICORE, a detecção de comunidades é realizada usando o Algoritmo de Louvain. O Algoritmo de Louvain é um algoritmo de maximização de modularidade que divide uma rede em comunidades. O algoritmo inicia atribuindo cada nó da rede a uma comunidade. Em seguida, o algoritmo itera sobre a rede, realocando os nós de uma comunidade para outra, de acordo com um critério de maximização de modularidade. O processo é repetido até que nenhuma melhoria na modularidade seja possível.

A de se notar que o DAMICORE executa uma parte do modelo CPA, mais especificamente até o passo 4,. pois ele não irá caminhar para a fase de otimização (Figura 7). Sua importância está na sua rapidez em convergir para uma árvore de consenso e na geração de modelos probabilísticos, representando as relações identificadas entre variáveis. Essas árvores podem ser utilizadas para identificar, por exemplo, as variáveis potencialmente mais relevantes para um conjunto de dados e problemas específicos, sem a necessidade de conhecimento prévio.

Figura 7 - Comparando modelos CPA/DAMICORE



Fonte: autoria própria

3 METODOLOGIA

Com o propósito de entender a distribuição geográfica do comércio varejista de alimentos, utilizamos três principais fontes de dados: o Portal SAN - SEGURANÇA ALIMENTAR E NUTRICIONAL vinculado ao MINISTÉRIO DA CIDADANIA, o IPVS - Índice elaborado pela Fundação SEADE, e a Base de Dados da RAIS 2016, com um filtro específico para estabelecimentos que prestam serviços de alimentação ou comercializam alimentos no varejo.

As bases de dados utilizadas foram (Portal SAN, 2023)(SAGE, 2023)(IPVS, 2023):

- Base de dados das Capitais por Grupo de Percentil e Renda Domiciliar.

- Base de dados das Capitais - Espaços Intraurbanos.

- Base de dados dos Municípios.

- Base de dados da RAIS 2016 com filtro "Estabelecimentos que prestam serviços de alimentação ou comercializam alimentos no varejo".

- IPVS também foi utilizado, fornecendo uma visão detalhada das condições de vida no interior do município.

A partir dessas bases de dados, gerou-se uma base única, sendo cada coluna as variáveis do problema, perfazendo um total de 132 colunas (variáveis); e 96 linhas, correspondentes a cada distrito na capital de São Paulo.

A partir dessa base gerada, queremos inferir sobre os dados e tentar achar uma relação entre o nível de renda, o IPVS e a oferta de alimentos saudáveis, que será o nosso objetivo.

Após a unificação das bases, se fez necessário fazer um processo de discretização dos dados.

As três bases de dados foram fundidas em um único arquivo, conforme estrutura apresentada no arquivo Descrição dos dados.docx, no Apêndice A. A coluna “COD_DIST” foi usada como chave para essa união.

Neste ponto cabe uma discussão: a discretização de dados é o processo de converter variáveis contínuas em categorias ou intervalos discretos. Quando se tem valores numéricos misturados com valores NaN (not a number) nos dados, é importante lidar com esses valores ausentes antes de realizar a discretização.

No caso em estudo, a quantidade de tipos de dados é diversificada. São necessárias adotar soluções de discretização para lidar, por exemplo, com os valores ausentes. Assim, as soluções adotadas foram as seguintes:

- Variáveis desconsideradas: As variáveis "V38", "V39" e "geometry" foram desconsideradas da base de dados por não trazer nenhuma informação relevante.
- Variáveis com valores NaN: As variáveis "V47" e "V48" têm valores NaN misturados com valores numéricos. Para lidar com esses valores, foi utilizado o método de imputação de valores, substituindo os valores NaN por 0.
- Variáveis com valores "{ñ class}": A variável "BairrosSP" tem valores "{ñ class}" misturados com valores numéricos. Para lidar com esses valores, foi utilizado o seguinte procedimento:
    - Substituir "{ñ class}" por NaN para que possam ser tratados como valores ausentes.
    - Em seguida, converter a coluna para o tipo numérico.
    - Por fim, aplicar técnicas de discretização, como criar intervalos fixos, para agrupar os valores em categorias.

As técnicas de discretização utilizadas foram:

- Binarização: Essa técnica converte uma variável contínua em uma variável binária, atribuindo um valor 0 para os valores abaixo de um determinado limite e 1 para os valores acima desse limite.
- Quantilização: Essa técnica divide a variável contínua em quatro intervalos iguais (quantis).
- One Hot Encoder: Essa técnica converte uma variável categórica em uma variável binária, criando uma coluna para cada categoria.

Após a unificação e a discretização dos dados, o pipeline do processamento foi o seguinte:

1- Da tabela original discretizada, fizemos o processo de reamostragem (bootstrapping), com n=30 reamostragens.

2- Instalação das bibliotecas do DAMICORE e os pacotes corretos relacionados ao igraph

3- Execução do DAMICORE e geração das várias possíveis estruturas em árvore (filograma ou árvores filogenéticas)

4- Visualizações envolvendo as diversas árvores pelo pacote ToyTree.

5- Identificação do membership (o DAMICORE usa o conceito de membership para identificar e analisar padrões de relacionamento entre entidades) e geração da árvore de consenso.

6- Geração dos clusters de variáveis e das tabelas de probabilidade condicional (modelos probabilísticos)

7- Executar o processo de uma nova amostragem (Sampling), usando os modelos probabilísticos obtidos, gerando uma amostra com n=100 e selecionando os 50% maiores ou menores valores, dependendo da variável analisada. Com base no objetivo do nosso projeto, vamos analisar as seguintes variáveis envolvidas:

- coluna 7: Menores valores de renda (Bairros mais pobres) - "Renda média domiciliar R$" (50% menores valores).
- coluna 102: IPVS (piores: de 3 para cima)- "num IPVS" (50% maiores valores).
- coluna 131: Estabelecimentos que vendem alimentos In Natura - "obj in_natura" (50% maiores valores).
- coluna 11: Número de feiras livres - "Número de feiras livres" (50% maiores valores).

8- A partir daí, com base nos objetivos, repete-se o passo 1 até o passo 6 novamente, gerando os clusters finais para análise.

Aqui cabe uma ressalva. Poderíamos continuar circulando o fluxo até perceber uma convergência, mas como nosso intuito é ter uma amostra rápida da situação, paramos num ciclo, somente.

4 ANÁLISE E DISCUSSÃO DOS DADOS

Vamos iniciar nossa discussão apresentando a árvore de consenso (vide apêndices A1, A2, A3 e A4) e os clusters gerados para cada variável envolvida no estudo, ou seja:

- "Renda média domiciliar R$"
- "num IPVS"
- "obj in_natura"
- "Número de feiras livres"

A estrutura de cada árvore de consenso se encontra no Apêndice deste documento.

Vamos analisar os clusters para cada situação:

1- Menores valores de renda (Bairros mais pobres): Renda média domiciliar R$:

c_0: ['geometry_d', 'num IPVS', 'IndEstabParticipaPAT_d', 'IndRaisNegativa_d', 'TamanhoEstabelecimento_d', 'Bares e outros estabs bebidas cnae 5611202', 'Lanchonetes, casas de chá, de sucos e similares cnae 5611203', 'Ambulantes de alimentação cnae 5612100', '48_Vulnerabilidade muito baixa', 'V10_Vulnerabilidade media', 'V16_d', 'V17_d', 'V19_d', 'V20_d', 'V46_d', 'V47_d']

c_1: ['QtdVínculosEstatutários_d', 'RegiõesAdmDF_d', 'COD_DIST', 'ID_d', 'CD_GEODI_d', 'TIPO_urbano', 'V38_d', 'V39_d', 'V40_d', 'V43_d']

c_2: ['UF_x', 'Código do Município IBGE', 'Densidade de estabelecimentos não saudáveis (ultraprocessados) 10 mil hab', 'UF_y_d', 'Município_x', 'Restaurantes cnae 5611201', 'Área', 'Nome da Área', 'V1_d', 'V8_Área não urbanizada de cidade ou vila', 'Renda média domiciliar R$', 'Grupo de percentis de acordo com a densidade de estabelecimentos saudáveis (in natura e misto)', 'Densidade de estabelecimentos saudáveis (in natura + misto) 10 mil hab']

c_3: ['Número de feiras livres', 'Hipermercados cnae 4711301', 'Supermercados cnae 4711302', 'Minimercados, mercearias e armazéns cnae 4712100', 'Padaria e confeitaria de revenda cnae 4721102', 'Varejista de laticínios e frios cnae 4721103', 'Varejista de doces, balas, bombons e semelhantes cnae 4721104', 'Açougues cnae 4722901', 'Peixarias cnae 4722902', 'Varejista de hortifrutigranjeiros cnae 4724500', 'Lojas de conveniência cnae 4729602', 'Produtos alimentícios não especificados cnae 4729699', 'Número de Domicílios', 'Número de Habitantes']

c_4: ['Número de estabelecimentos In natura na base da RAIS', 'Número de estabelecimentos mistos na base da RAIS', 'Número de estabelecimentos ultraprocessados na base da RAIS', 'Total de estabelecimentos de alimentação na base da RAIS']

c_5: ['V8_Zona rural, exclusive aglomerado rural', 'V8_Área urbana isolada', 'V8__Área urbanizada de cidade ou vila', 'V9_N', 'V9_S', 'V10_Baixissima vulnerabilidade']

c_6: ['CNAE2.0Subclasse_d', 'V14_d', 'V24_d', 'V25_d', 'V28_d', 'V31_d']

c_7: ['V12_d', 'V13_d', 'V18_d', 'V21_d', 'V22_d', 'V23_d', 'V26_d', 'V27_d', 'V29_d', 'V30_d']

c_8: ['QtdVínculosCLT_d', 'IndAtividadeAno_d', 'Cantinas cnae 5620103', 'V2_d', 'V3_d', 'V4_d', 'V5_d', 'V6_d', 'V15_d']

c_9: ['V49_d', 'BairrosFortaleza_d', 'CNAE2.0Classe_d', 'CNAE95Classe_d', 'IndCEIVinculado_d', 'Município_y_d', 'V48_d']

c_10: ['TIPO_rural', 'V10_Vulnerabilidade baixa', 'V32_d', 'V33_d', 'V44_d']

c_11: ['V34_d', 'V36_d', 'V37_d', 'V41_d', 'V42_d']

c_12: ['BairrosRJ_d', 'QtdVínculosAtivos_d', 'IndSimples', 'NaturezaJurídica_d', 'TipoEstab_Cei', 'TipoEstab_Cei_d', 'TipoEstab_Cnpj', 'Alimentos preparados para consumo domiciliar cnae 5620104', 'AREA_d', 'CD_GEODM_discr', 'V62_d', 'V8_Aglomerado rural de extensão urbana', 'V8_Aglomerado rural isolado - outros aglomerados', 'V35_d', 'V45_d']

c_13: ['id_d', 'BairrosSP_d']

c_14: ['V10_Não classificado', 'V10_Vulnerabilidade alta (Urbanos)', 'V10_Vulnerabilidade muito alta (aglomerados subnormais urbanos)', 'V11_d']

2- IPVS (piores: de 3 para cima): num IPVS

c_0: ['RegiõesAdmDF_d', 'COD_DIST', 'ID_d', 'CD_GEODI_d', 'TIPO_urbano', 'V38_d', 'V39_d', 'V40_d', 'V43_d', 'V45_d', 'V47_d']

c_1: ['UF_x', 'Município_x', 'Área']

c_2: ['Código do Município IBGE', 'Densidade de estabelecimentos não saudáveis (ultraprocessados) 10 mil hab', 'Número de feiras livres', 'QtdVínculosEstatutários_d', 'Número de estabelecimentos In natura na base da RAIS', 'UF_y_d', 'Número de estabelecimentos mistos na base da RAIS', 'Número de estabelecimentos ultraprocessados na base da RAIS', 'Total de estabelecimentos de alimentação na base da RAIS', 'Restaurantes cnae 5611201', 'Nome da Área', 'V1_d', 'Número de Domicílios', 'Número de Habitantes', 'Renda média domiciliar R$', 'Grupo de percentis de acordo com a densidade de estabelecimentos saudáveis (in natura e misto)', 'Densidade de estabelecimentos saudáveis (in natura + misto) 10 mil hab']

c_3: ['Supermercados cnae 4711302', 'Açougues cnae 4722901', 'Peixarias cnae 4722902']

c_4: ['Hipermercados cnae 4711301', 'Minimercados, mercearias e armazéns cnae 4712100', 'Padaria e confeitaria de revenda cnae 4721102', 'Varejista de laticínios e frios cnae 4721103', 'Varejista de doces, balas, bombons e semelhantes cnae 4721104', 'Varejista de

hortifrutigranjeiros cnae 4724500', 'Lojas de conveniência cnae 4729602', 'Produtos alimentícios não especificados cnae 4729699']

c_5: ['geometry_d', 'num IPVS', 'id_d', 'BairrosSP_d', 'QtdVínculosAtivos_d', 'IndEstabParticipaPAT_d', 'TamanhoEstabelecimento_d', 'Lanchonetes, casas de chá, de sucos e similares cnae 5611203', 'V10_Vulnerabilidade muito alta (aglomerados subnormais urbanos)', 'V14_d', 'V25_d', 'V46_d']

c_6: ['V49_d', 'BairrosFortaleza_d', 'CNAE2.0Classe_d', 'CNAE95Classe_d', 'QtdVínculosCLT_d', 'IndAtividadeAno_d', 'IndCEIVinculado_d', 'IndRaisNegativa_d', 'Município_y_d', 'CNAE2.0Subclasse_d', 'Cantinas cnae 5620103', 'V6_d', 'V15_d', 'V48_d']

c_7: ['TIPO_rural', 'V10_Vulnerabilidade baixa', 'V32_d', 'V33_d', 'V34_d', 'V44_d']

c_8: ['V36_d', 'V37_d', 'V41_d', 'V42_d']

c_9: ['V2_d', 'V3_d', 'V4_d', 'V5_d', 'V62_d', 'V8_Zona rural, exclusive aglomerado rural', 'V8_Área não urbanizada de cidade ou vila', 'V8_Área urbana isolada', 'V8__Área urbanizada de cidade ou vila', 'V9_N', 'V9_S', 'V10_Baixissima vulnerabilidade']

c_10: ['BairrosRJ_d', 'IndSimples', 'NaturezaJurídica_d', 'TipoEstab_Cei', 'TipoEstab_Cei_d', 'TipoEstab_Cnpj', 'Alimentos preparados para consumo domiciliar cnae 5620104', 'AREA_d', 'CD_GEODM_discr', 'V8_Aglomerado rural de extensão urbana', 'V8_Aglomerado rural isolado - outros aglomerados', 'V35_d']

c_11: ['Bares e outros estabs bebidas cnae 5611202', 'Ambulantes de alimentação cnae 5612100', 'V10_Vulnerabilidade media', 'V16_d', 'V17_d', 'V19_d', 'V20_d', 'V23_d', 'V24_d', 'V27_d', 'V28_d', 'V31_d']

c_12: ['48_Vulnerabilidade muito baixa', 'V12_d', 'V13_d', 'V18_d', 'V21_d', 'V22_d', 'V26_d', 'V29_d', 'V30_d']

c_13: ['V10_Não classificado', 'V10_Vulnerabilidade alta (Urbanos)', 'V11_d']

3- Estabelecimentos que vendem alimentos In Natura: obj in_natura

c_0: ['Número de feiras livres', 'Hipermercados cnae 4711301', 'Supermercados cnae 4711302', 'Minimercados, mercearias e armazéns cnae 4712100', 'Padaria e confeitaria de revenda cnae 4721102', 'Varejista de laticínios e frios cnae 4721103', 'Varejista de doces, balas, bombons e semelhantes cnae 4721104', 'Açougues cnae 4722901', 'Peixarias cnae 4722902', 'Lojas de conveniência cnae 4729602', 'Produtos alimentícios não especificados cnae 4729699']

c_1: ['Código do Município IBGE', 'Número de estabelecimentos In natura na base da RAIS', 'Número de estabelecimentos mistos na base da RAIS', 'Número de estabelecimentos ultraprocessados na base da RAIS', 'Total de estabelecimentos de alimentação na base da

RAIS', 'Varejista de hortifrutigranjeiros cnae 4724500', 'Número de Domicílios', 'Número de Habitantes', 'Renda média domiciliar R$']

c_2: ['UF_x', 'Densidade de estabelecimentos não saudáveis (ultraprocessados) 10 mil hab', 'QtdVínculosEstatutários_d', 'RegiõesAdmDF_d', 'UF_y_d', 'Município_x', 'Restaurantes cnae 5611201', 'Área', 'CD_GEODI_d', 'TIPO_urbano', 'Nome da Área', 'V1_d', 'V8_Zona rural, exclusive aglomerado rural', 'V8_Área não urbanizada de cidade ou vila', 'V8_Área urbana isolada', 'V8__Área urbanizada de cidade ou vila', 'V9_N', 'V9_S', 'V10_Baixissima vulnerabilidade', 'Grupo de percentis de acordo com a densidade de estabelecimentos saudáveis (in natura e misto)', 'V38_d', 'Densidade de estabelecimentos saudáveis (in natura + misto) 10 mil hab', 'V39_d', 'V40_d', 'V43_d']

c_3: ['IBGESubsetor_d', 'CEPEstab_d', 'in_natura_d', 'ultraprocessado_d', 'misto_d', 'obj in_natura_d']

c_4: ['COD_DIST', 'ID_d']

c_5: ['geometry_d', 'num IPVS', 'id_d', 'BairrosSP_d', 'BairrosRJ_d', 'QtdVínculosAtivos_d', 'IndEstabParticipaPAT_d', 'Bares e outros estabs bebidas cnae 5611202', 'Lanchonetes, casas de chá, de sucos e similares cnae 5611203', 'Ambulantes de alimentação cnae 5612100', 'V8_Aglomerado rural isolado - outros aglomerados', 'V10_Não classificado', 'V10_Vulnerabilidade alta (Urbanos)', 'V10_Vulnerabilidade muito alta (aglomerados subnormais urbanos)', '48_Vulnerabilidade muito baixa', 'V10_Vulnerabilidade media', 'V11_d', 'V16_d', 'V19_d', 'V45_d']

c_6: ['TamanhoEstabelecimento_d', 'V17_d', 'V20_d', 'V25_d', 'V46_d', 'V47_d']

c_7: ['QtdVínculosCLT_d', 'IndAtividadeAno_d', 'IndRaisNegativa_d', 'CNAE2.0Subclasse_d', 'Cantinas cnae 5620103', 'V2_d', 'V4_d', 'V5_d', 'V6_d', 'V14_d', 'V15_d']

c_8: ['V49_d', 'BairrosFortaleza_d', 'CNAE2.0Classe_d', 'CNAE95Classe_d', 'IndCEIVinculado_d', 'Município_y_d', 'TIPO_rural', 'V3_d', 'V10_Vulnerabilidade baixa']

c_9: ['V32_d', 'V33_d', 'V36_d', 'V37_d']

c_10: ['V34_d', 'V41_d', 'V42_d', 'V44_d', 'V48_d']

c_11: ['V24_d', 'V28_d', 'V31_d']

c_12: ['V12_d', 'V13_d', 'V18_d', 'V21_d', 'V22_d', 'V23_d', 'V26_d', 'V27_d', 'V29_d', 'V30_d']

c_13: ['IndSimples', 'NaturezaJurídica_d', 'TipoEstab_Cei', 'TipoEstab_Cei_d', 'TipoEstab_Cnpj', 'Alimentos preparados para consumo domiciliar cnae 5620104', 'AREA_d', 'CD_GEODM_discr', 'V62_d', 'V8_Aglomerado rural de extensão urbana', 'V35_d']

4- Número de feiras livres: "Número de feiras livres"

c_0: ['Número de feiras livres', 'Hipermercados cnae 4711301', 'Supermercados cnae 4711302', 'Minimercados, mercearias e armazéns cnae 4712100', 'Padaria e confeitaria de revenda cnae 4721102', 'Varejista de laticínios e frios cnae 4721103', 'Varejista de doces, balas, bombons e semelhantes cnae 4721104', 'Açougues cnae 4722901', 'Peixarias cnae 4722902', 'Varejista de hortifrutigranjeiros cnae 4724500', 'Lojas de conveniência cnae 4729602', 'Produtos alimentícios não especificados cnae 4729699', 'Número de Domicílios', 'Número de Habitantes', 'Renda média domiciliar R$']

c_1: ['Densidade de estabelecimentos não saudáveis (ultraprocessados) 10 mil hab', 'UF_y_d', 'Restaurantes cnae 5611201', 'Nome da Área', 'V1_d', 'V8_Zona rural, exclusive aglomerado rural', 'V8_Área não urbanizada de cidade ou vila', 'V8_Área urbana isolada', 'V8__Área urbanizada de cidade ou vila', 'V9_N', 'V9_S', 'V10_Baixissima vulnerabilidade', 'Grupo de percentis de acordo com a densidade de estabelecimentos saudáveis (in natura e misto)', 'Densidade de estabelecimentos saudáveis (in natura + misto) 10 mil hab']

c_2: ['UF_x', 'Código do Município IBGE', 'QtdVínculosEstatutários_d', 'RegiõesAdmDF_d', 'Município_x', 'Área', 'COD_DIST', 'ID_d', 'CD_GEODI_d', 'TIPO_urbano', 'V38_d', 'V39_d', 'V40_d', 'V43_d']

c_3: ['IBGESubsetor_d', 'CEPEstab_d', 'in_natura_d', 'ultraprocessado_d', 'misto_d', 'obj in_natura_d']

c_4: ['num IPVS', 'id_d', 'BairrosSP_d', 'IndEstabParticipaPAT_d', 'IndRaisNegativa_d', 'CNAE2.0Subclasse_d', 'TamanhoEstabelecimento_d', 'Bares e outros estabs bebidas cnae 5611202', 'Lanchonetes, casas de chá, de sucos e similares cnae 5611203', 'Ambulantes de alimentação cnae 5612100', 'Cantinas cnae 5620103', 'V10_Vulnerabilidade muito alta (aglomerados subnormais urbanos)', '48_Vulnerabilidade muito baixa', 'V10_Vulnerabilidade media', 'V14_d', 'V16_d', 'V17_d', 'V19_d', 'V20_d', 'V24_d', 'V25_d', 'V28_d', 'V31_d', 'V46_d', 'V47_d']

c_5: ['CNAE95Classe_d', 'QtdVínculosCLT_d', 'IndAtividadeAno_d', 'V2_d', 'V4_d', 'V5_d', 'V6_d', 'V15_d']

c_6: ['BairrosFortaleza_d', 'CNAE2.0Classe_d', 'IndCEIVinculado_d', 'Município_y_d', 'TIPO_rural', 'V10_Vulnerabilidade baixa']

c_7: ['V49_d', 'V3_d']

c_8: ['V32_d', 'V33_d', 'V34_d', 'V36_d', 'V37_d']

c_9: ['V41_d', 'V42_d', 'V44_d', 'V48_d']

c_10: ['geometry_d', 'BairrosRJ_d', 'QtdVínculosAtivos_d', 'IndSimples', 'NaturezaJurídica_d', 'TipoEstab_Cei', 'TipoEstab_Cei_d', 'TipoEstab_Cnpj', 'Alimentos preparados para consumo domiciliar cnae 5620104', 'AREA_d', 'CD_GEODM_discr', 'V62_d', 'V8_Aglomerado rural de extensão urbana', 'V8_Aglomerado rural isolado - outros aglomerados', 'V35_d', 'V45_d']

c_11: ['V12_d', 'V13_d', 'V18_d', 'V21_d', 'V22_d', 'V23_d', 'V26_d', 'V27_d', 'V29_d', 'V30_d']

c_12: ['V10_Não classificado', 'V10_Vulnerabilidade alta (Urbanos)', 'V11_d']

c_13: ['Número de estabelecimentos In natura na base da RAIS', 'Número de estabelecimentos mistos na base da RAIS', 'Número de estabelecimentos ultraprocessados na base da RAIS', 'Total de estabelecimentos de alimentação na base da RAIS']

Numa primeira abordagem, vemos que a partir de 13 clusters de variáveis gerados (c_1 a c_13), para cada variável de nosso interesse de estudo, ou seja, variável aparece no cluster c_3. Ou seja, 'Renda média domiciliar R$', 'num IPVS', 'obj in_natura_d' e 'Número de feiras livres' num conjunto de 132 variáveis, aplicando o DAMICORE conseguimos reduzir nosso espaço de busca para uma tabela com um reduzido número de variáveis correlacionadas. variáveis.

Apesar de só termos uma relação direta entre as variáveis 'Número de feiras livres' e 'Renda média domiciliar R$', no caso 4 de análise de "Número de feiras livres", mesmo assim, para todos os casos, reduzimos o espaço de busca de correlação entre as variáveis.

Caberia agora a presença de um especialista para fazer uma análise mais detalhada das correlações desses grupos.

5 CONCLUSÃO E TRABALHOS FUTUROS

EDAs são algoritmos de otimização evolutiva que usam modelos probabilísticos para estimar os valores das variáveis e amostrar novas soluções. Esse processo tem mostrado ser eficaz na exploração do espaço de busca de problemas combinatórios.
Os modelos probabilísticos mais adequados para EDAs variam de acordo com as características do problema. Por exemplo, problemas com variáveis contínuas podem ser mais bem resolvidos com modelos probabilísticos que assumem distribuição contínua, como a distribuição normal. Problemas com variáveis discretas podem ser mais bem resolvidos com modelos probabilísticos que assumem distribuição discreta, como a distribuição binomial.
Outra vantagem foi trabalhar com dados complexos e heterogêneos multifontes.

Vimos tambem que o algoritmo DAMICORE (DAta MIning of Code REpositories) é uma ferramenta poderosa para a análise de grandes conjuntos de dados complexos e heterogêneos, onde se tem pouca informação a respeito. Ele oferece uma série de vantagens em relação a outros métodos de análise, incluindo:

- Eficiência: O DAMICORE é um algoritmo eficiente que pode analisar grandes conjuntos de dados complexos e heterogêneos em um tempo relativamente curto. Isso é possível devido ao uso de técnicas de aprendizado de máquina para reduzir a dimensionalidade dos dados e acelerar o processo de análise.

- Precisão: O DAMICORE é um algoritmo preciso que pode identificar padrões e relacionamentos complexos em grandes conjuntos de dados complexos e heterogêneos. Isso é possível devido ao uso de técnicas de mineração de dados para identificar padrões e correlações nos dados.
- Flexibilidade: O DAMICORE é um algoritmo flexível que pode ser adaptado para atender a uma ampla gama de necessidades de análise. Isso é possível devido à sua estrutura modular, que permite aos usuários escolherem os módulos e técnicas de análise que melhor atendem às suas necessidades.

Os EDAs têm sido aplicados a uma ampla gama de problemas, mas ainda há muitos problemas reais que poderiam ser investigados. Estudos de caso em problemas reais podem ajudar a avaliar o desempenho dos EDAs em situações do mundo real.

Os modelos probabilísticos que são usados nos EDAs são baseados em modelos probabilísticos clássicos, como a distribuição normal e a distribuição binomial. O desenvolvimento de novos modelos probabilísticos que sejam mais adequados para problemas específicos pode melhorar o desempenho dos EDAs.

REFERÊNCIAS

(The State of the World , 2023) The State of Food Security and Nutrition in the World 2023. Disponível em: https://www.fao.org/3/cc3017en/online/state-food-security-and-nutrition-2023/food-security-nutrition-indicators.html. Acesso em: 2 Dez 2023.

(IBGE, 2019) Instituto Brasileiro de Geografia e Estatística (IBGE). Pesquisa Nacional de Saúde (PNS): 2019. Rio de Janeiro: IBGE, 2020. Disponível em: https://www.ibge.gov.br/estatisticas/ sociais/saude/9160-pesquisa-nacional-de-saude.html. Acesso em: 2 Dez 2023.

(SAN, 2019) 2º Inquérito Nacional sobre Insegurança Alimentar no Contexto da Pandemia da Covid-19 no Brasil – Rede Brasileira de Pesquisa em Soberania e SAN. Disponível em: https://pesquisassan.net.br/2o-inquerito-nacional-sobre-inseguranca-alimentar-no-contexto-da-pandemia-da-covid-19-no-brasil/. Acesso em: 3 Dez 2023.

(The State of the World , 2022) The State Of Food Security And Nutrition In The World - Repurposing Food And Agricultural Policies To Make Healthy Diets More Affordable. Disponível em: https://www.fao.org/3/cc0639en/cc0639en.pdf. Acesso em: 3 Dez 2023.

(Portal SAN, 2023) Portal SAN - SEGURANÇA ALIMENTAR E NUTRICIONAL. Disponível em: https://aplicacoes.mds.gov.br/sagirmps/portal-san/. Acesso em: 3 Dez 2023.

(SAGI, 2023) Monitoramento SAGI. Disponível em:]https://aplicacoes.mds.gov.br/sagirmps/portal-san/artigo.php?link=23. Acesso em: 3 Dez 2023.

(Painel Global, 2017) Melhoria da nutrição através do aprimoramento dos ambientes alimentares. Resumo de políticas nº 7. Londres, Reino Unido: Painel Global sobre Agricultura e Sistemas Alimentares para a Nutrição. Disponível em: https://www.glopan.org/wp-content/uploads/2019/06/Ambientes-Alimentares-documento_0.pdf. Acesso em: 3 Dez. 2023.

(IPVS, 2023) IPVS - Índice Paulista de Vulnerabilidade Social. Disponível em: http://catalogo.governoaberto.sp.gov.br/dataset/21-ipvs-indice-paulista-de-vulnerabilidade-social#disqus_thread. Acesso em: 3 Dez. 2023.

(CONSEA, 2023) Relatório Do Plano Estadual De Segurança Alimentar E Nutricional Sustentável Vigência – 2019-2023. Disponível em: https://consea.agricultura.sp.gov.br/uploads/conferencia/vi/relatorio_plano_estadual_de_seguranca_alimentar.pdf. Acesso em: 3 Dez 2023.

(DeJong, 2008) JONG, K. A. D. Evolutionary computation : A unified approach. Cambridge: MIT Press, 256 p., 2006.

(Schwefel, 1996) SCHWEfel, Hans-Paul; MÄNNER, Reinhard. Metaheuristics: Theory and Applications. New York: Wiley, 1996.

(Raven et al, 2007), RAVEN, Peter H.; JOHNSON, George B.; SOLTIS, Dennis E.; SOLTIS, Peter M. Filogenia: Métodos e Aplicações. Rio de Janeiro: Guanabara Koogan, 2007.

(Goldberg, 2002) Goldberg, D. E. Genetic Algorithms in Search, Optimization, and Machine Learning (2nd ed.). New York, NY: Addison-Wesley.

(Gaspar-Cunha, 2012) Gaspar-Cunha, A. Takahashi, R.; Antunes, C. Manual de computação evolutiva e metaheurística. Coimbra: Imprensa da Universidade de Coimbra, 2012.

(Soares, 2014) Soares, Antonio Helson Mineiro. "Algoritmos de estimação de distribuição baseados em árvores filogenéticas." Universidade de São Paulo, 2014. http://www.teses.usp.br/teses/disponiveis/55/55134/tde-25032015-111952/. Acesso em: 25 Nov 2023.

(Harik et al, 1999) Harik, G. R., Lobo, F. G., Goldberg, D. E., & Cantu-Paz, E. (1999). The Behavior of Genetic Algorithms on Deceptive Problems: No Free Lunch Theorems Revisited. In J. J. Merelo Guervós, J. S. Aguilar González, & J. M. Merelo Guervós (Eds.), Parallel Problem Solving from Nature - PPSN IV (pp. 286-295). Berlin, Germany: Springer-Verlag.

(Pelikan et al, 2003) Pelikan, M., Goldberg, D. E., & Cantu-Paz, E. Genetic Algorithms vs. Bayesian Optimization Algorithms: A Theoretical and Empirical Comparison. In H.-M. Voigt, W. Ebeling, I. Rechenberg, & J. Schwefel (Eds.), Parallel Problem Solving from Nature - PPSN VIII (pp. 586-595). Berlin, Germany: Springer-Verlag.

(Mühlenbein, 1997) MÜHLENBEIN, H. The equation for response to selection and its use for pre- diction. Evolutionary Computation, 5 (3), 303–346.

(Larrañaga, 2001) Larrañaga, P.; Lozano, J. A. Estimation of distribution algorithms: A new tool for evolutionary computation (genetic algorithms and evolutionary computation). Springer, 2001.

(Cancino, 2007) Cancino, W. , & Delbem, A. Inferring phylogenies by multi-objective evolutionary algorithms. International Journal of Information Technology and Intelligent Computing, 2 , 1–26.

(Felsenstein, 2003) Felsenstein, J. Inferring phylogenies (2nd). Sinauer Associates.

(Donetti, 2004) Donetti, L. , & Muñoz, M. A. Detecting network communities: A new systematic and efficient algorithm. Journal of Statistical Mechanics: Theory and Experiment, 2004 (10), P10012.

(Han et al, 2012) HAN, J.; PEI, J.; KAMBER, M. Data mining: concepts and techniques. 3. ed. Waltham, MA: Morgan Kaufmann, 2012.

(Kraskov et al 2003) Kraskov, A., Stogbauer, H., Andrzejak, R., & Grassberger, P. Hierarchical clustering based on mutual information. Arxiv preprint q-bio/0311039.

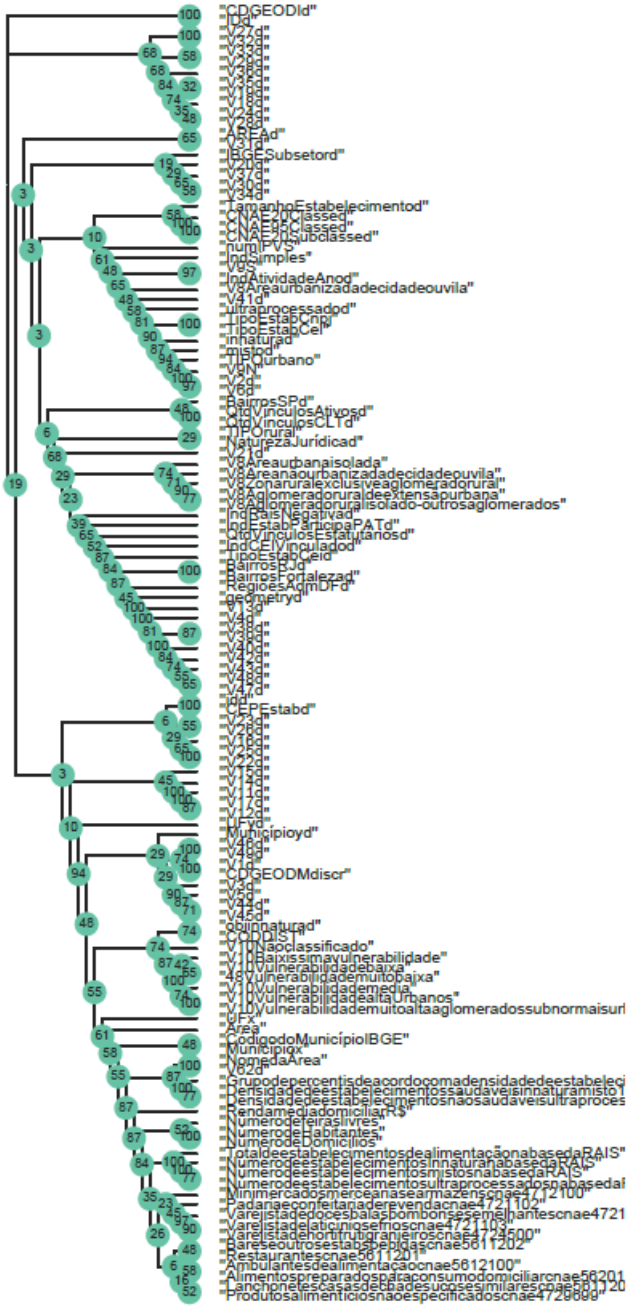
(Cilibrasi, 2005) Cilibrasi, R. , & Vitányi, P. M. B. Clustering by compression. IEEE Transactions on Information Theory, 51 , 1523–1545.

[(Crocomo, 2013) Crocomo, M. K., Martins, J. P., & Delbem, A. C. B. Decomposition of black-box optimization problems by community detection in Bayesian networks. International Journal of Nature Computing Research (IJNCR), 3 (4), 1–19. doi: 10.4018/jncr. 2012100101.

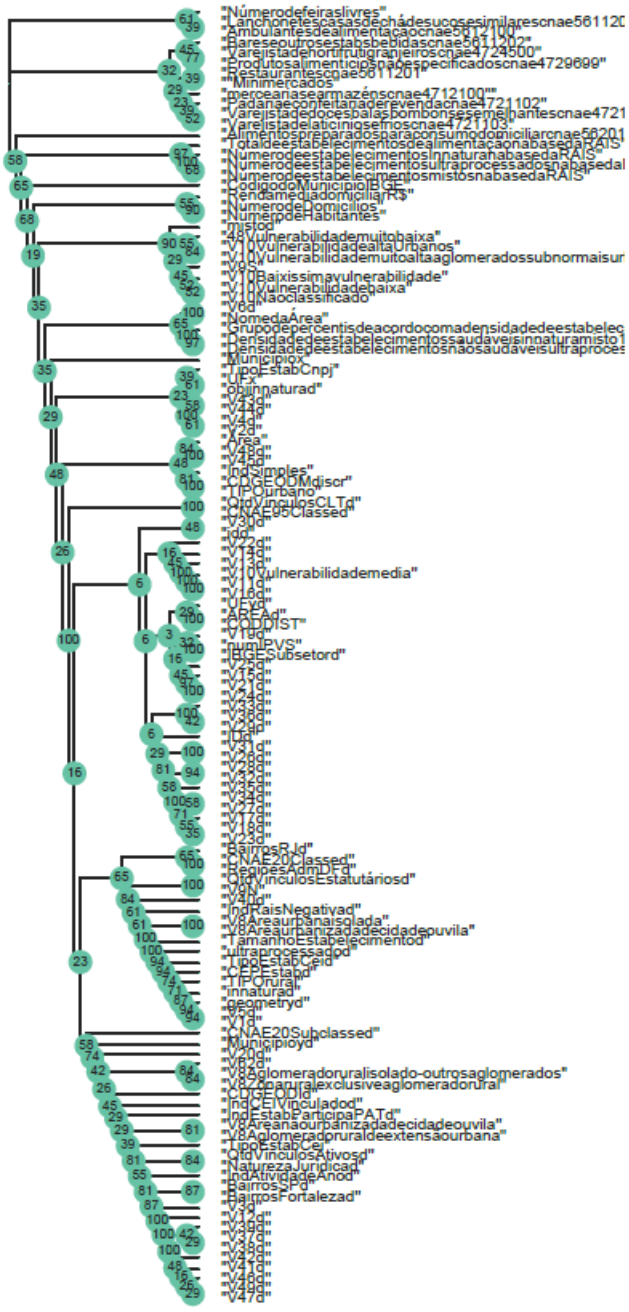
(Lipschutz, 2004) Lipschutz, S., & Lipson, M. Matemática Discreta: Coleção Schaum . Bookman Cia Ed. URL: http://books.google.com.br/books?id=2S9bwDmD1P0C.

(Sanches, 2011) Sanches, A., Cardoso, J.M., Delbem, A.C., 2011b. Identifying merge-beneficial software kernels for hardware implementation. In: 2011 International Conference on Reconfigurable Computing and FPGAs. pp. 74–79. http://dx.doi.org/10.1109/ReConFig. 2011.51.

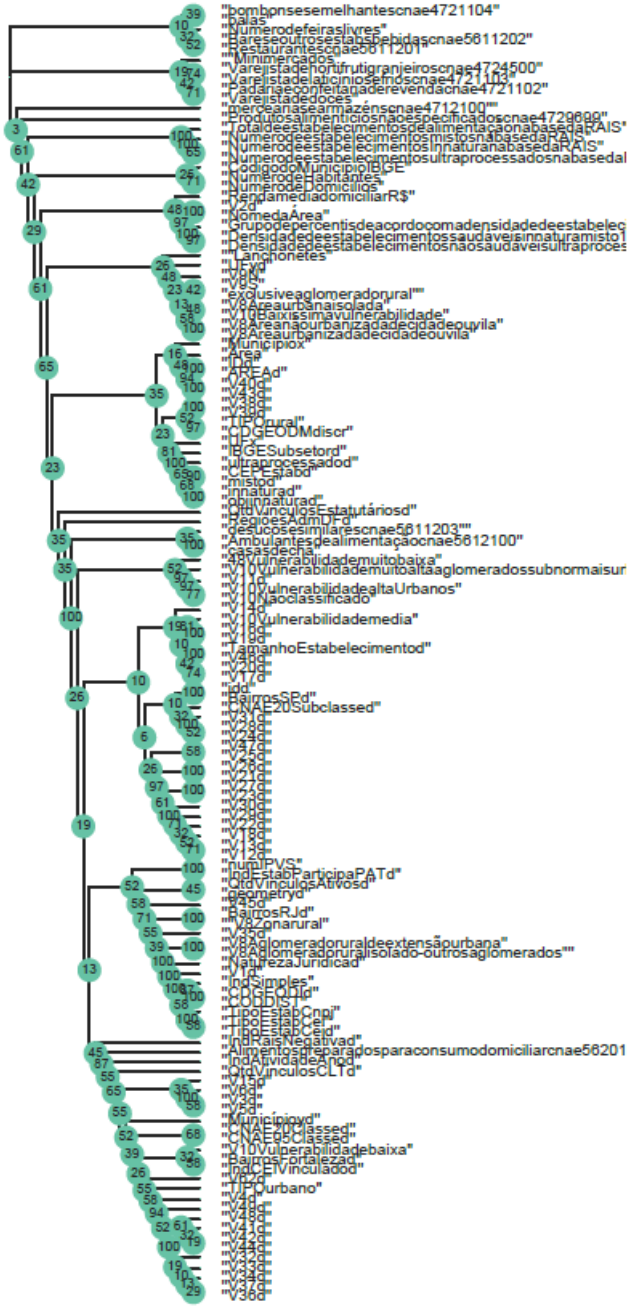
APENDICE A: Árvore de cosenso para renda baixa



APENDICE B: Árvore de cosenso para IPVS baixo



APENDICE C: Árvore de cosenso para estabelecimentos com alimentos “in natura”



APENDICE D: Árvore de cosenso para feiras livres