Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Mineração Visual de Dados em Notas Fiscais do Consumidor Eletrônicas

Frederico de Paiva Lenza

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Orientador
Prof. Dr. Vinícius Ruela Pereira Borges

Brasília
2020

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Mineração Visual de Dados em Notas Fiscais do Consumidor Eletrônicas

Frederico de Paiva Lenza

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Prof. Dr. Vinícius Ruela Pereira Borges (Orientador)
CIC/UnB

Prof.ª Dr.ª Aretha Barbosa Alencar
DACOM/UTFPR

Dr. Luís Paulo Faina Garcia
CIC/UnB

Prof. Dr. Marcelo Grandi Mandelli
Coordenador do Bacharelado em Ciência da Computação

Brasília, 11 de dezembro de 2020

Dedicatória

Dedico este trabalho a todas as pessoas que me deram suporte durante o período em que estive na universidade. Especialmente a minha família, amigos e ao meu orientador.

Agradecimentos

Agradeço à minha família que sempre me deu suporte. Ao meu orientador Vinicius que sempre foi compreensivo e esteve presente quando precisei de sua ajuda. Aos meus amigos que sempre estiveram próximos me ajudando de diversas formas. E a todos com quem pude ter contato durante esses anos na universidade, agradeço por tudo e espero que minha influência tenha sido igualmente positiva na vida de todos.

Resumo

Uma grande quantidade de notas fiscais de consumidor eletrônicas, associadas às compras em estabelecimentos comerciais, atacado e varejo, são geradas diariamente no Brasil. Nos dados das notas fiscais, existem alguns tipos de fraude relacionados com a evasão tributária, que é definida como a total ou parcial intenção de se isentar de pagar um tributo. Esta monografia propõe um processo de visualização exploratória, com o objetivo de auxiliar os especialistas em tarefas de auditoria fiscal, visando detectar fraudes e anomalias em dados financeiros. O processo foi formulado de forma que o especialista análise dados financeiros e que possuem atributos de diferentes tipos por meio de visualizações baseadas no posicionamento de pontos, considerando três diferentes técnicas de projeção multidimensional: Multidimensional Scaling, Isometric Mapping e t-distributed Stochastic Neighbor Embedding. As fraudes e anomalias nos dados financeiros podem ser identificadas pelo especialista ao interpretar os padrões e as relações de similaridade nas representações gráficas obtidas, como também pode-se manipular os pontos por meio de técnicas de interação, com possibilidade de utilizar algoritmos de agrupamento com a finalidade de enriquecer a análise. Nos experimentos, foram utilizados conjuntos de dados provenientes de notas fiscais do consumidor eletrônicas do Distrito Federal e de compras de cartão de crédito. As representações gráficas produzidas pela técnica de visualização t-SNE apresentaram melhor qualidade em relação às demais, sendo possível identificar as notas fiscais mais similares e que possuem tributação parecida, como também notas fiscais que possuem anomalias e que podem ser indícios de fraude. O processo de visualização exploratória mostrou ser potencialmente útil para auxiliar o especialista no entendimento dos padrões globais e locais nos dados por meio da interação com as representação gráficas obtidas.

Palavras-chave: Visualização exploratória de dados, mineração visual de dados, projeções multidimensionais, detecção de fraudes, notas fiscais do consumidor eletrônicas.

Abstract

A great quantity of electronic receipts, associated with purchases in commercial establishments, wholesale and retail, are generated daily in Brazil. There are some types of fraudulent behaviors related to electronic receipts, these behaviors are defined as the total or partial intention of exempting yourself from paying a tribute. This study proposes a visual exploration process aiming at supporting specialists in the task detecting frauds and anomalies in transactional data. The process was created in a way that enables the specialist to visualize transactional data that present attributes of different types using multidimensional projection algorithms, such as Multidimensional Scaling, Isometric Mapping and t-distributed Stochastic Neighbor Embedding. Anomalies in transactional data can be identified by the specialist when interpreting patterns and similarity relationships embedded in the obtained graphical layouts. The layouts can also be manipulated to a certain degree with the usage of techniques such as zoom and filter. Optionally, the layout can also be clustered to reveal hidden patterns found by unsupervised machine learning algorithms. The datasets used in the experiments were from the electronic tax invoice data gathered in the Federal District and a German credit card dataset. The graphical representations generated through t-SNE had the best quality from the other techniques utilized, being possible to identify data clustered together with similar data as well as potential evidence of anomalies. The visual exploration process showed to be useful to support the specialist in understading the global and local data patterns by means of interactive resources with the obtained layouts.

Keywords: Visual data exploration, visual data mining, multidimensional projections, fraud detection, eletronic tax invoices

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

As notas fiscais do consumidor eletrônicas (NFCe) são documentos fiscais emitidos para o consumidor final no ato de venda de um produto, que foram criadas com o objetivo de oferecer uma alternativa totalmente eletrônica de documentos fiscais, efetivamente substituindo os documentos utilizados no varejo até então [5]. Dessa forma, é possível reduzir os custos de obrigações acessórias aos contribuintes e possibilitar o aprimoramento do controle fiscal pelas Administrações Tributárias. Essas organizações visam garantir que a tributação ocorra de maneira uniforme a todos, em conformação com as regras correspondentes, uma vez que os próprios indivíduos que pagam tributos devidamente são aqueles que mais sentem os reflexos negativos da sonegação de impostos [6].

A sonegação de impostos (ou fraude fiscal) pode ser definida como a intenção de eximir-se, total ou parcialmente do pagamento de um tributo [7]. No caso das NFCes, existem vários meios pelos quais a fraude pode ocorrer. Um desses meios é o uso do Código de Situação Tributária (CST) incompatível com o produto, reduzindo assim o Imposto sobre Circulação de Mercadorias e Prestação de Serviços (ICMS) aplicado. Existe também a possibilidade de que o contribuinte realize um cálculo incorreto do valor do imposto, declarando um preço menor do que o preço real de um produto, fazendo com que se aplique uma alíquota de ICMS menor do que a devida.

Diariamente, uma grande quantidade de NFCes são geradas devido às operações de compras no atacado e no varejo, fazendo com que as tarefas de auditoria e de controle fiscal, quando realizadas por especialistas humanos, sejam inviáveis em relação ao elevado esforço e custo empreendidos. Além disso, frequentemente existem padrões implícitos em conjuntos de dados que passam despercebidos em uma análise manual e que estão sujeitos à erros de interpretação sem a utilização de ferramentas apropriadas [8]. Dessa forma, com o intuito de automatizar essas tarefas, faz-se necessária a utilização de abordagens computacionais baseadas em Mineração de Dados (MD), que visam extrair conhecimento útil e potencialmente relevante em conjuntos de dados, que crescem tanto em tamanho,

como em complexidade [9].

A MD consiste de uma etapa do processo de descoberta de conhecimento em bases de dados (*Knowledge Discovery in Databases* - KDD) [10], que define de uma sequência de etapas visando a extração de informações potencialmente úteis e relevantes em conjuntos de dados. O processo de KDD pode ser descrito pelas seguintes etapas: Seleção, pré-processamento, transformação, mineração de dados e interpretação [11]. Inicialmente, um subconjunto representativo de dados é selecionado do conjunto original. Em seguida, os dados são preparados para as etapas posteriores de acordo com uma metodologia de limpeza de dados específica ao domínio do problema, etapas conhecidas como pré-processamento e transformação. Seguido a isso vem a MD, que pode consistir de diversas tarefas, como detecção de dados atípicos, aplicação de algoritmos de agrupamento, regressão estatística e sumarização. Por fim, os dados e as abstrações criadas são interpretadas e o modelo é validado.

Como parte do processo de KDD, a Visualização de Dados (VD) é uma área interdisciplinar que propõe a utilização de abstrações visuais com o intuito de amplificar a cognição, aquisição ou o uso de conhecimento [12]. Essas abstrações são representações gráficas dos dados, ou *layouts*, que consistem em um mapeamento dos dados originais para elementos gráficos, como por exemplo linhas, pontos e formas geométricas. No processo de KDD, a VD pode ser integrada de três diferentes formas: (1) para visualizar antecipadamente os dados a serem analisados; (2) para ajudar no entendimento dos resultado de um processo de mineração de dados; (3) para auxiliar a análise de resultados parciais do processo de extração de conhecimento [13].

Particularmente, a detecção de fraudes e anomalias no presente trabalho é uma tarefa desafiadora, pois as instâncias do conjunto de dados de notas fiscais considerado nesta pesquisa não possuem informações de rótulo, inviabilizando o emprego de modelos de classificação ou outras abordagens supervisionadas. Por isso, a visualização se mostra uma ferramenta robusta para transmitir ideias, devido ao importante papel que desempenha na cognição humana [14].

Assim sendo, um processo de visualização pode ser especialmente útil no contexto de uma tarefa de detecção de fraudes, pois a VD pode ser considerada como o método mais intuitivo de validar agrupamentos [8] uma vez que ações fraudulentas tendem a exibir padrões de comportamento que permitem a sua partição em grupos com alta consistência entre si [15]. Dessa forma, a VD serve como um suporte ao especialista encarregado de encontrar dados atípicos, provendo a ele um entendimento visual e intuitivo dos padrões e relações de similaridade nos dados.

Nesse cenário, um processo de visualização exploratória de dados pode ser apropriado para a descoberta de conhecimento nas notas fiscais. O especialista pode iterativamente

elaborar hipóteses, gerar os *layouts* por meio das técnicas de visualização e utilizar recursos de interação para auxiliar na interpretação dos padrões globais e locais dos dados conforme as tarefas e objetivos previamente estabelecidos [16]. Nesse sentido, uma decisão importante para a visualização exploratória se refere às técnicas de visualização a serem empregadas. O foco dessa pesquisa foram as visualizações baseadas no posicionamento de pontos, em especial, as projeções multidimensionais [17]. As projeções são capazes de preservar as relações de similaridade dos dados originais no *layout* e por meio dessas projeções, se torna possível o descobrimento de conhecimento no conjunto de dados.

A presente pesquisa propõe um processo de visualização exploratória interativa de dados visando auxiliar o especialistas e auditores nas tarefas de identificar fraudes e anomalias em NFCes do Distrito Federal. O processo de visualização consiste em cinco etapas, envolvendo: o pré-processamento de dados; a criação de uma matriz de dissimilaridade a partir de um subconjunto representativo das notas fiscais; a utilização de um algoritmo de agrupamento (opcional) para a identificação e comparação das notas fiscais; geração dos *layouts* por meio das técnicas de visualização baseadas em projeções multidimensionais que utilizam a matriz de dissimilaridade calculada, e; visualização interativa dos dados com participação do especialista.

1.1 Hipótese de Pesquisa

A pesquisa descrita nesta monografia buscará responder a seguinte hipótese de pesquisa:

"É possível auxiliar a compreensão de um auditor fiscal com relação aos dados de NFCes por meio de um processo de visualização exploratória?"

1.2 Objetivos

O objetivo principal desse projeto é estudar e propôr um processo de visualização exploratória para a descoberta de conhecimento em conjuntos de dados relacionados com notas fiscais do consumidor eletrônicas, visando auxiliar os especialistas na detecção de fraudes e anomalias. Para cumprir esse objetivo geral, pode-se definir os seguintes objetivos específicos:

- Investigar as características das NFCes e explorar as funções de distância para viabilizar o cálculo das dissimilaridades entre elas;
- Estudar as projeção multidimensionais que podem ser incluídas no processo de visualização exploratória para gerar *layouts* intuitivos e que expressem as estruturas globais e locais das notas fiscais;

- Realizar experimentos de forma a validar o conhecimento obtido a partir do processo de visualização exploratória, verificando a possibilidade de detecção de fraudes e anomalias em notas fiscais.

1.3 Organização

Neste texto, o destaque é dado ao processo de visualização e as técnicas envolvidas. Esta monografia está organizada da seguinte maneira:

- O Capítulo 2 descreve a fundamentação teórica da pesquisa, detalhando conceitos relevantes para o entendimento de processamento de dados, visualização e mineração visual de dados;
- O Capítulo 3 revisa os artigos relacionados da literatura, com foco nas pesquisas de detecção de fraudes e visualização de dados;
- O Capítulo 4 descreve o processo de visualização exploratória para descoberta de conhecimento nos conjuntos de dados considerados nesta pesquisa: as notas fiscais eletrônicas do consumidor do Distrito Federal e um conjunto de transações de cartões de crédito;
- O Capítulo 5 apresenta os resultados experimentais e o conhecimento obtido a partir do processo de visualização exploratória nos conjuntos de dados mencionados;
- O Capítulo 6 conclui a pesquisa descrita nesta monografia e discute perspectivas de trabalhos futuros.

Capítulo 2

Fundamentação Teórica

A quantidade de dados armazenados por organizações e negócios está em constante crescimento devido à inovações em tecnologias de coleta, armazenamento e manutenção de dados. Ademais, uma característica que os conjuntos de dados geralmente compartilham é a sua grande quantidade de atributos, o que resulta em bancos de dados de alta dimensionalidade. Como exemplo, é possível citar a automação de processos financeiros como a criação de NFCes, a coleta de dados feita por empresas de marketing, censos demográficos, sistemas de monitoramento baseados em sensores, etc.

De forma a extrair informações de grandes conjuntos de dados, técnicas de Mineração Visual de Dados (MVD) provaram ser de grande valor [14]. Além de serem utilizadas para extrair informações úteis, são aplicadas também para auxiliar a identificação de padrões, tendências e evidenciar os relacionamentos implícitos contidos nos dados.

2.1 Fundamentos sobre Dados

Dados são conjuntos de valores quantitativos ou qualitativos sobre um objeto, que podem ser obtidos a partir de imagens, textos, sensores etc. De acordo com Bruce et al. (2020) [18], dados podem ser categorizados como:

- **Dados quantitativos:** São dados que podem ser contados, mensurados e expressos com números. São definidos de forma rígida, sem muito espaço para interpretação e obtidos através de experimentos, pesquisas e testes. Podem ser subdivididos em dados discretos, como inteiros; e dados contínuos, que podem ser infinitamente divididos em partes menores, como pontos flutuantes.
- **Dados qualitativos:** São dados não estruturados ou semiestruturados em natureza. Não podem ser expressos como um número, como por exemplo palavras, objetos, figuras, observações e símbolos. Variáveis categóricas são definidas como dados

qualitativos, apesar de poderem ter valores numéricos, operações normalmente feitas sob dados numéricos não fazem sentido se feitas em variáveis categóricas.

Nesta monografia, formaliza-se um conjunto de dados $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}\}$, em que uma instância multidimensional $\mathbf{x_i} = (x_{i,1}, x_{i,2}, ..., x_{i,m})$ é caracterizada por m atributos, que podem ser categóricos, nominais, ordinais etc [19]. Nesse sentido, adota-se a representação dos dados conforme o Modelo Tabular, que consiste de matrizes simples, bidimensionais, compostas por instâncias de dados, muitas vezes de atributos diferentes. No modelo tabular, geralmente cada linha compõe uma unidade de dado e cada coluna se refere a um atributo.

Em processos de mineração de dados e visualização, é comum que os dados sejam utilizados em formato tabular, como por exemplo, o espaço de características associado aos dados, ou pela matriz de dissimilaridades calculada pelo cálculo da distância entre pares de instâncias.

2.1.1 Métricas de Dissimilaridade

De maneira geral, dissimilaridades correspondem a números não negativos $d(\mathbf{x_i}, \mathbf{x_j})$ que são pequenos quando $\mathbf{x_i}$ e $\mathbf{x_j}$ são semelhantes e se tornam grandes quando i e j são diferentes. Medidas de dissimilaridade são usualmente simétricas e o valor dessa métrica de um objeto em relação a si mesmo é zero. Métricas de dissimilaridade tem aplicação em vários algoritmos de aprendizado de máquina, e podem ser utilizadas para a criação de matrizes de dissimilaridade.

Matrizes de dissimilaridade são estruturas de dados utilizadas como entrada para outros algoritmos [20], onde cada instância corresponde a uma medida de dissimilaridade entre pares de objetos. Uma medida de dissimilaridade é calculada a partir de uma função de distância. Por fim, é importante notar que matrizes de dissimilaridade apresentam uma complexidade espacial de $\mathcal{O}(n^2)$, o que pode representar um grande gargalo dependendo da aplicação.

Distância Euclidiana

A distância euclidiana entre dois pontos representa a distância entre dois pontos em um plano cartesiano, calculada utilizando o teorema de pitágoras, definido na Equação 2.1. Com o uso dessa função, o espaço euclidiano torna-se um espaço métrico. Essa equação pode ser aplicada a espaços de diferentes dimensões com fórmulas semelhantes.

$$d(x_i, x_j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2}. \tag{2.1}$$

Distância Manhattan

A distância manhattan é definida como a soma das diferenças absolutas entre coordenadas em um plano cartesiano. Mais formalmente, a distância d_1 entre dois vetores $\mathbf{x_i}$ e $\mathbf{x_j}$ em um espaço vetorial m-dimensional é definida como

$$d(\mathbf{x_i}, \mathbf{x_j}) = ||\mathbf{x_i} - \mathbf{x_j}||_1 = \sum_{k=1}^{m} |x_{i,k} - x_{j,k}|, \tag{2.2}$$

onde (xi, xj) são vetores

$$x_i = (x_{i,1}, x_{i,2}, ..., x_{i,n}) \; e \; q = (x_{j,1}, x_{j,2}, ..., x_{j,n}). \tag{2.3}$$

Distância de Gower

A distância de Gower [21] permite que usuários extraiam informações de conjuntos de dados, cujas instâncias apresentam atributos numéricos, categóricos, binários, etc. O cálculo da função de dissimilaridade entre duas instâncias é dado pela média ponderada da contribuição de cada variável. Como mostra a Equação 2.4:

$$d_{ij} = d(i, j) = \frac{\sum_{k=1}^{p} w_k \delta_{ij}^{(k)} d_{ij}^{(k)}}{\sum_{k=1}^{p} w_k \delta_{ij}^{(k)}}. \tag{2.4}$$

Em outras palavras, $\mathbf{d_{ij}}$ é uma média ponderada de $\mathbf{d_{ij}^{(k)}}$ com pesos $\mathbf{w_k}\delta_{\mathbf{ij}}^{(\mathbf{k})}$, onde $\mathbf{w_k}$ representa o peso da k-ésima variável, $\delta_{\mathbf{ij}}^{(\mathbf{k})}$ pode assumir os valores de 0 ou 1, e $\mathbf{d_{ij}^{(k)}}$ representa a k-ésima variável.

A contribuição $\mathbf{d_{ij}^{(k)}}$ de uma variável categórica ou binária vale 0 se os valores forem iguais, 1 se forem diferentes. Para variáveis de tipos numéricos, a contribuição $\mathbf{d_{ij}^{(k)}}$ assume o valor da distância manhattan (Equação 2.2) dividida pelo intervalo da variável. Note que se a contribuição $\mathbf{d_{ij}^{(k)}}$ permanece no intervalo $[0, 1]$, a dissimilaridade $\mathbf{d_{ij}}$ também permanecerá.

2.2 Análise de Agrupamentos

A análise de agrupamentos pode ser definida como um conjunto de técnicas utilizadas para agrupar objetos em conjuntos relacionados entre si chamados de *clusters*. Na ausência de dados classificados, os algoritmos de agrupamento podem ser úteis para gerar modelos concisos dos dados, que podem ser interpretados como um resumo do modelo generativo dos dados não-rotulados [22]. É utilizado para tarefas tais como mineração de dados, reconhecimento de padrões, análise de dados estatística, análise de imagens, compressão

de dados e aprendizado de máquina em geral. Esses algoritmos envolvem um processo que consiste em: Formular um problema; selecionar uma medida de distância apropriada; escolher um algoritmo de agrupamento; decidir no número de grupos previamente e então após o funcionamento do algoritmo, validar os seus resultados. O objetivo final de um algoritmo de agrupamento é garantir que instâncias de um conjunto de dados sejam posicionadas no mesmo grupo de instâncias similares [23].

Algoritmos de agrupamento são divididos em algoritmos de particionamento, hierárquicos, de malha, baseados em modelos e baseados em densidade. Esse trabalho explora algoritmos de agrupamento focando em algoritmos de particionamento. Métodos de particionamento envolvem a criação de subdivisões de dados que dependem de certos critérios objetivos, tais como a minimização do erro quadrado [24]. Um algoritmo de particionamento utilizado com frequência em aplicações práticas é o k-means [25], que basicamente consiste na definição de grupos tal que a variação entre-grupos (ou variação dentro de grupos) seja minimizada [26]. O algoritmo padrão define a variação dentro de grupos como a soma das distâncias euclidianas quadradas entre itens e o centróide correspondente, em que um centróide corresponde ao centro geométrico de um grupo [27].

2.2.1 Partitioning Around Medoids

Embora o k-means seja amplamente utilizado, nesta pesquisa decidimos utilizar uma alternativa robusta do k-means chamada k-medoids, também conhecido como Partitioning Around Medoids (PAM). O algoritmo PAM difere do k-means por estabelecer grupos por meio da identificação de medóides ao invés de centróides [26]. Um medóide é caracterizado por ser um objeto dentro de um grupo que contém uma dissimilaridade mínima entre ele e os outros objetos do grupo, dessa forma maximizando a coesão dentro do grupo. São os pontos mais próximos do centro de um grupo que são necessariamente objetos dentro do conjunto de dados. Um centróides difere de um medóide pelo fato de que centróides podem não ser necessariamente objetos no conjunto de dados.

Dessa forma, utilizaremos o algoritmo PAM que se mostra menos sensível a dados atípicos, pois minimiza a soma das dissimilaridades entre pares ao invés das distâncias euclidianas quadradas. Além disso, também é considerado um algoritmo bem adequado para realizar o agrupamento sobre matrizes de dissimilaridade [28], tais como as matrizes geradas com a distância de Gower.

O algoritmo k-medoids ou *Partitioning Around Medoids* (PAM) é um algoritmo de aprendizado de máquina não supervisionado que funciona por particionamento tal como k-means, porém os grupos são formados ao redor de medóides, que são instâncias dentro do conjunto de dados [29].

Sua complexidade temporal é dada por $\mathcal{O}(k * (n-k)^2)$, porém uma implementação ingênua que recompute a função de custo em toda iteração terá $\mathcal{O}(n^2k^2)$. O custo computacional pode ser ainda reduzido para $\mathcal{O}(n^2)$ através da partição do custo, tal que algumas computações possam ser divididas ou ignoradas [30]. Por fim, a qualidade dos agrupamentos gerados pelo algoritmo podem ser validados utilizando algumas métricas como o coeficiente de silhueta.

2.2.2 Coeficiente de Silhueta

O coeficiente de silhueta pode ser utilizado para a validação da qualidade de um agrupamento, combinando métricas de coesão e separação para objetos e agrupamentos [19]. Seu valor pode variar entre -1 e 1, sendo que um valor negativo é indesejável, pois indica que a coesão entre objetos de um mesmo agrupamento é baixa.

O cálculo de do coeficiente pode ser feito em três etapas:

1. Para o i-ésimo objeto, calcular sua distância média em relação a todos os outros objetos do agrupamento; nomear esse valor de $\mathbf{a_i}$.
2. Para o i-ésimo objeto e qualquer agrupamento que não contenha o objeto, calcular a distância média do objeto para com todos os outros objetos do agrupamento. Encontrar o valor mínimo com respeito a todos os clusters; nomear esse valor de $\mathbf{b_i}$.
3. Para o i-ésimo objeto, o coeficiente de silhueta vale $\mathbf{s_i} = (\mathbf{b_i} - \mathbf{a_i})/max(\mathbf{a_i}, \mathbf{b_i})$.

Um exemplo de gráfico com coeficientes de silhueta é fornecido na Figura 2.1, neste caso, a análise é ambivalente entre 2 e 3 grupos.

2.3 Visualização da Informação

Visualização pode ser definida como a comunicação de informação através do uso de representações gráficas [1]. Pode ser interessante considerar o número de tipos de dados e visualizações com as quais seres humanos lidam diariamente, como por exemplo um mapa de GPS, um raio X ou um gráfico meteorológico. Em todos os casos, representações gráficas são usadas como auxílio ao entendimento de alguma informação textual ou verbal.

O ser humano é uma entidade inerentemente visual e utiliza com frequência suas habilidades de percepção para auxiliar o seu processo de tomada de decisão. Porém, algumas informações não são imediatamente visíveis, como as informações implícitas em um conjunto de dados.

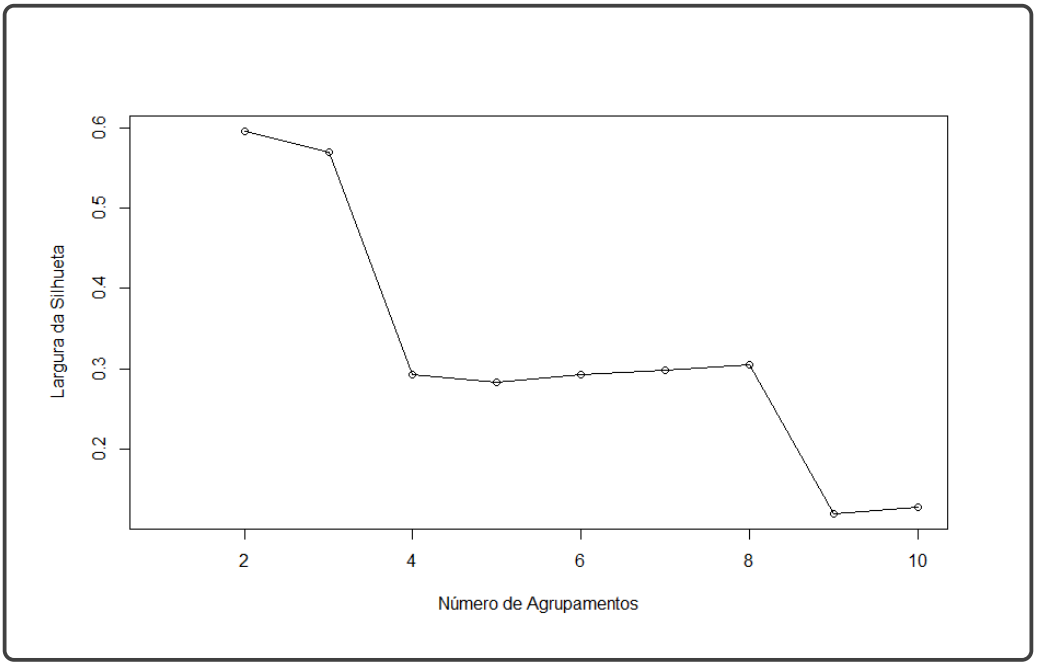


Figura 2.1: Exemplo de gráfico de silhueta do conjunto de dados de NFCes.

O volume de dados produzidos cresce constantemente, dados de navegação e diversas interações são armazenados em bancos de dados, com intuito de se obter vantagem competitiva. Desse modo onde mesmo simples transações, chamadas de telefone ou acessos à internet são registrados em bancos de dados, aumenta também o potencial de descobrimento de potenciais informações implícitas. Desse modo, surge também uma demanda crescente por ferramentas e técnicas que auxiliem na comunicação dessas informações de maneira efetiva. Com o objetivo de suprir essa demanda, o campo de visualização de dados propõe o uso de processos de visualização, como pode ser visto na Figura 2.2.

Um processo de visualização consiste de etapas para visualizar dados de maneira efetiva, em que o usuário se faz presente em todos os estágios [1]. As etapas são descritas abaixo e se assemelham com as etapas de um processo de KDD:

- **Pré-processamento:** Ao iniciar o processo, os dados podem necessitar de tratamento pela presença de anomalias, como dados ausentes, atributos com formatações diferentes, dados redundantes, etc. Desse modo, o primeiro passo do processo é tratar os dados e os mapear para tipos fundamentais de forma que sejam manipuláveis por técnicos de visualização. Em grandes conjuntos de dados, pode ser necessário fazer um processo de amostragem, filtragem, agregação ou particionamento.
- **Mapeamentos visuais:** Uma vez que os dados estejam devidamente tratados, o especialista a cargo de criar a visualização decide as características visuais, como geometria e cor da representação gráfica. As decisões tomadas nesta etapa são de

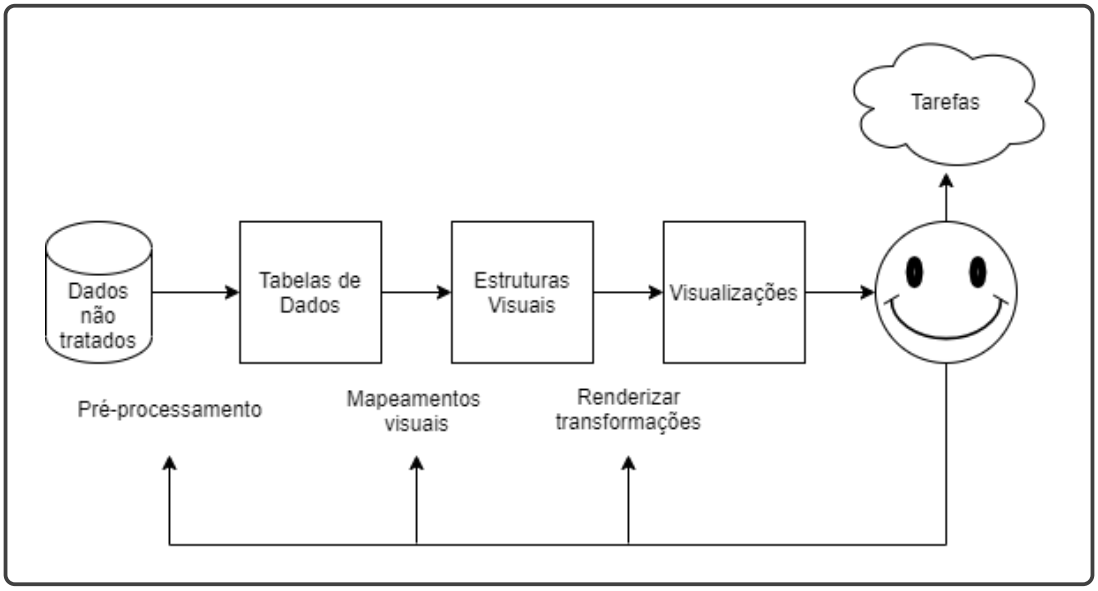


Figura 2.2: Fluxograma detalhando um processo de visualização básico. Adaptado de [1].

suma importância, pois uma visualização de baixa qualidade pode, além de não comunicar informações, levar a interpretações errôneas dos dados.

- **Renderizar transformações:** O estágio final corresponde ao mapeamento de dados geométricos para a imagem. Isso inclui fazer uma interface com uma *Application Programming Interface* (API). Nesta etapa se decidem os parâmetros de visualização, técnicas de *shading*, etc.

2.3.1 Técnicas de Visualização

Existem na literatura diversas técnicas de visualização com aplicações em campos de conhecimento como biologia [31], astronomia [32] até ciências sociais [33]. Com o tempo, novas técnicas de visualização são criadas para atender demandas diferentes, sendo que alguns problemas complexos requerem a aplicação de representações gráficas específicas para possibilitar uma visualização eficiente, como por exemplo o uso de mapas em aparelhos de GPS.

Keim [16] designa um esquema de classificação para sistemas de visualização baseado em três dimensões: tipos de dados a ser visualizados, técnicas de visualização, e métodos de interação/distorção. Em técnicas de visualização, são traçadas distinções entre representações 2D/3D, transformadas geometricamente, baseadas em ícones, densas de pixels.

Gráficos de Dispersão

Um gráfico de dispersão (ou gráfico X-Y) é uma representação gráfica de duas dimensões em que cada marcador (que pode variar entre diversos símbolos) representa uma instância de do conjunto de dados associado. A posição do marcador no gráfico normalmente indica o seu valor, porém, em algumas aplicações como em visualizações de projeções multidimensionais, a posição relativa dos dados expressam significado, como as relações de similaridade dos dados, além de padrões locais e globais.

Gráficos de dispersão são úteis para a examinação de relacionamentos e correlações entre variáveis. Por exemplo, é possível observar em um gráfico de dispersão que algumas variáveis tais como *oferta* e *demanda*, apresentam uma dependência mútua, ou seja, a mudança do valor de uma variável tem a capacidade de exercer influência sobre o valor de outra variável. Além disso, os gráficos de dispersão são utilizados com frequência para plotar regressões estatísticas e gráficos de correlação. Um exemplo é dado na Figura 2.3, onde é possível observar a correlação entre as instâncias do conjunto de dados *mtcars* [2].

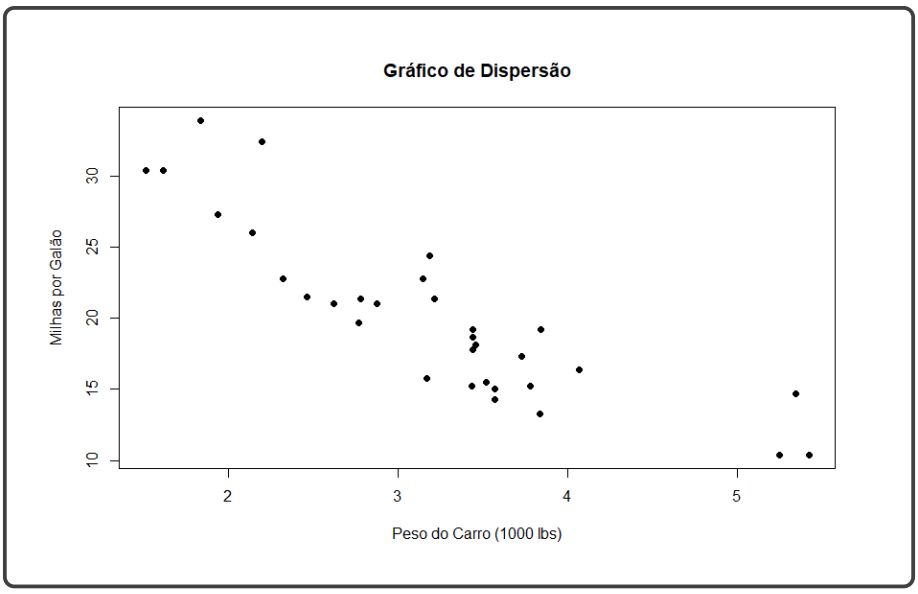


Figura 2.3: Exemplo de gráfico de dispersão do conjunto de dados *mtcars* [2].

Além do gráfico de dispersão, existem outras técnicas de visualização clássicas que são populares em tarefas de análise visual de dados, como o mapa de calor [34] e as coordenadas paralelas [35]. No entanto, essas técnicas apresentam limitações para produzir *layouts* intuitivos e que preservem os padrões e as estruturas de dados multidimensionais, dessa forma, técnicas de visualização alternativas foram criadas de modo a suprir essas limitações.

Uma alternativa se refere ao uso de visualizações baseadas no posicionamento de pontos no espaço visual, que podem ser categorizadas em projeções multidimensionais ou árvores

de similaridades [36]. Nessa monografia, serão consideradas as visualizações baseadas em projeções multidimensionais, porque demonstram utilidade em processos de visualização de dados por similaridade devido ao modo intuitivo representam os dados, sendo assim interessantes a uma tarefa de detecção de anomalias.

2.3.2 Projeções Multidimensionais

Projeções multidimensionais são algoritmos que mapeiam dados m-dimensionais em um espaço p-dimensional, com $p = \{1, 2, 3\}$, de forma a preservar a estrutura inerente dos dados e suas relações de distância. Dessa forma, os dados podem ser visualizados de forma intuitiva, revelando informações e padrões implícitos.

Formalmente, uma técnica de projeção multidimensional é definida como [37]:

Definição 1 (Projeção Multidimensional) *Seja* $\boldsymbol{X}$ *um conjunto de objetos em* $\mathbb{R}^m$ *com* $\delta : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ *um critério de proximidade entre objetos em* $\mathbb{R}^m$*, e* $\boldsymbol{Y}$ *um conjunto de pontos em* $\mathbb{R}^p$ *para* $p = \{1, 2, 3\}$ *e* $d : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ *um critério de proximidade em* $\mathbb{R}^p$*. Uma técnica de projeção multidimensional pode ser descrita como uma função* $f : \mathbf{X} \to \mathbf{Y}$ *que visa tornar* $|\delta(\mathbf{x_i}, \mathbf{x_j}) - d(f(\mathbf{x_i}, f(\mathbf{x_j}))|$ *o mais próximo de zero,* $\forall \mathbf{x_i}, \mathbf{x_j} \in \mathbf{X}$*.*

Embora a literatura possua diversas maneiras de categorizar as projeções multidimensionais, nessa monografia adota-se a categorização proposta por Paulovich et al. [36]:

- **Force Directed Placement:** são técnicas que simulam um sistema composto por objetos sob a ação de forças de atração e repulsão. São exemplares as técnicas do *Spring Model* [38], *Hybrid Model* [39] e *Force Scheme* [37].
- **Multidimensional Scaling:** são técnicas utilizadas para mapear um espaço multidimensional em um espaço de dimensionalidade reduzida, porém preservando a sua estrutura e a relação de distância entre instâncias. São exemplares as técnicas *Classical Scaling* [40], *Isometric Mapping* [41] e *Sammon's Mapping* [40].
- **Redução de Dimensionalidade:** são técnicas utilizadas para mapear um espaço multidimensional em um espaço de dimensionalidade reduzida, preservando alguma característica do espaço original, no caso do presente trabalho, relações de distância. São exemplares as técnicas *Principal Component Analysis* [42], *Local Linear Embedding* [43] e *t-distributed Stochastic Neighbor Embedding* [44].

Multidimensional Scaling

Multidimensional Scaling (MDS) é uma técnica de projeção multidimensional não linear, que cria um mapa de posições relativas dos objetos de conjunto de dados, dada apenas uma matriz de distâncias como entrada. É uma forma de visualizar o nível de similaridade de objetos em um conjunto de dados. A técnica também é utilizada para traduzir informações sobre distâncias entre pares de pontos em uma configuração de n pontos mapeada em um espaço cartesiano abstrato [45]. Assim sendo, dada uma matriz de proximidade, o algoritmo MDS posiciona objetos em um espaço n-dimensional de tal forma a preservar a distância entre esses objetos.

O MDS pode ser métrico, reproduzindo as distâncias originais dos dados, ou pode ser não métrico, assumindo que os postos matriciais são conhecidos. Dessa forma, o MDS não métrico produz um mapa que reproduz esses postos, sendo que as distâncias não são reproduzidas. O MDS métrico reproduz relacionamentos lineares entre os dados, enquanto que o MDS não métrico reproduz uma série de curvas que dependem apenas do valor dos postos.

O algoritmo contém duas etapas. A primeira etapa consiste em converter uma matriz de entrada D em uma matriz de produto cartesiano, ou matriz Gram B. A segunda etapa, que produz o gargalo do algoritmo, consiste na completa decomposição espectral da matriz B que apresenta uma complexidade de $\mathcal{O}(n^3)$ [46]. Um exemplo de *layout* gerado utilizando uma visualização baseada em MDS com o conjunto de dados Íris [2] pode ser visto na Figura 2.4. O conjunto Íris possui 150 instâncias caracterizadas por quatro atributos e categorizadas em três espécies, que representam as cores dos pontos.

Isometric Mapping

Isometric Mapping (ISOMAP) é um algoritmo de projeção multidimensional não linear utilizado para computar uma transformação de um espaço altamente dimensional para um espaço quasi-isométrico de menor dimensão. O algoritmo estende o MDS aplicando o conceito de distâncias geodésicas impostas por um grafo ponderado. Dessa forma, se comparado com o MDS clássico, sua diferença ocorre na construção da sua matriz de distância. Enquanto que o MDS utiliza distâncias euclidianas, o ISOMAP utiliza a distância entre pontos que são o peso de um grafo valorado. Essa distância tem o potencial de capturar mais adequadamente a estrutura implícita dos dados do que a distância euclidiana.

O algoritmo é composto por três estágios: o primeiro estágio consiste na procura pelo vizinho mais próximo; o segundo estágio, a procura pelo menor caminho em um grafo; e o terceiro estágio, a decomposição parcial de autovalores. Para o primeiro estágio, a

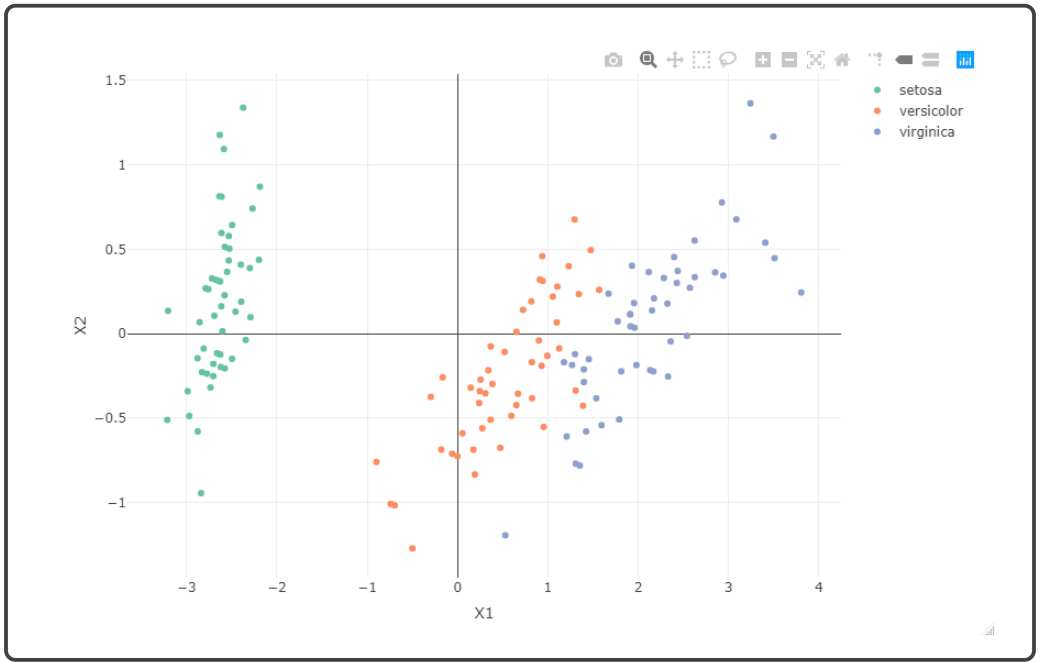


Figura 2.4: Gráfico gerado via MDS do conjunto de dados Íris.

complexidade é $\mathcal{O}[mlog(k)Nlog(n)]$ para k primeiros vizinhos de n pontos em m dimensões. Já na segunda etapa, para a procura do menor caminho em um grafo são utilizados ou o algoritmo de *Dijkstra*, que é $\mathcal{O}[n^2(k + log(n))]$ ou *Floyd-Warshall* que é $\mathcal{O}[n^3]$. Por fim, a terceira etapa contém uma complexidade de $\mathcal{O}[dn^2]$, em que d corresponde aos d maiores autovalores do kernel. A complexidade geral do algoritmo então corresponde a $\mathcal{O}[mlog(k)Nlog(n)] + \mathcal{O}[n^2(k + log(n))] + \mathcal{O}[dn^2]$ [47]. Um exemplo de um *layout* gerado pela visualização baseada em ISOMAP utilizando o conjunto de dados *Íris* pode ser visto na Figura 2.5.

Redução de Dimensionalidade

Reduzir a dimensionalidade de um conjunto de dados consiste em fazer uma conversão de um espaço altamente dimensional em um espaço de dimensões reduzidas que preferivelmente mantenha a estrutura do conjunto de dados original. Algoritmos nesse campo de estudo são normalmente utilizados quando o processamento de um conjunto de dados de alta dimensionalidade é indesejável [48], tais como conjuntos de dados massivos e com grandes quantidade de atributos.

Esses algoritmos são importantes pois tratam de várias características desagradáveis de dados altamente dimensionais, tais como a maldição da dimensionalidade [49] e dados esparsos. Além disso, também permitem a visualização de conjuntos de dados através

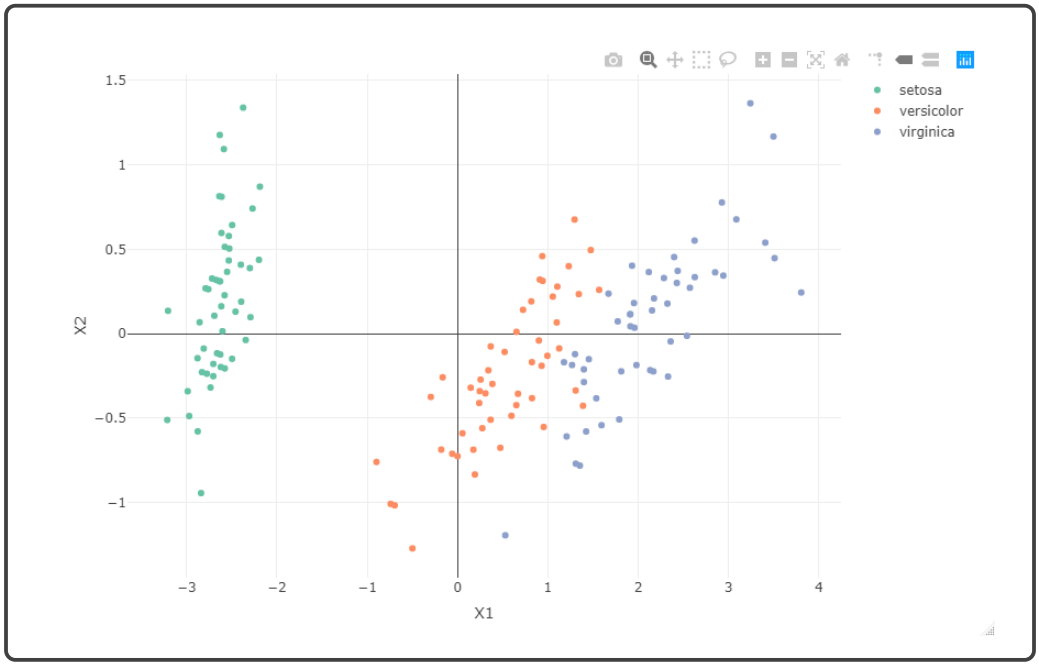


Figura 2.5: Gráfico gerado via ISOMAP do conjunto de dados *Íris* [2].

da redução das dimensões de conjuntos de dados para 2 ou 3 dimensões, que é o foco explorado nesta pesquisa.

Os algoritmos de redução de dimensionalidade são divididos em lineares e não lineares [48]. Algoritmos lineares reduzem as dimensões de dados como uma combinação linear das variáveis originais. Esses algoritmos são aplicáveis quando os dados pertencem a um subespaço linear, e assim sendo, as variáveis originais são substituídas por um conjunto menor de variáveis. Algoritmos não lineares são empregados quando os dados contém relacionamentos não lineares entre si. Dessa forma, a representação com menor número de dimensões é alcançada preservando as distâncias originais entre os dados.

t-Distributed Stochastic Neighbor Embedding

Introduzido por van de Maaten e Hinton em 2008 [44], t-distributed Stochastic Neighbor Embedding (t-SNE) é uma técnica de redução de dimensionalidade não supervisionada e não linear, utilizada para a exploração visual de dados altamente dimensionais por meio do mapeamento de dados em duas ou três dimensões. Essa técnica é uma variação do *Stochastic Neighbor Embedding* [50] que se mostra muito mais fácil de otimizar e produz melhores visualizações devido a sua tendência de diminuir o acúmulo de pontos no meio do mapa. Sobretudo, em comparação com outros métodos não paramétricos de visualização, t-SNE também tende a produzir melhores representações e apresenta uma performance melhor em relação à outras técnicas de visualização [44].

O algoritmo é eficiente não apenas para capturar a estrutura local de um conjunto de dados de alta dimensionalidade, mas também para encontrar a estrutura global dos dados, apresentando diversos grupos em várias escalas. Em particular, a maioria das técnicas de redução de dimensionalidade não é capaz de obter tanto a estrutura local quanto a estrutura global de um conjunto de dados.

A técnica de modo geral funciona convertendo a afinidade de pontos em probabilidades. As afinidades no espaço original são representadas por uma distribuição normal, enquanto que as afinidades na nova representação dos dados são representadas por uma distribuição t-student. Isso permite ao algoritmo ter uma sensibilidade à estrutura local dos dados, além de tornar possível a visualização de padrões que se situam em diversos diferentes grupos contidos no conjunto.

Dada uma matriz D entre objetos de entrada, o algoritmo calcula um coeficiente de similaridade no espaço original p_{ij} conforme mostra a Equação (2.5):

$$p_{j|i} = \frac{exp(-||D_{ij}||^2/2\sigma_i^2)}{\sum_{k \neq i} exp(-||D_{ij}||^2/2\sigma_i^2)}, \tag{2.5}$$

que então se torna simétrico a partir da Equação 2.7:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n} \tag{2.6}$$

σ é escolhido para cada objeto de tal forma que o parâmetro de perplexidade de $p_{j|i}$ tenha um valor que seja próximo ao valor definido pelo usuário. O valor do parâmetro perplexidade controla quantos vizinhos são levados em consideração quando é feita a construção do *embedding* no espaço de baixa dimensionalidade. Para o espaço de baixa dimensionalidade, é utilizada a distribuição de Cauchy (t-student com um grau de liberdade) dada pela Equação (2.7):

$$q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k \neq l}(1 + ||y_k - yt||^2)^{-1}} \tag{2.7}$$

Através da mudança de objetos no *embedding* para minimizar a divergência de *Kullback-Leibler* (K-L) entre as distribuições q_{ij} e p_{ij}, um mapa é criado que enfatiza a estrutura de pequena escala, devido a assimetria da divergência de K-L. A distribuição de t-student é escolhida de modo a evitar o problema da aglomeração, pois em um espaço com muitas dimensões, existem potencialmente vários objetos equidistantes com uma distância moderada de um certo objeto, mais do que poderia ser levado em conta num espaço de poucas dimensões. Dessa forma, a distribuição t-student é utilizada para espalhar esses objetos no novo espaço dimensional.

A minimização da divergência de K-L pode ser feita por meio de um gradiente descendente, o que leva a um gargalo no algoritmo quando aplicado à grandes conjuntos de dados devido a complexidade de $\mathcal{O}(n^2)$ do gradiente. Sendo assim, as implementações comumente utilizadas em bibliotecas de ciência de dados contém a otimização de *Barnes-Hut* que funciona através de dois mecanismos: Primeiramente o algoritmo aproxima as similaridade por 0 na distribuição p_{ij}, em que as entradas não nulas são processadas encontrando $3 * perplexidade$ vizinhos e utilizando uma busca eficiente em árvore. Em seguida, utiliza-se o algoritmo de *Barnes-Hut* na computação do gradiente que aproxima grandes distâncias utilizando uma estrutura de dados de árvore chamada quadtree. Essa aproximação é controlada por um parâmetro, em que menores valores levam a aproximações mais exatas. Essa implementação do algoritmo com o uso da otimização de *Barnes-Hut* possui uma complexidade temporal de $\mathcal{O}(nlog(n))$ para cada iteração [51]. Um exemplo de *layout* gerado pela técnica de visualização t-SNE utilizando o conjunto de dados *Íris* pode ser visto na Figura 2.6.

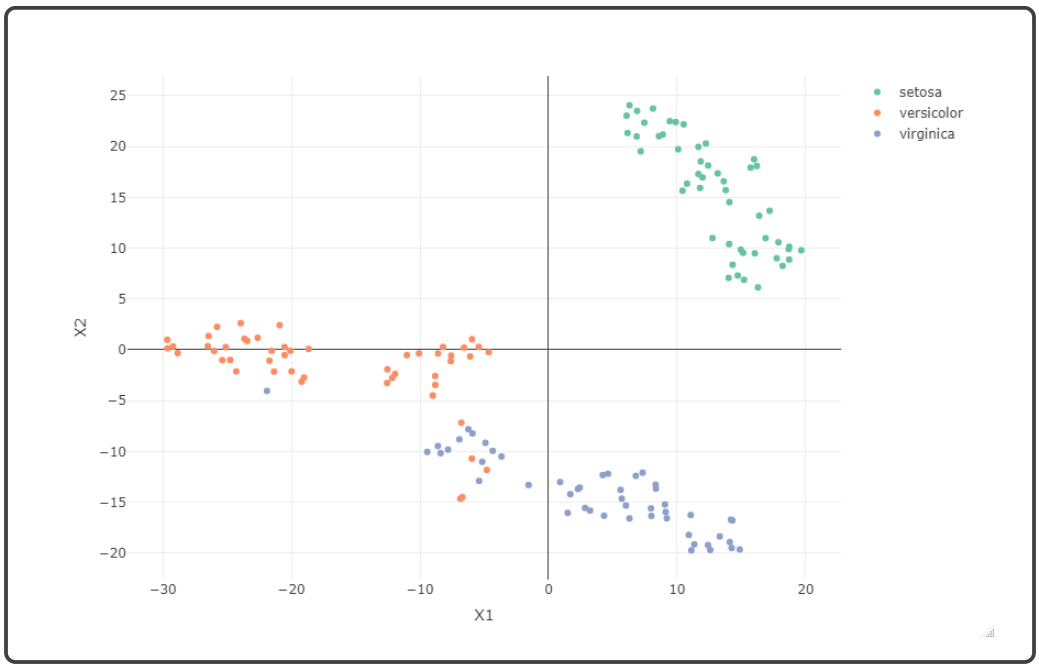


Figura 2.6: Gráfico gerado via t-SNE do conjunto de dados *iris* [2].

2.4 Considerações Finais

Neste capítulo foi apresentada uma base teórica sobre as técnicas aplicadas no presente trabalho. É possível notar que um processo de visualização é uma tarefa complexa, em que cada etapa do processo possui relevância e influência sobre o resultado final. O pré-processamento é de importante e se trata da etapa que mais consome tempo em um

processo de análise de dados [52]. Técnicas como algoritmos de redução de dimensionalidade e projeções multidimensionais também contém diversas características que devem ser compreendidas para uma visualização bem sucedida, como por exemplo os seus parâmetros. Por fim, o modo como um analista representa os dados também é importante, pois a tarefa de comunicar informações de maneira efetiva depende do uso de representações gráficas adequadas.

Capítulo 3

Revisão da Literatura

Este capítulo apresenta a descrição e análise de quatro artigos científicos da área de visualização de dados com a finalidade de auxiliar o processo de detecção de fraudes.

3.1 Trabalhos relacionados

Dilla e Raschke [53] desenvolveram um *framework* teórico para prever quando e como investigadores devem usar técnicas de visualização para detectar transações fraudulentas. Os autores declaram que analistas de dados financeiros estão percebendo a importância de processos de visualização como auxílio a tarefa de detecção de fraudes, porém reconhecem que estudos na área são extremamente limitados. Dessa forma, utilizando a teoria do ajuste cognitivo, desenvolveram o *framework* como forma de apresentar um conjunto de proposições e hipóteses de pesquisa para trabalhos futuros, com o objetivo de auxiliar na descoberta de circunstâncias onde processos de visualização interativa possam aumentar a eficiência dos analistas e a sua efetividade em detectar fraudes.

Chang et al. [54] apresenta uma aplicação chamada *WireVis*, que propõe um conjunto de visualizações coordenadas, em que cada visualização é baseada na identificação de uma palavra-chave específica em transações eletrônicas. As visualizações propostas descrevem relações entre os perfis de usuários e palavras-chave ao longo do tempo. Ademais, os autores introduzem uma técnica de procura-pelo-exemplo, que revela perfis com padrões de transações similares.

O software foi colocado em prática e teve boas primeiras impressões de profissionais da área. Com o *feedback* recebido, os analistas propuseram novas funcionalidades ao software, como deixar o programa naturalmente extensível para novas técnicas. Dessa forma, o analista pode considerar diferentes abordagens de agrupamentos dependendo dos seus objetivos. O sistema se mostra robusto e promissor, pois pode ser conectado a

um conjunto de dados massivo sem perda de performance, preservando alta interatividade com o analista.

Sun et al. [3] propõe um sistema chamado *FraudVis* para analisar visualmente dados fraudulentos com o uso de algoritmos de aprendizado de máquina não supervisionado. O sistema auxilia os analistas a entenderem os algoritmos utilizados e também fornece suporte ao ajuste fino dos parâmetros. Os algoritmos utilizados analisam fraudes por um viés temporal, correlação intragrupo e intergrupo, seleção de características e perspectivas individuais de cada usuário. No artigo, os autores relatam a solução de dois estudos de casos reais de detecção de dados atípicos utilizando o sistema proposto.

O trabalho apresenta alguns conceitos interessantes nos seus estudos de casos, como por exemplo, as características observáveis em dados atípicos. Um desses foi o chamado *"Silence and Burst"*, que descreve comportamentos atípicos que podem ser intermitentes, isto é, tais comportamentos podem estar presentes ou não em diferentes momentos. Outra característica interessante apresentada foi a *"Correlation Property for Intra-Group"* que descreve como grupos de comportamento aberrante podem ser identificados por meio das suas intra-relações. Por fim, eles também relatam *"Inter-group Analysis"*, quer dizer, as relações que usuários legítimos têm com outros usuários legítimos tende a espalhá-los em um gráfico, porém, usuários de um grupo fraudulento tendem a apresentar grupos com maior coesão, indicando que seus nós são realmente similares de alguma forma. O sistema é detalhado na Figura 3.1.

Deng e Ruan [4] propõe uma aplicação denominada *FraudJudger* para detectar usuários fraudulentos de forma semi-supervisionada, ou seja, com um número limitado de dados rotulados. Os autores admitem que a criação de conjuntos de dados rotulados é uma tarefa extremamente dispendiosa, dessa forma, propõem uma aplicação com a capacidade de aprender as representações latentes de usuários através de dados não rotulados, fazendo uso de *Adversarial Autoencoders* (AAE). Ademais, a aplicação encontra outros padrões fraudulentos utilizando algoritmos de agrupamento e faz a visualização desses agrupamentos utilizando o algoritmo t-SNE. Os *layouts* gerados são vistos na Figura 3.2.

Uma característica importante do sistema é a sua capacidade de atualizar os rótulos de seus conjuntos de dados de forma independente, sendo assim diminuindo a necessidade de trabalho manual para rotular os dados ao longo do tempo. Por fim, os autores realizaram experimentos em conjuntos de dados reais de transações eletrônicas e obtiveram bons resultados, considerando seu sistema como melhor do que outros sistemas de detecção de fraudes previamente existentes.

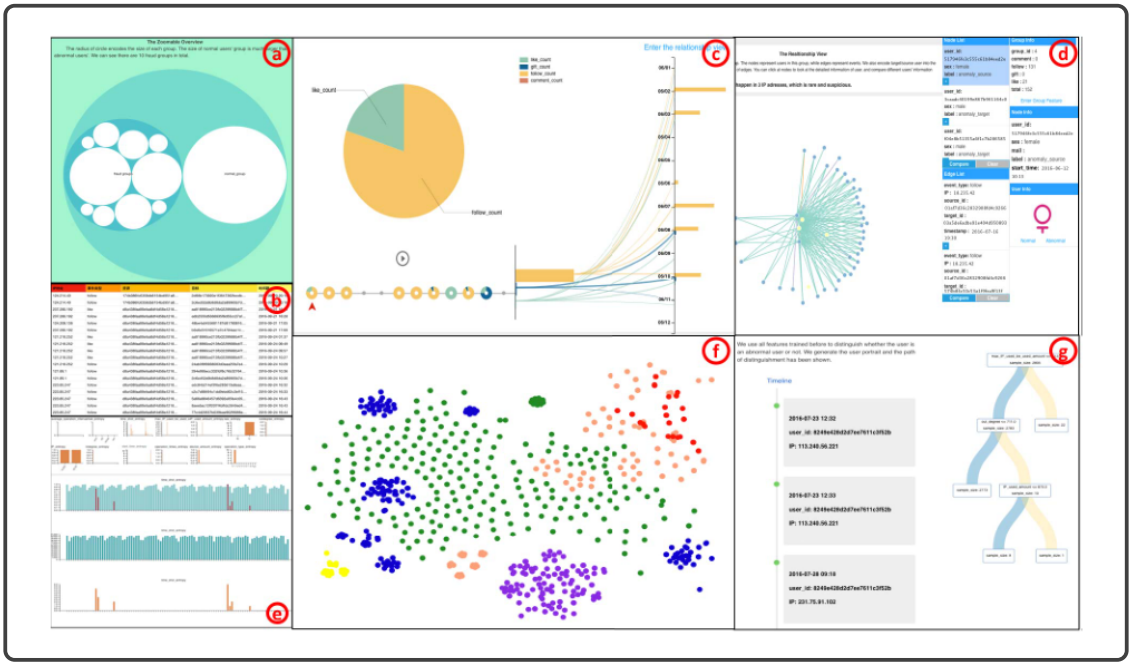


Figura 3.1: Imagem detalhando as interfaces do sistema *FraudVis* [3]. (a) Grupos de fraude global, (b) Dados não tratados, (c) Visualização de atividades de grupos em uma sequência temporal de seus comportamentos, (d) Visualização da interação entre usuários indicando seus relacionamentos dentro dos grupos, (e) Comparação das características que contribuem mais para a detecção do resultado em diferentes escalas, (f) Comparação inter-grupo em cinco grupos mais similares, (g) Visualização em gráfico de árvore.

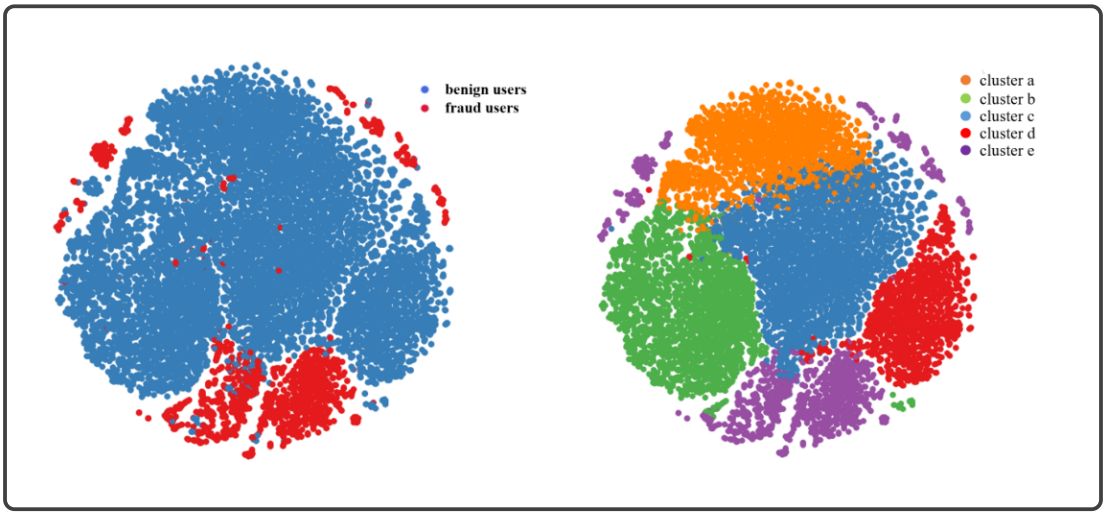


Figura 3.2: Visualizações geradas com o uso do t-SNE pela aplicação *FraudJudger* [4].

3.2 Considerações Finais

Os artigos apresentados neste capítulo são todos da área de visualização de dados com foco em detecção de fraudes. Dos quatro artigos apresentados, três são de aplicação prá-

tica, onde sistemas computacionais são criados a fim de atender demandas específicas de detecção de dados atípicos; e um apresenta um viés teórico, propondo avançar o conhecimento da área através da criação de uma análise para auxiliar trabalhos futuros em sua tomada de decisão.

Os três sistemas de aplicação prática propostos demonstram a efetividade da utilização de algoritmos de visualização em conjunto com algoritmos de agrupamento, o que é de grande importância para o presente trabalho, especialmente a aplicação do algoritmo t-SNE e sua aplicação na visualização de agrupamentos. Além disso, os trabalhos lidam com tipos de fraudes específicas e criam processos de visualização de forma a atender demandas igualmente específicas.

No presente trabalho, o processo de visualização adotado foi criado de forma a atender a demanda de detectar fraudes em um conjunto de dados de NFCes, porém, esse conjunto de dados não foi rotulado como os trabalhos revisados, nem préviamente explorado, o que torna este trabalho pioneiro. Devido a ausência de um subconjunto rotulado, torna-se impossibilita a validação dos resultados, além de diminuir o número de técnicas de análise de dados aplicáveis ao processo. Dessa forma, o processo de visualização aplicado se foca na descoberta de conhecimento, visando dar auxílio a tarefa de detecção de fraudes.

Capítulo 4

Metodologia

Neste capítulo será apresentada a metodologia adotada nesta pesquisa para a elaboração de um processo de visualização exploratória de dados a fim de auxiliar a detecção de fraudes em NFCes. Devido às tarefas de detecção de fraudes e de anomalias serem semelhantes, a presente pesquisa estudou e empregou técnicas de visualização que apropriadas nesse contexto. Adicionalmente, outro conjunto de dados relacionados com fraudes em cartão de crédito foi considerado para propósitos de análise a validação do processo de visualização proposto.

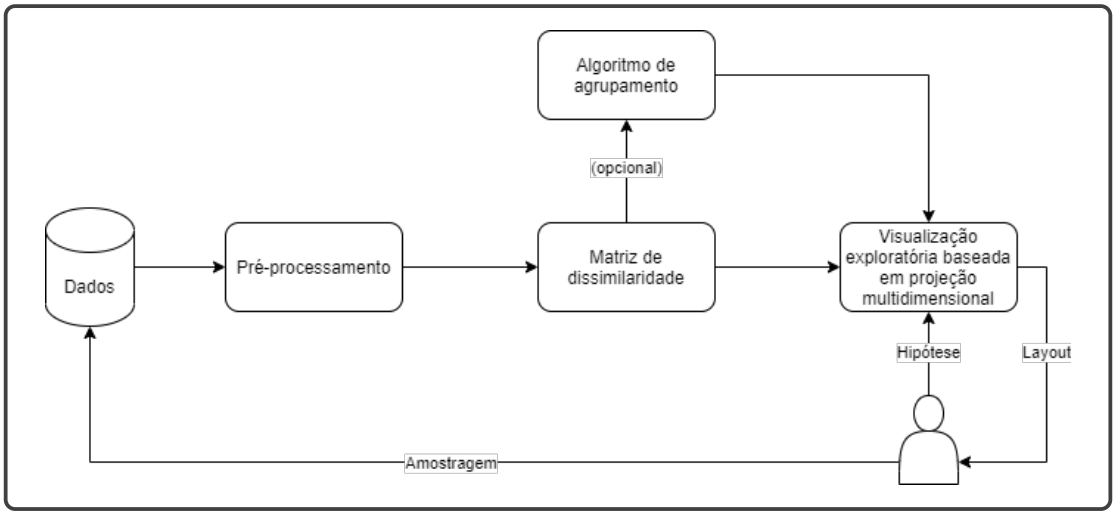


Figura 4.1: Fluxograma detalhando o processo de visualização adotado.

As etapas do processo de visualização proposto podem ser vistas na Figura 4.1. No início, uma amostra de dados é selecionada pelo especialista. Em seguida, os dados são carregados e pré-processados, de forma que as instâncias com atributos ausentes são retiradas do subconjunto de dados. Em seguida, uma matriz de dissimilaridade Gower é criada a partir dos dados já tratados para codificar dados de tipos diferentes em uma estrutura de dados contendo a informação de distância entre pares de pontos, realçando

padrões e relações de similaridade entre dados. Na próxima etapa, um algoritmo de agrupamento pode ser aplicado a matriz de dissimilaridade de forma a encontrar padrões de forma não supervisionada, pois o conjunto de dados em questão não contém rótulos que permitam o uso de algoritmos supervisionados. Finalmente, a visualização exploratória é realizada sob os dados seguindo o mantra de Shneiderman, permitindo a interação do analista com os *layouts* obtidos das técnicas de visualização [16].

4.1 Conjunto de Dados

A Tabela 4.1 apresenta os dois conjuntos de dados que foram considerados no presente trabalho para validar o processo de visualização exploratória proposto. O primeiro conjunto é apresentado na Tabela 4.2 e contém dados de NFCes do Distrito Federal geradas no ano de 2018. Por questões de confidencialidade, as identificações dos indivíduos envolvidos nas transações foram criptografadas. Já o segundo conjunto, mostrado na Tabela 4.3 consiste de dados rotulados sobre portadores de cartões de crédito. Ambos os conjuntos de dados são multidimensionais e contém atributos de diferentes tipos. Uma breve descrição de cada atributo contido no conjunto de dados de NFCes é dado a seguir:

- **Descrição**: A descrição do produto. Não segue uma estrutura fixa, de modo que o vendedor tem a capacidade de descrever o produto da forma como desejar.
- **CFOP**: Código Fiscal de Operações e Prestações. Delimita detalhes sobre a operação realizada. CFOP pode ser utilizado para indicar operações de exportação, devolução, entrega de domicilio, etc.
- **NCM**: Nomenclatura Comum do Mercosul. É uma nomenclatura regional para a categorização de mercadorias adotada pelo Brasil, Argentina, Paraguai e Uruguai desde 1995. Delimita a class de produtos à qual o produto pertence.
- **CST**: Código de Situação Tributária. Determina de qual forma o produto será tributado. Por exemplo, o produto pode ter uma tributação integral, parcial ou ser isento.
- **Unidade**: Unidade de medida do produto.
- **Quantidade**: Quantidade vendida do produto.
- **Demi**: Data referente à venda.
- **Valor Unitário**: Valor de uma unidade do produto.

Conjunto de Dados	Instâncias	Atributos
Notas Fiscais do Consumidor Eletrônicas	1.597.830	9
Statlog (German Credit Card)	1.000	20

Tabela 4.1: Tabela descrevendo os conjuntos de dados utilizados nos experimentos.

Atributo	Tipo
CFOP	Qualitativo
NCM	Qualitativo
CST	Qualitativo
DESCRIÇÃO	Cadeira de Caracteres
UNIDADE	Qualitativo
QUANTIDADE	Numérico
DEMI	Data (dd/mm/yy)
VL_UNITARIO	Numérico

Tabela 4.2: Atributos do conjunto de dados de NFCes.

Atributo	Tipo	Atributo	Tipo
over_draft	Qualitativo	credit_usage	Numérico
credit_history	Qualitativo	purpose	Qualitativo
current_balance	Numérico	average_credit_balance	Qualitativo
employment	Qualitativo	location	Qualitativo
location	Numérico	personal_status	Qualitativo
other_parties	Qualitativo	residence_since	Numérico
property_magnitude	Qualitativo	cc_age	Numérico
other_payment_plans	Qualitativo	housing	Qualitativo
existing_credits	Numérico	job	Qualitativo
num_dependents	Numérico	own_telephone	Qualitativo
foreign_worker	Qualitativo	class	Qualitativo

Tabela 4.3: Atributos do conjunto de dados de Statlog.

4.2 Pré-processamento dos Dados

A etapa de pré-processamento se inicia com o carregamento do conjunto de dados de interesse, em que no caso do conjunto das NFCes, deve-se obter uma amostra de NFCes relacionadas a produtos que possivelmente contenham indícios de fraude. Como existem limitações no que se refere à complexidade espacial do cálculo da matriz de uma dissimilaridades, o tamanho da amostra escolhido para a maioria dos experimentos foi de 1.000 instâncias, porém foi possível realizar experimentos em amostras de até 10.000 instâncias no computador pessoal utilizado.

Para o pré-processamento das NFCes, as instâncias com dados do tipo Código de Situação Tributária (CST) ausentes foram removidas da análise, pois constituem dados de

prestação de serviço que não são o foco do presente trabalho e aumentariam a complexidade da análise. Além disso, os dados temporais não são levados em conta devido ao fato de que as amostras eram aleatórias e considerou-se que as datas da compra não são relevantes para a detecção de fraudes neste caso específico.

Por sua vez, como o conjunto de dados Statlog é rotulado, foi utilizado o critério de ganho de informação e entropia para determinar os atributos mais relevantes em relação ao atributo de categoria (*class*). Os atributos que apresentam um ganho de informação nulo são retirados da análise. Em uma situação real, com a ausência de dados rotulados, seria possível realizar uma tarefa semelhante por dedução e geração de hipóteses.

Após o pré-processamento dos dados, é criada uma matriz de distância de Gower a partir dos conjuntos de dados, que representa informações de dados categóricos e numéricos em uma estrutura de dados de tamanho $n \times n$, em que cada instância representa a dissimilaridade entre pares de dados. Essa matriz é utilizada como entrada para um algoritmo de agrupamento PAM, a fim de revelar informações sobre padrões implícitos nos dados, embora o uso dessas informações para realçar as visualizações seja opcional e dependa da vontade do analista. Para a escolha do número de grupos do algoritmo de agrupamento, foi utilizada a noção de curva de silhueta, em que geralmente o maior valor de silhueta indicaria uma maior qualidade de agrupamento devido à maior coesão entre os pontos em um mesmo grupo. Essa métrica foi utilizada, ao contrário de outras como o *Elbow Method*, pois os gráficos de silhueta proporcionam análises menos ambivalentes em relação ao número de agrupamentos [55].

Em seguida, a matriz de distâncias é utilizada como entrada para uma técnica de visualização baseada em projeção multidimensional, de forma a transformá-la em uma estrutura de dados de duas ou três dimensões para sua posterior visualização. Os métodos escolhidos para esses experimentos foram o *Multidimensional Scaling* (MDS) métrico, o *Isometric Mapping* (ISOMAP) e o *t-Distributed Stochastic Neighborhood Embedding* (t-SNE). As técnicas de projeção multidimensional foram escolhidas devido a sua capacidade de receber matrizes de dissimilaridade como entrada, visto que outros algoritmos da mesma classe, como o *Principal Component Analysis* [42], não operam sobre matrizes de dissimilaridade. Ademais, o t-SNE é uma técnica que se tornou popular em diversas pesquisas [4] relacionadas à visualização e ciência de dados. Por isso, uma maior ênfase foi dada ao t-SNE e ao seu ajuste fino de parâmetros, de forma a variar os valores do número de iterações, como também os valores de perplexidade e a semente [56].

Primeiramente, é relevante ressaltar a importância do número de iterações, pois o algoritmo revela *layouts* distintos antes e depois de alcançar um estado estável. A perplexidade é um parâmetro que consiste de uma medida de informação, que é definida como $Perp(P_i) = 2^{H(P_i)}$, em que $H(P_i)$ é o valor da entropia de Shannon em bits. Para o

algoritmo t-SNE, esse parâmetro pode ser visto como o número efetivo de vizinhos mais próximos, de forma semelhante a vários outros algoritmos de aprendizagem de máquina. Finalmente, como o t-SNE é um algoritmo estocástico e realiza uma etapa de otimização com um gradiente descendente iniciado com um valor aleatório, é possível que cada execução do algoritmo gere um *layout* um pouco diferente. Dessa forma, é recomendável que se execute o algoritmo mais de uma vez e selecione o *layout* com menor valor de divergência K-L.

4.3 O Processo de Visualização Exploratória

A visualização exploratória de dados consiste na integração do elemento humano ao processo de visualização, promovendo sua flexibilidade, criatividade e conhecimento [16]. Deste modo, o especialista em questão pode aplicar sua percepção aos dados de maneira visual e obter compreensão, levantar hipóteses, tirar conclusões e interagir com os dados de maneira intuitiva.

A validação das hipóteses também pode ser realizada por meio do processo de exploração de dados, utilizando técnicas de estatística e aprendizado de máquina. No entanto, a visualização exploratória apresenta vantagens sobre essas alternativas, pois é intuitiva e não requer uma compreensão de parâmetros matemáticos complexos.

Como o processo proposto tem como propósito principal a descoberta de conhecimento em conjuntos de NFCes, o auditor ou especialista deve elaborar uma hipótese de pesquisa ou definir objetivos específicos e seguir as etapas propostas por Daniel E. Keim, que correspondem ao Mantra de Shneiderman [57] que são detalhadas adiante: *Visão global inicial*, *ampliação e filtragem* e *detalhes-sob-demanda*.

4.3.1 Visão Global Inicial

Ao iniciar o processo, o usuário obtém uma visão global dos dados ao analisar o *layout* gerado pela técnica de visualização baseada em projeção multidimensional. A Figura 4.2 ilustra essa etapa ao mostrar o *layout* obtido ao utilizar a técnica t-SNE para visualizar os dados do conjunto Statlog. Dessa forma, é possível identificar padrões relevantes conforme o posicionamento dos pontos no referido *layout*, podendo decidir se concentrar em grupos de pontos que estejam relacionados aos seus objetivos.

O especialista tem acesso às técnicas de interação com os *layouts*, possibilitando a descoberta de conhecimento durante o processo exploratório. Nesse sentido, pode-se atribuir cores aos pontos (símbolos) no espaço visual conforme os atributos categóricos dos dados ou alterar as representações simbólicas das instâncias, de modo a obter uma separação visual dos dados. A Figura 4.3 ilustra essa separação, que foi atribuída pelo algoritmo de

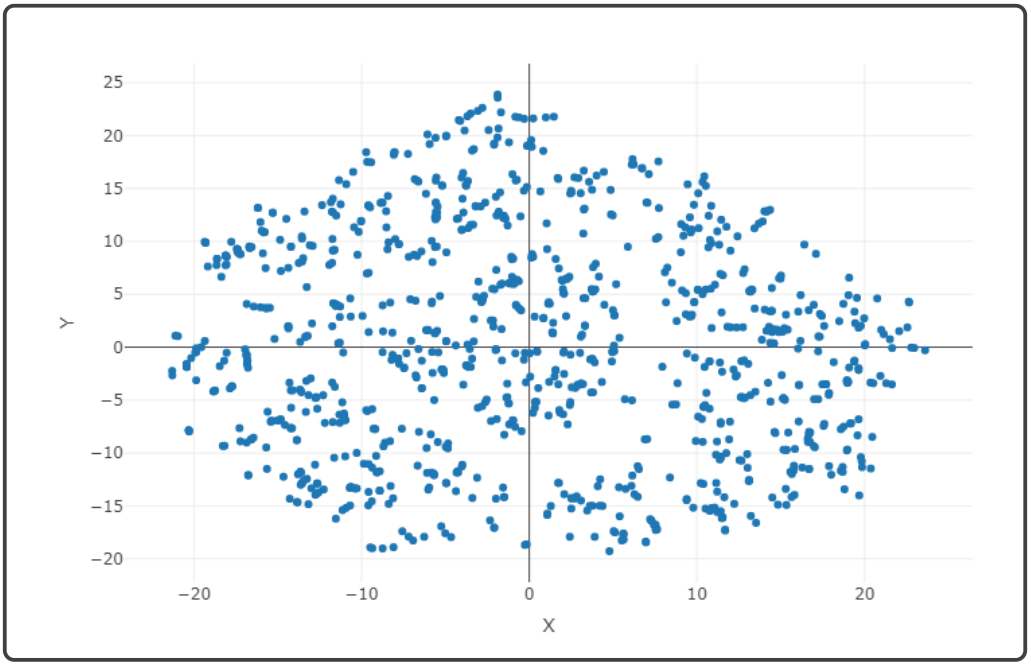


Figura 4.2: Exemplo de representação gráfica do conjunto de dados Statlog, gerada via t-SNE.

agrupamento, separando as instâncias em 7 grupos distintos. Já a Figura 4.4 apresenta a coloração dos pontos no *layout* de acordo os valores do atributo *over_draft*.

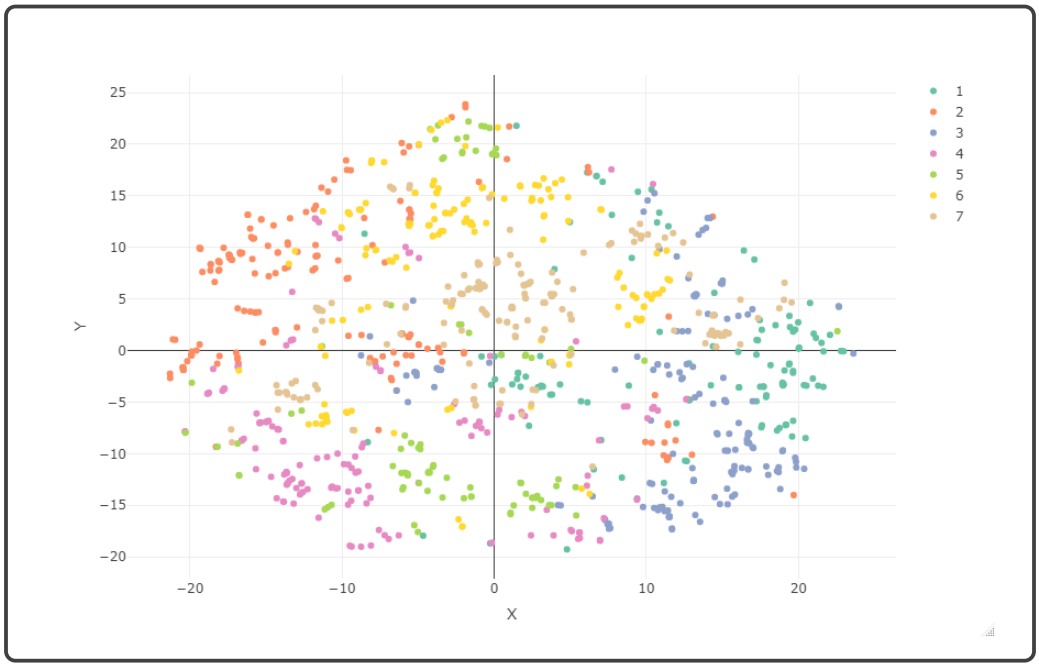


Figura 4.3: Exemplo de representação gráfica do conjunto de dados Statlog com coloração dada pelo algoritmo de agrupamento PAM com 7 grupos distintos.

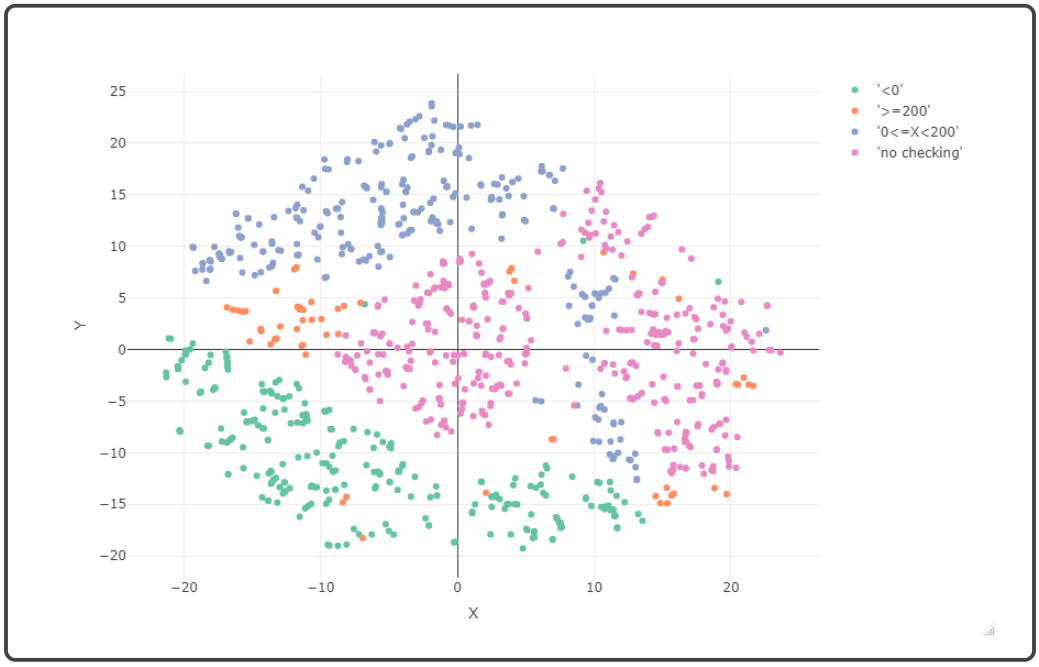


Figura 4.4: Exemplo de representação gráfica do conjunto de dados Statlog com coloração dada pelo atributo *over_draft*.

A partir do *layout* gerado utilizando a visualização t-SNE, é importante que o especialista tenha em mente algumas particularidades do algoritmo de forma a interpretar efetivamente seus resultados [56]:

- O t-SNE reproduz as relações de similaridade entre os dados visualmente;
- O tamanho relativo de agrupamentos na representação gráfica não possui significado;
- A distância entre agrupamentos bem separados pode não ter significado.

4.3.2 Ampliação e Filtro

As técnicas e recursos de interação do usuário com o *layout* podem ser empregados para auxiliar o especialista na análise de áreas ou grupos de pontos específicos do espaço visual original. A Figura 4.5 demonstra o uso da ferramenta de seleção em uma área do *layout*, enquanto que a Figura 4.6 apresenta o *layout* ampliado obtido dessa área, possibilitando que uma análise concentrada no grupo de pontos associado. Além disso, o especialista também tem a opção de filtragem por atributos para simplificar o *layout*, conforme mostra a Figura 4.7.

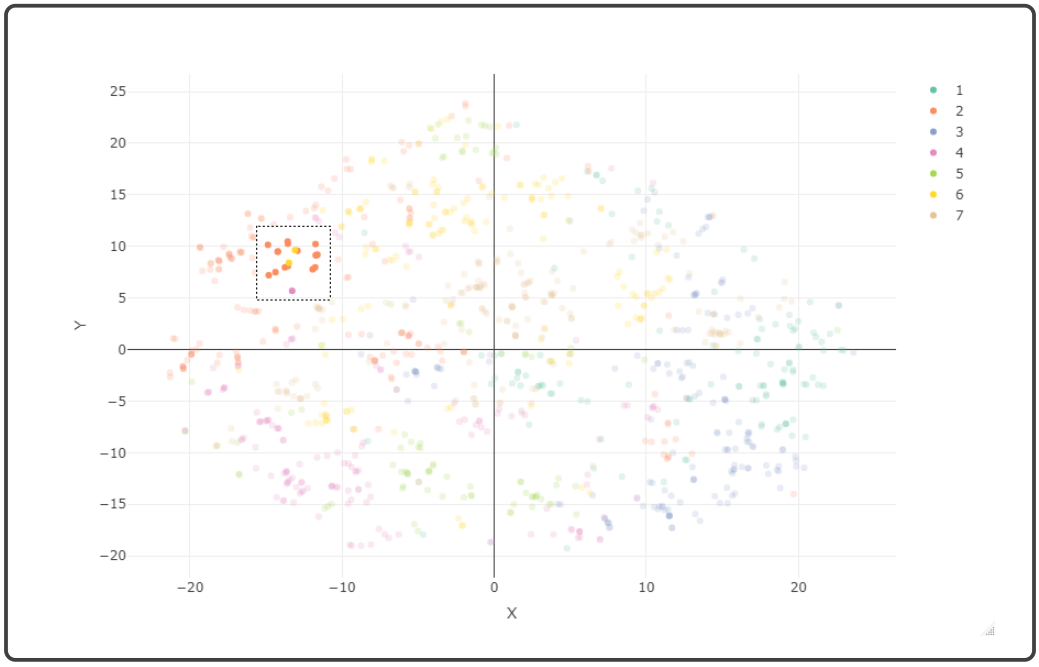


Figura 4.5: Exemplo de utilização da ferramenta de seleção no conjunto de dados Statlog.

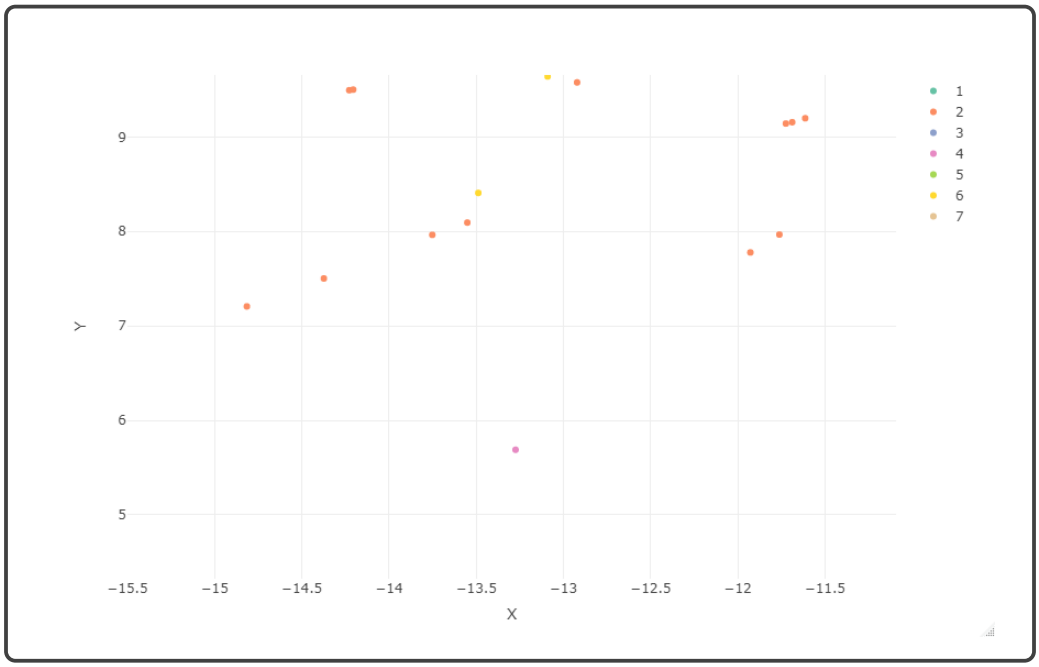


Figura 4.6: Exemplo de utilização da ferramenta de ampliação no conjunto de dados Statlog.

4.3.3 Detalhes-sob-demanda

Nessa etapa, a visualização exploratória prossegue de forma interativa, permitindo a obtenção de detalhe-sob-demanda, ilustrado na Figura 4.8. Nesse caso, o especialista pode

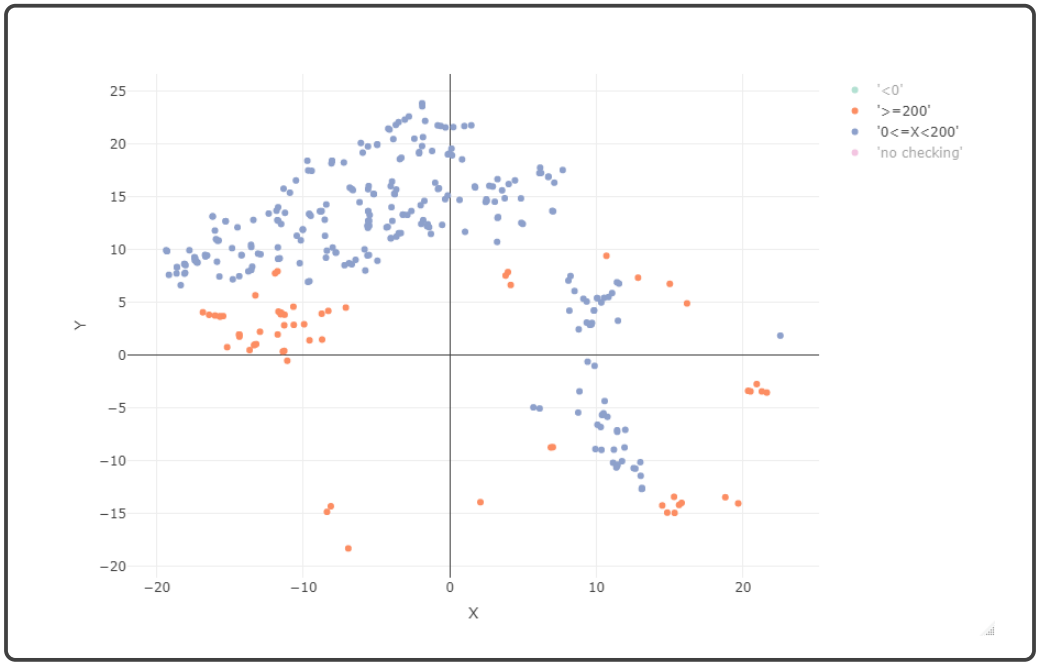


Figura 4.7: Exemplo de utilização da ferramenta de filtragem no conjunto de dados Statlog.

selecionar pontos no *layout* e obter informações estatísticas ou originais para auxiliar na tarefa de interpretação, como a descrição dos produtos envolvidos ou seus valores de compra nas respectivas transações.

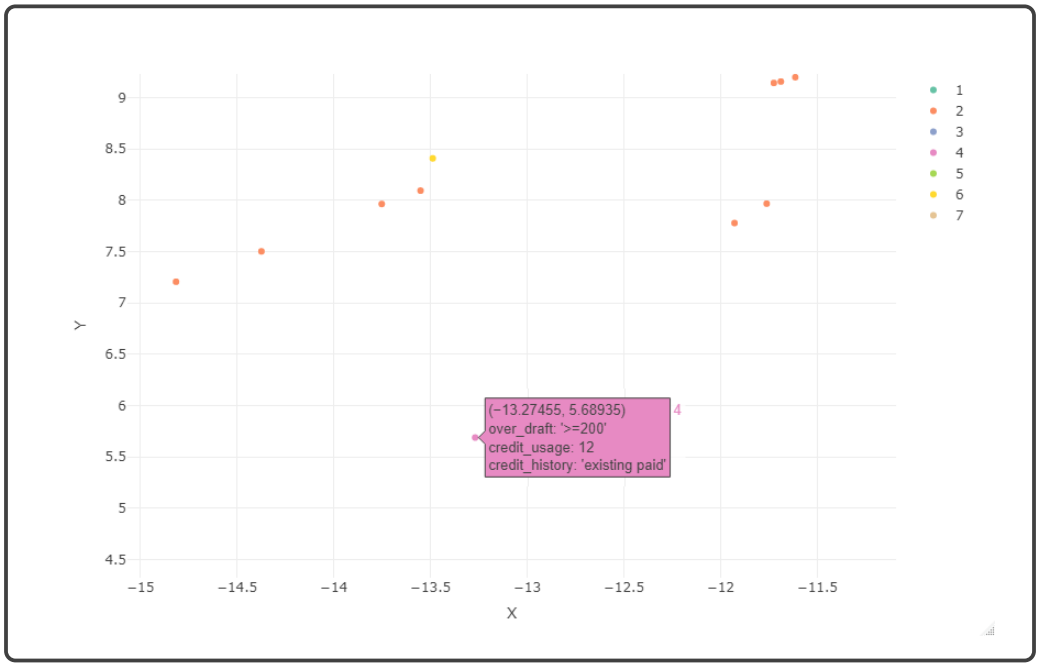


Figura 4.8: Exemplo de detalhes-sob-demanda no conjunto de dados Statlog.

Resumidamente, a partir do processo de visualização proposto, o especialista a cargo da detecção de fraudes obtém uma representação gráfica dos dados, em que o posicionamento dos pontos no espaço visual se baseia e reflete as relações de similaridade dos dados originais. Dessa forma, a descoberta de conhecimento é resultado da visualização exploratória, que envolve a interpretação e interação do usuário com os *layouts* gerados. Assim, o especialista pode observar indícios de fraude ao analisar os grupos formados nos gráficos, uma vez que as notas fiscais similares tendem a se posicionar próximas umas das outras, associado ao uso de recursos interativos, como coloração dos pontos conforme atributos específicos (CST ou NCM), ou com detalhamento da descrição das compras associadas.

4.4 Considerações Finais

Neste capítulo foi apresentada a metodologia adotada no presente trabalho, que emprega o uso de vários algoritmos de áreas diversas da ciência da computação e estatística para criar um processo de visualização exploratória. No processo proposto, a ênfase é colocada na interação entre o especialista e a representação gráfica gerada, de modo que a facilitar a exploração visual dos dados e o descobrimento de conhecimento.

A partir de uma hipótese de pesquisa, o especialista à frente do processo tem a capacidade de gerar representações gráficas dos dados. Ao analisar essas representações, ele obtém uma compreensão intuitiva acerca da estrutura e das relações implícitas ao conjunto de dados. Essa compreensão pode ser utilizada para criar novas hipóteses e a partir delas gerar novas representações ou pode ser utilizada para tirar conclusões que apoiem a tomada de decisões. Ademais, várias técnicas de visualização podem ser utilizadas a fim de possibilitar a análise por diferentes perspectivas, permitindo ao especialista o exercício de sua criatividade e intuição na processo de descobrimento de conhecimento.

Capítulo 5

Resultados Experimentais

Neste capítulo serão apresentados os resultados experimentais do presente trabalho, descrevendo detalhes de implementação e informações relevantes para a reprodução da pesquisa.

5.1 Ambiente de Testes

Um computador pessoal foi utilizado para a execução dos experimentos descritos nesta monografia e a linguagem de programação R, em conjunto com bibliotecas da linguagem presentes no *Comprehensive R Archive Network* (CRAN), foi escolhida. As especificações do computador e do software utilizados são dadas na Tabela 5.1. As bibliotecas da linguagem R aplicadas nos experimentos são apresentadas na Tabela 5.2.

Hardware	Especificações
CPU	Intel(R) Core(TM) i5-6400 CPU @ 2.70Ghz
GPU	NVIDIA GeForce GTX 970
Memória	8 GB DDR3 @ 1600 Mhz
Disco	HDD 1 TB, 3.5' SATA
Sistema Operacional	Windows 10 Pro
Linguagem de Programação	R version 4.0.3 (2020-10-10)
IDE	RStudio Desktop

Tabela 5.1: Especificações do computador pessoal utilizado nos experimentos.

5.2 Ajuste de Parâmetros

Nesta seção, os experimentos foram realizados visando obter os parâmetros apropriados para a geração de *layouts* que facilitem a interpretação dos dados pelos especialistas.

Biblioteca	Versão	Fonte
dplyr	1.0.2	[58]
cluster	2.1.0	[59]
Rtsne	0.15	[51]
plotly	4.9.2.1	[60]
vegan	2.5.6	[61]
stats	4.0.3	[2]

Tabela 5.2: Especificações das bibliotecas de R que foram utilizadas nos experimentos.

Inicialmente, buscou-se variar os parâmetros relacionados ao número de iterações, perplexidade e valores de semente (*seeds*).

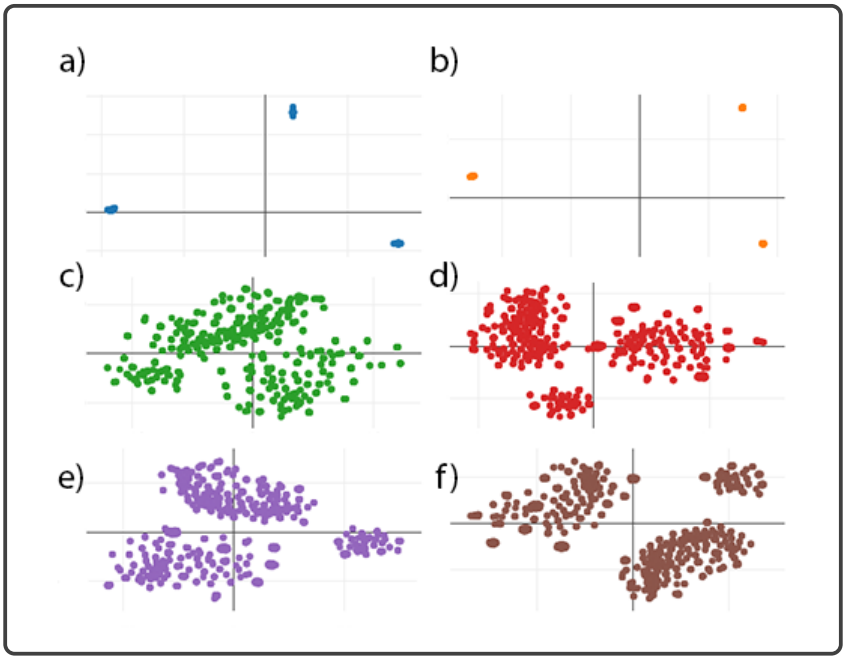


Figura 5.1: *Layouts* gerados pela técnica t-SNE utilizando o conjunto de dados de NFCes variando-se a quantidade de iterações: (a) 100 iterações; (b) 200 iterações; (c) 300 iterações; (d) 400 iterações; (e) 500 iterações; (f) 1000 iterações.

As Figuras 5.1 e 5.2 mostram os *layouts* gerados variando-se o número de iterações. Ao executar o t-SNE com 100 e 200 iterações para o conjunto de dados de NFCes, como mostram as Figuras 5.1 (a-b), pode-se verificar uma grande sobreposição de pontos que prejudica a análise de estruturas locais nos dados. Para os conjuntos de dados Statlog, percebe-se nas Figuras 5.2 (a-c) que o posicionamento dos pontos não se altera para 100, 200 e 300 iterações. Nas Figuras 5.1 (c-f) e 5.2 (d-f), percebe-se a formação de grupos de pontos, que não se altera significativamente com o passar das iterações.

Ao executar o t-SNE variando o valor da semente, diferentes valores de divergência K-L foram obtidos. No entanto, os *layouts* gerados não apresentaram mudanças significativas

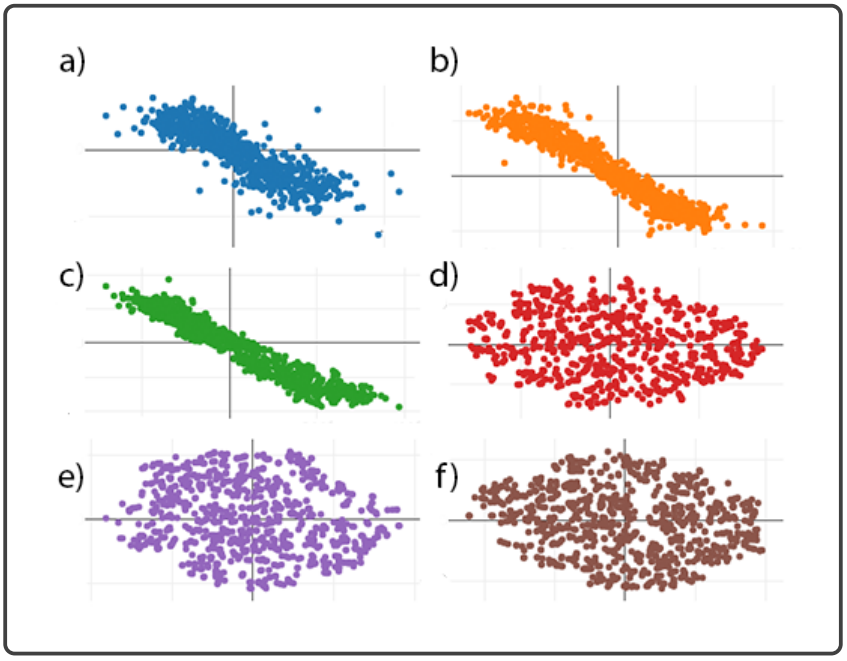


Figura 5.2: *Layouts* gerados pela técnica t-SNE utilizando o conjunto de dados Statlog variando-se a quantidade de iterações: (a) 100 iterações; (b) 200 iterações; (c) 300 iterações; (d) 400 iterações; (e) 500 iterações; (f) 1000 iterações.

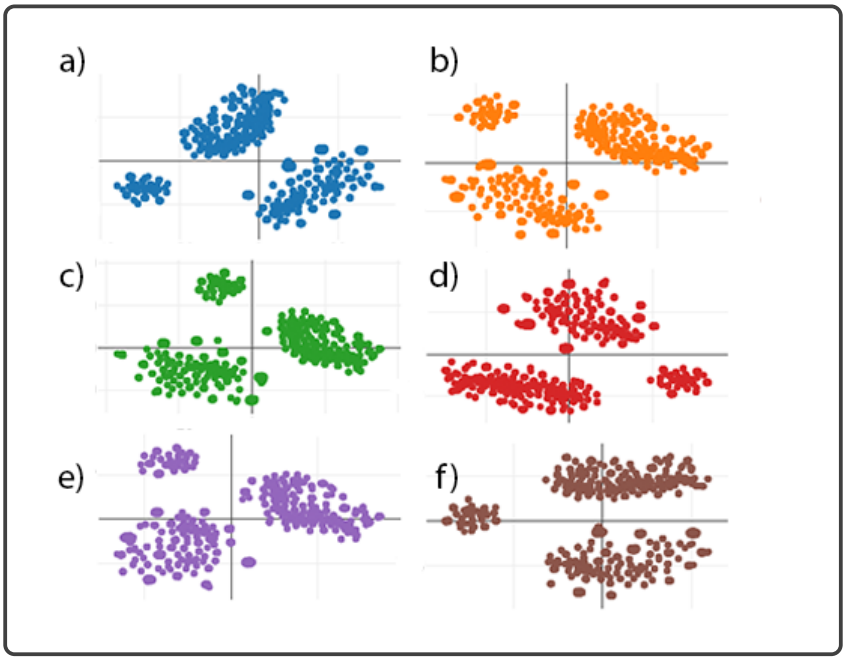


Figura 5.3: *Layouts* gerados pela técnica t-SNE utilizando o conjunto de dados de NFCes variando-se o valor da semente para se obter diferentes valores de divergência K-L: (a) 0.655; (b) 0.620; (c) 0.626; (d) 0.638; (e) 0.635; (f) 0.626.

no posicionamento dos pontos, como demonstrado nas Figuras 5.3 e 5.4.

Variando-se o parâmetro de perplexidade, observa-se que a estrutura global dos dados

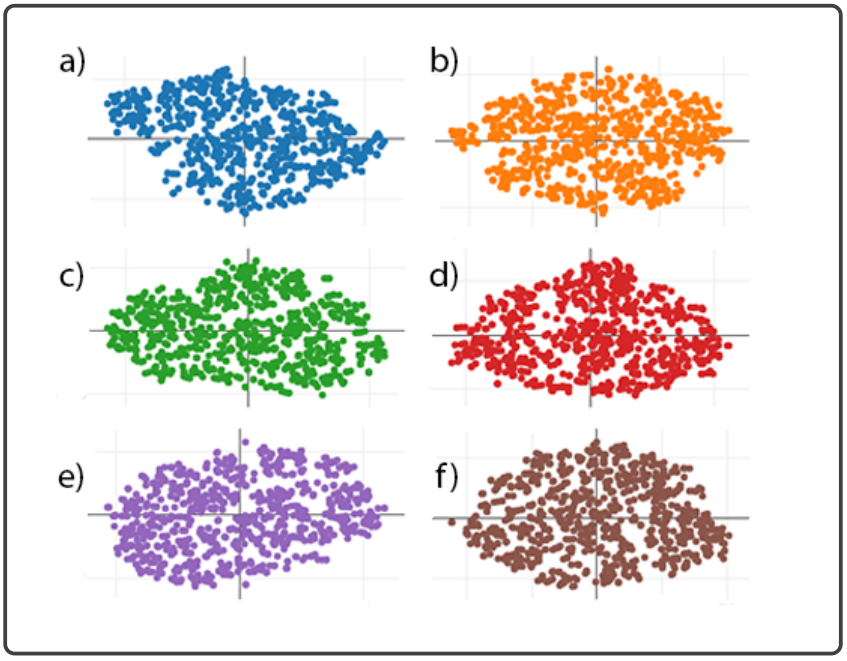


Figura 5.4: *Layouts* gerados pela técnica t-SNE utilizando o conjunto de dados Statlog variando-se o valor da semente para se obter diferentes valores de divergência K-L: (a) 1.361; (b) 1.430; (c) 1.398; (d) 1.386; (e) 1.433; (f) 1.387.

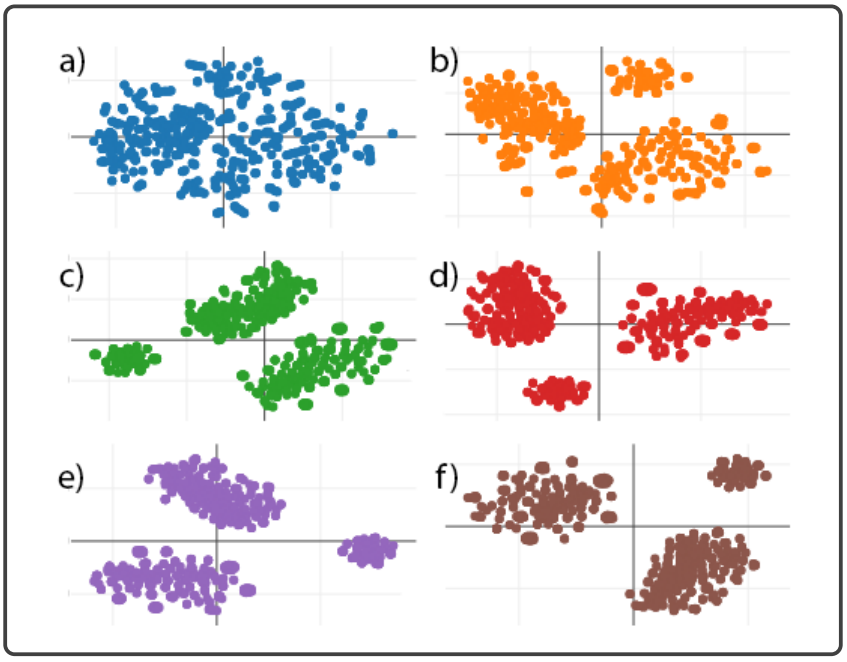


Figura 5.5: *Layouts* gerados pela técnica t-SNE utilizando o conjunto de dados de NFCes variando-se o parâmetro perplexidade: (a) 05 perplexidade; (b) 15 perplexidade; (c) 25 perplexidade; (d) 35 perplexidade; (e) 45 perplexidade; (f) 50 perplexidade.

não está completa até que o parâmetro tenha alcançado o valor 15, como na Figura 5.5 (b). Na Figura 5.5 (a), não há uma distinção entre os três maiores agrupamentos de

dados, porém já na Figura 5.5 (b), é possível observar a formação dos agrupamentos. Nos experimentos realizados com o conjunto de dados Statlog, não se observa mudanças significativas com a alteração do parâmetro, como ilustrado na Figura 5.6.

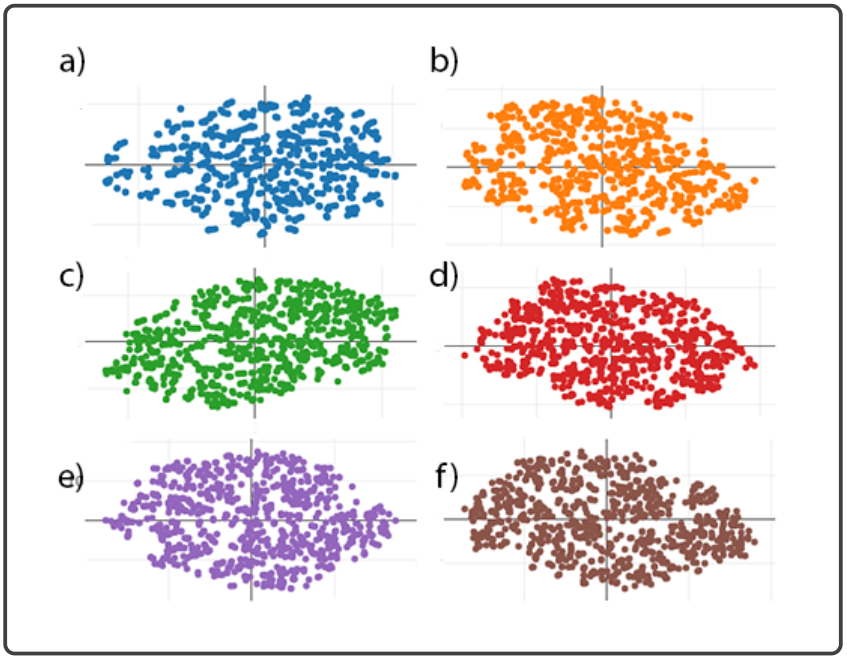


Figura 5.6: *Layouts* gerados pela técnica t-SNE utilizando o conjunto de dados Statlog variando-se o parâmetro perplexidade: (a) 05 perplexidade; (b) 15 perplexidade; (c) 25 perplexidade; (d) 35 perplexidade; (e) 45 perplexidade; (f) 50 perplexidade.

Os *layouts* obtidos nesses experimentos indicaram que os valores de parâmetros apropriados para a técnica de visualização t-SNE foram 30 para perplexidade e 1.000 para número de iterações, pois a partir desses valores é seguro afirmar, baseado nos *layouts* gerados, que o posicionamento dos dados não sofrerá mudanças significativas com um acréscimo desses parâmetros. Como o valor da divergência K-L não afetou significativamente os *layouts* obtidos, o valor da semente foi mantido nos experimentos realizados em seguida.

Adicionalmente ao algoritmo t-SNE, foram utilizados o MDS e o ISOMAP como técnicas alternativas de projeção multidimensional para gerar outros *layouts*. As Figuras 5.7 (a-c) apresentam os *layouts* obtidos pelas técnicas MDS, ISOMAP e t-SNE para o conjunto de dados das NFCes, em que as cores dos pontos estão associadas aos rótulos obtidos pelo algoritmo de agrupamento PAM, enquanto que as Figuras 5.8 (a-c) atribuem as cores aos pontos de acordo com os atributos categóricos “CFOP”, “CST” e “NCM”, respectivamente.

É possível notar que os todas as técnicas revelam a estrutura global dos dados, porém os *layouts* gerados pelo t-SNE além de revelar a estrutura global dos dados, também

revelam a estrutura local. No conjunto de dados de NFCes, a estrutura local revelada pelo t-SNE é evidenciada pela presença de produtos de mesmo NCM, como por exemplo bananas de tipos distintos, se aglomerando em um mesmo grupo coeso.

Por sua vez, as Figuras 5.9 e 5.10 ilustram os *layouts* considerando o conjunto de dados Statlog. Na Figura5.9 (a-c) os *layouts* obtidos pelas técnicas de visualização MDS, t-SNE, ISOMAP, respectivamente apresentam as cores do pontos atribuídas conforme o atributo "over draft". Na Figura 5.10, esses mesmos *layouts* tiveram seus pontos coloridos conforme os rótulos obtidos pelo algoritmo PAM.

A utilização da biblioteca *plotly* [60] possibilitou a geração de *layouts* interativos em duas e três dimensões, como é possível observar na Figura 5.11. A biblioteca permitiu a ampliação de imagens; filtragem de dados com base em características pré-selecionadas; e a apresentação de detalhes-sob-demanda, como está demonstrado na Figura 5.12. Assim sendo, foi possível observar empiricamente que os *layouts*, principalmente aquele gerado pela técnica t-SNE, preservam as relações entre os dados, pois dados similares foram agrupados com alta coesão intragrupo.

Constatou-se que as visualizações geradas pelo t-SNE foram melhores em revelar as estruturas implícitas dos dados em relação às visualizações baseadas no ISOMAP e MDS. Os algoritmos de agrupamentos foram úteis em conjunção com a visualização no contexto do processo, podendo enriquecer a análise dos padrões dos dados.

5.3 Validação da Hipótese de Pesquisa

Com o objetivo de validar a hipótese de pesquisa apresentada, uma simulação do ponto de vista do especialista foi executada utilizando o processo de visualização exploratório proposto considerando uma amostra de 10.000 instâncias do conjunto de dados de NFCes. Nesta simulação, foi decidido que a coloração dos dados representariam tanto os agrupamentos de dados, gerados pelo algoritmo de agrupamento PAM, visto na Figura 5.13, quanto o código de situação tributária (CST), visto na Figura 5.14. O número de agrupamentos escolhido para o processo é decidido com a utilização do coeficiente de silhueta , apresentado na Figura 5.15. Por fim, o t-SNE é utilizado para a criação do *layout*, em que os parâmetros utilizados foram perplexidade igual a 30 e 1000 iterações.

Visão Global Inicial

A partir do *layout* gerado, inicia-se o processo de visualização exploratória, observando-se os grupos e as instâncias presentes em cada grupo. De modo geral, a estrutura global dos dados se divide em 3 agrupamentos maiores e diversos agrupamentos menores. De acordo com os maiores agrupamentos, é possível observar que a maior distinção ocorre

devido ao CST dos produtos, porém é também visível que alguns grupos menores se encontram próximos de grupos de CST distinto, como por exemplo, o grupo de dados de coloração azul próximos aos dados de coloração laranja, vistos na Figura 5.16. É importante salientar que o algoritmo de agrupamento difere da coloração fornecida pelo CST em alguns grupos, o que pode ser relevante para a análise.

Ampliação e Filtragem

Com a utilização da técnica de ampliação, observada na Figura 5.17, o especialista é capaz de concentrar sua análise em um subconjunto de dados menor. Assim sendo, não há a necessidade de utilizar a técnica de filtragem.

Detalhes sob Demanda

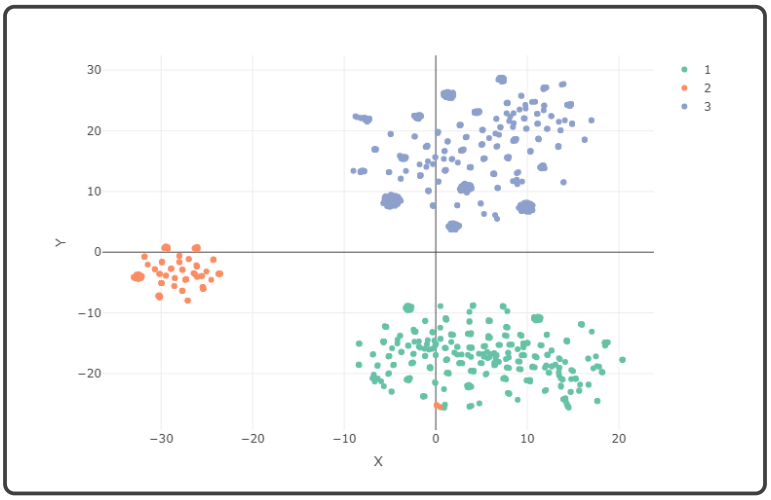
A partir da área de análise ampliada, é possível então se concentrar em instâncias individuais de dados e seus detalhes. Os detalhes são obtidos ao se posicionar o cursor sobre a instância desejada. Na Figura 5.18 observa-se uma instância cuja descrição é definida por "&FJ CAR TIO JORGE", que se refere a um tipo/marca de feijão. Próximo a essa instância, observa-se "FEIJAO CARIOCA CRIST" que também está relacionada a um produto de feijão. No entanto, apesar desses produtos apresentarem os mesmos NCM e CFOP, apresentam CSTs diferentes, o que poderia representar indícios de inconsistência na tributação. A partir dessa informação, o especialista pode então validar sua hipótese ou continuar sua análise.

5.4 Discussão

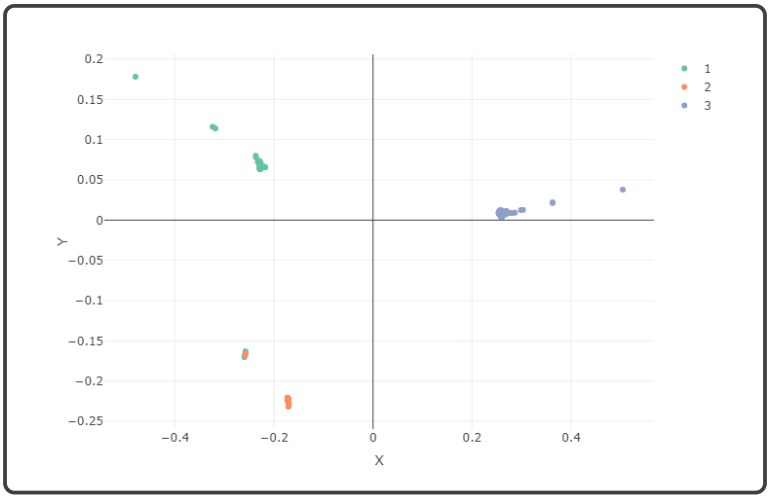
O processo de visualização exploratória proposto apresenta as vantagens de possibilitar a visualização de um conjunto de dados multidimensional e de atributos de diferentes tipos, enfatizando as relações de similaridade entre os dados. A visualização do conjunto de dados nesta representação permite ao especialista obter uma compreensão intuitiva acerca da estrutura e dos padrões contidos nos dados, além de ter a capacidade de interagir utilizando técnicas como seleção, ampliação e filtragem. Dessa forma, o especialista é integrado ao processo de exploração, possibilitando o exercício de sua flexibilidade, criatividade e conhecimento.

No entanto, o processo também apresenta limitações, como o gargalo criado pela complexidade espacial relativa à criação de uma matriz de dissimilaridade, o que limita a escalabilidade das representações. De forma geral, o processo foi proposto como uma etapa em um processo maior de análise de dados, em que outros tratamentos já foram

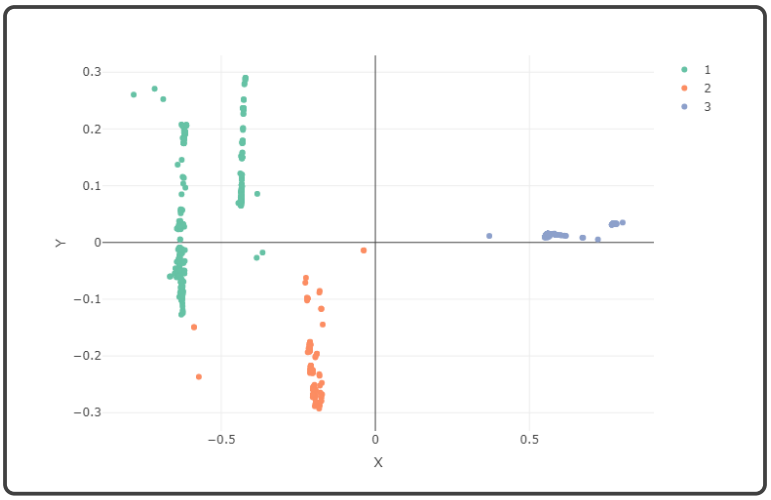
aplicados aos dados e então se torna oportuna a visualização de um subconjunto menor de forma a auxiliar a tarefa de detecção de fraudes.



(a) Amostra de 1.000 instâncias, via t-SNE, com a coloração dada pelos agrupamentos gerados pelo algoritmo PAM.

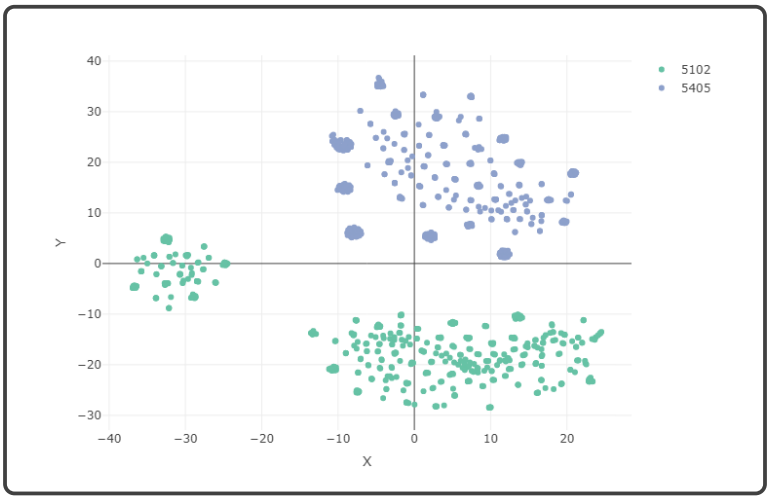


(b) Amostra de 1.000 instâncias, via MDS, com a coloração dada pelos agrupamentos gerados pelo algoritmo PAM.

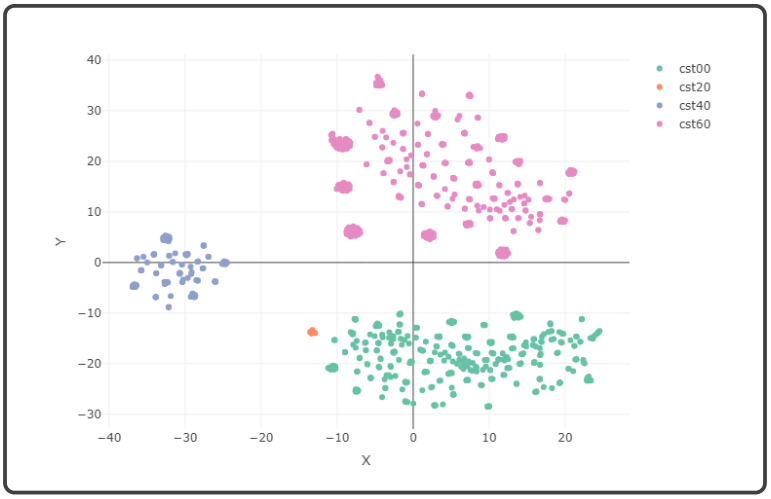


(c) Amostra de 1.000 instâncias, via ISOMAP, com a coloração dada pelos agrupamentos gerados pelo algoritmo PAM.

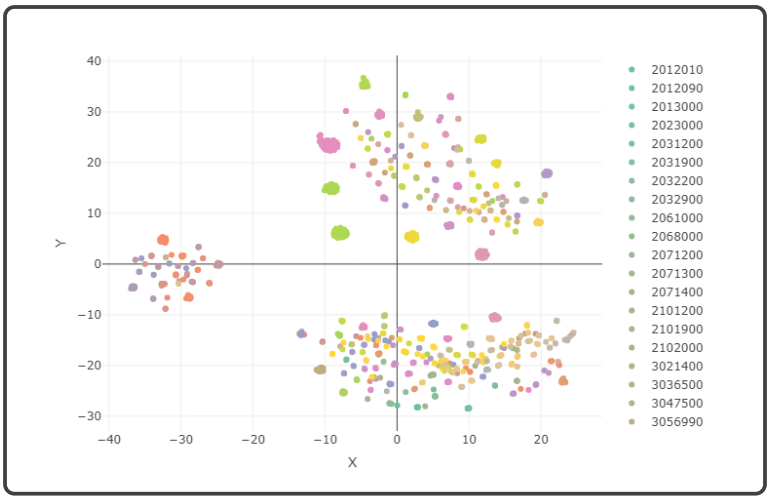
Figura 5.7: *Layouts* do conjunto de dados de NFCes em duas dimensões.



(a) Amostra de 1.000 instâncias, via t-SNE, com coloração atribuída pelo atributo CFOP.

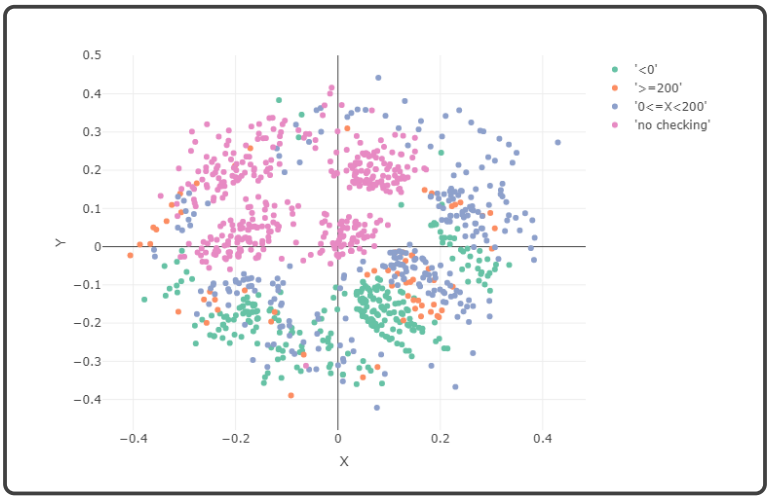


(b) Amostra de 1.000 instâncias, via t-SNE, com coloração atribuída pelo atributo CST.

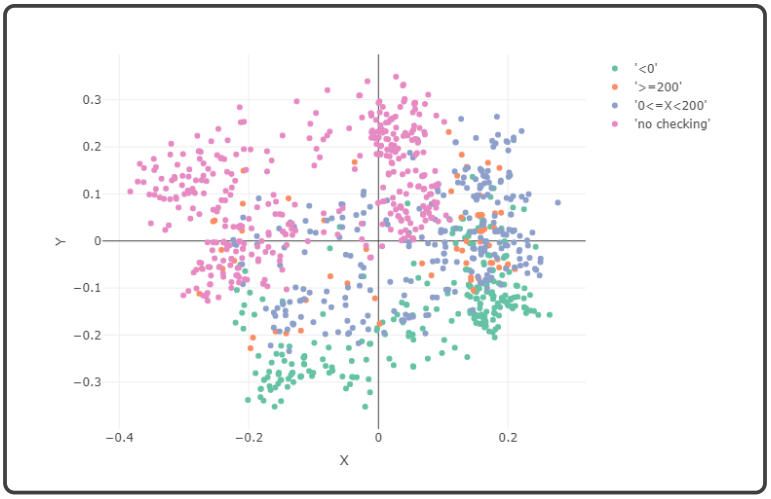


(c) Amostra de 1.000 instâncias, via t-SNE, com coloração atribuída pelo atributo NCM.

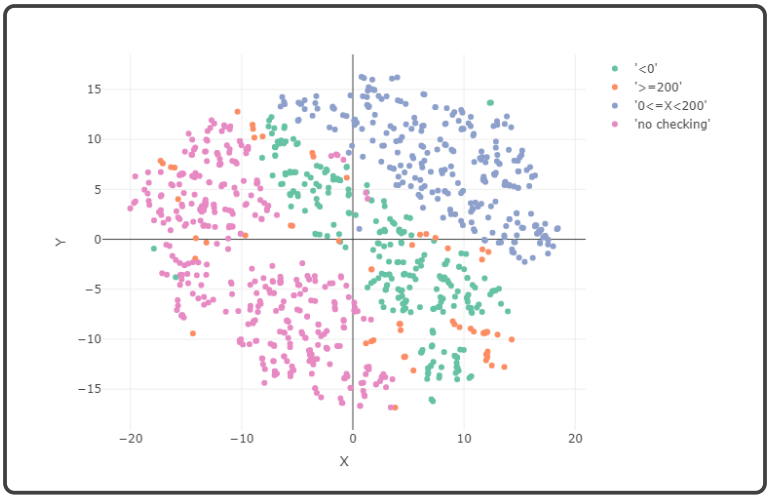
Figura 5.8: *Layouts* do conjunto de dados de NFCes em duas dimensões.



(a) Amostra de 1.000 instâncias, via MDS, com coloração atribuída pelo atributo “over draft”.

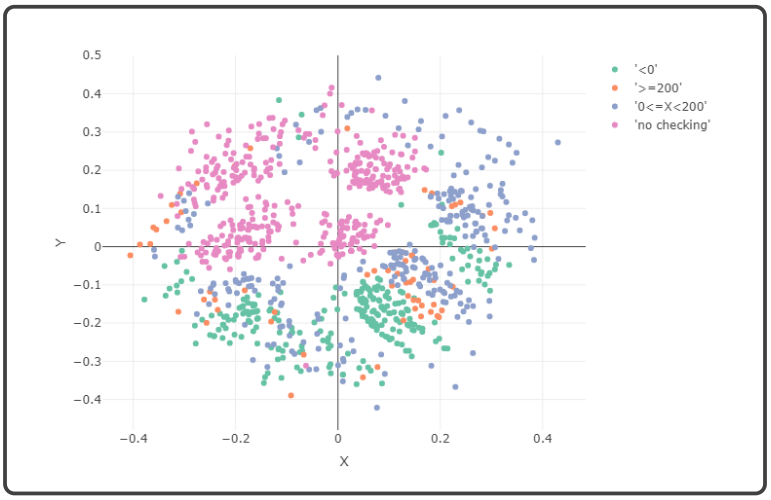


(b) Amostra de 1.000 instâncias, via ISOMAP, com coloração atribuída pelo atributo “over draft”.

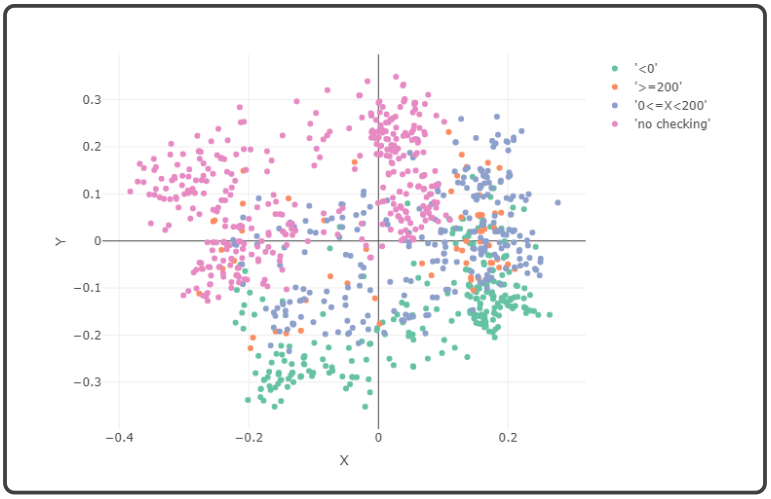


(c) Amostra de 1.000 instâncias, via t-SNE, com coloração atribuída pelo atributo “over draft”.

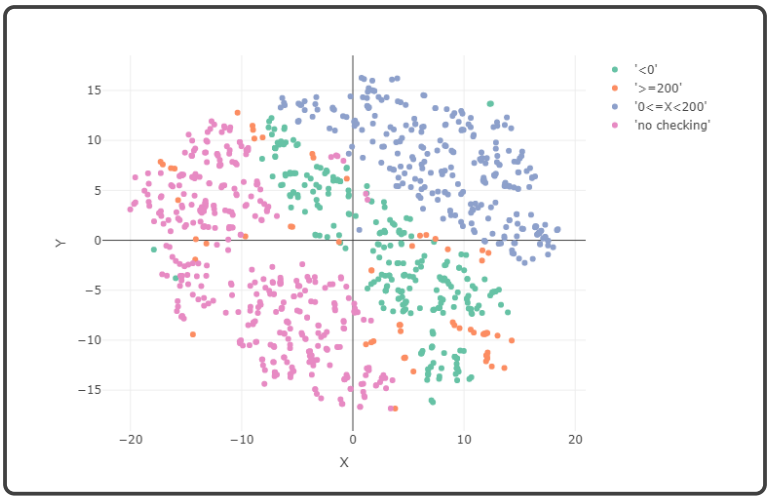
Figura 5.9: *Layouts* do conjunto de dados Statlog em duas dimensões.



(a) Amostra de 1.000 instâncias, via MDS, com coloração atribuída pelos agrupamentos gerados pelo algoritmo PAM.

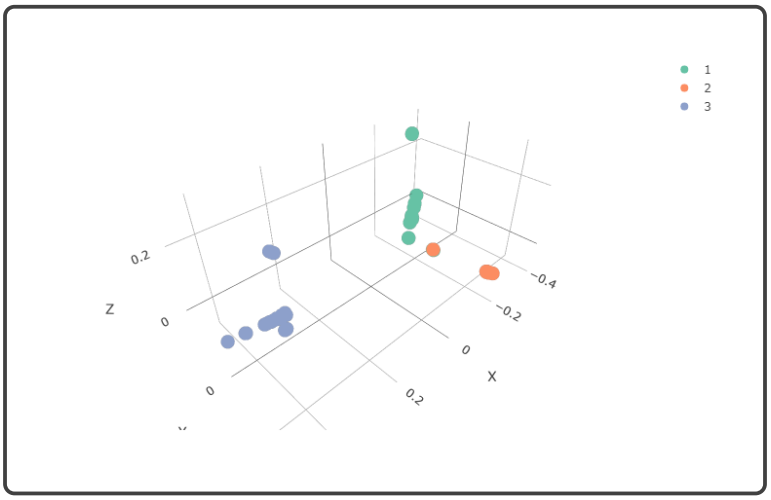


(b) Amostra de 1.000 instâncias, via ISOMAP, com coloração atribuída pelos agrupamentos gerados pelo algoritmo PAM.

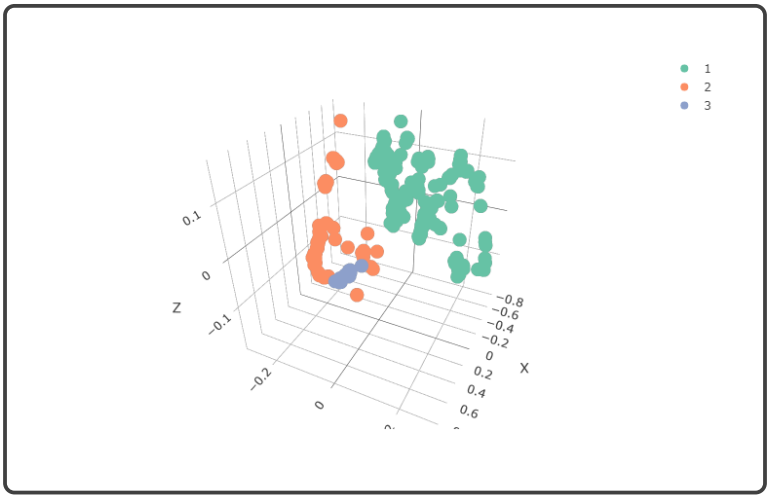


(c) Amostra de 1.000 instâncias, via t-SNE, com coloração atribuída pelos agrupamentos gerados pelo algoritmo PAM.

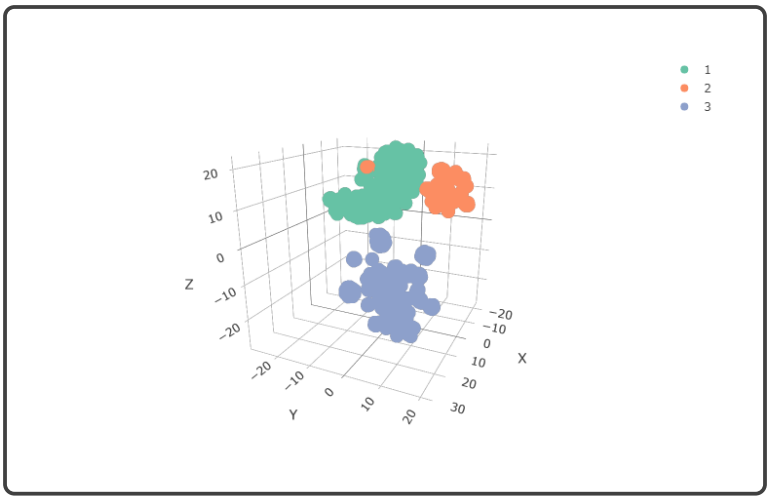
Figura 5.10: *Layouts* do conjunto de dados Statlog em duas dimensões.



(a) Amostra de 1.000 instâncias, via MDS, com a coloração atribuída pelos agrupamentos gerados pelo algoritmo PAM.



(b) Amostra de 1.000 instâncias, via ISOMAP, com a coloração atribuída pelos agrupamentos gerados pelo algoritmo PAM.



(c) Amostra de 1.000 instâncias, via t-SNE, com a coloração atribuída pelos agrupamentos gerados pelo algoritmo PAM.

Figura 5.11: *Layouts* do conjunto de dados de NFCes em três dimensões.

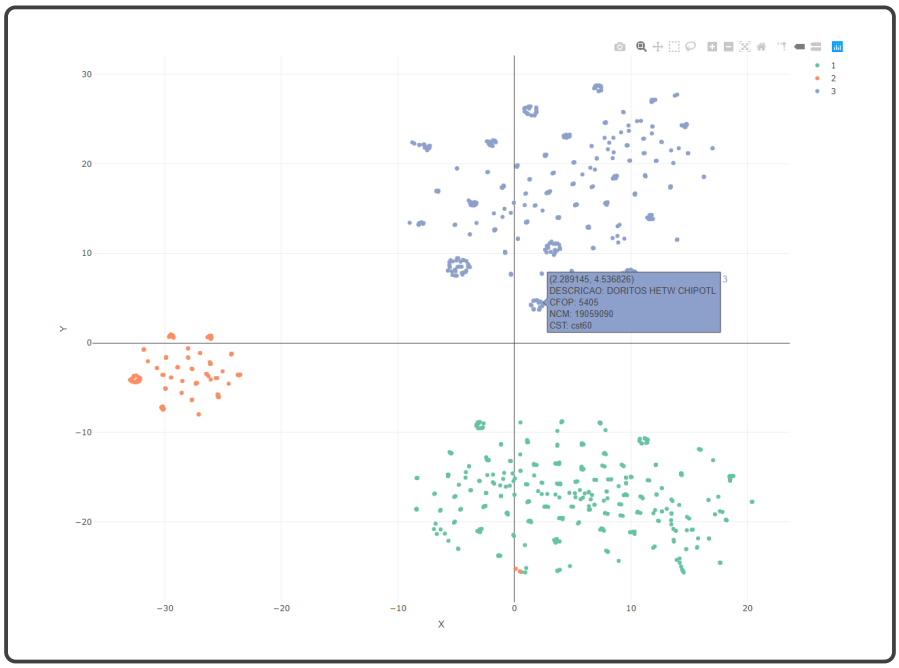


Figura 5.12: *Layout* gerado via t-SNE, demonstrando detalhes-sob-demanda.

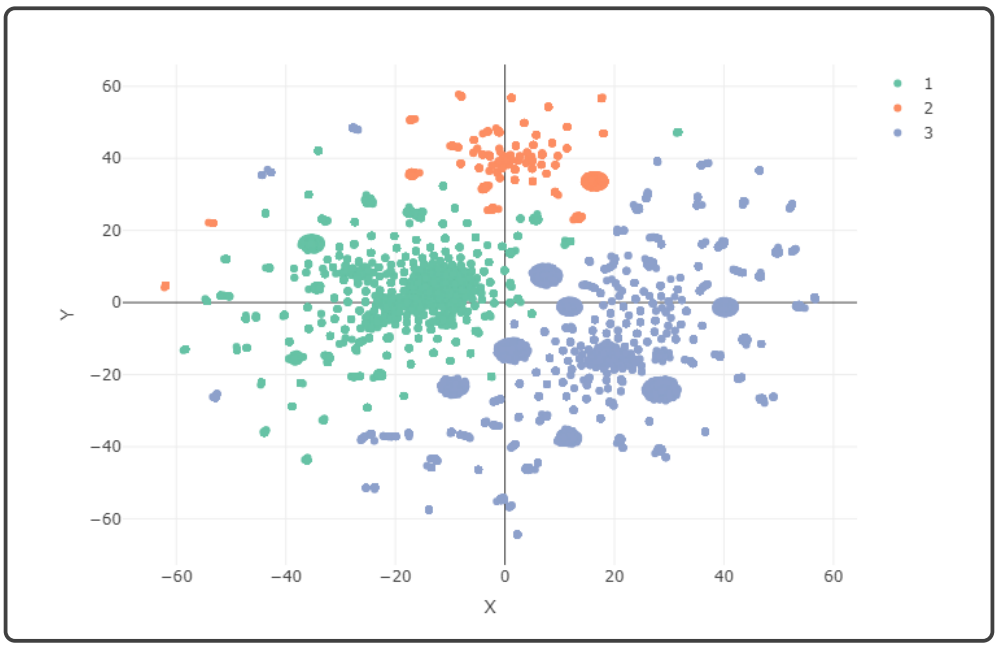


Figura 5.13: Processo de visualização exploratória, coloração atribuída pelos agrupamentos gerados pelo algoritmo PAM.

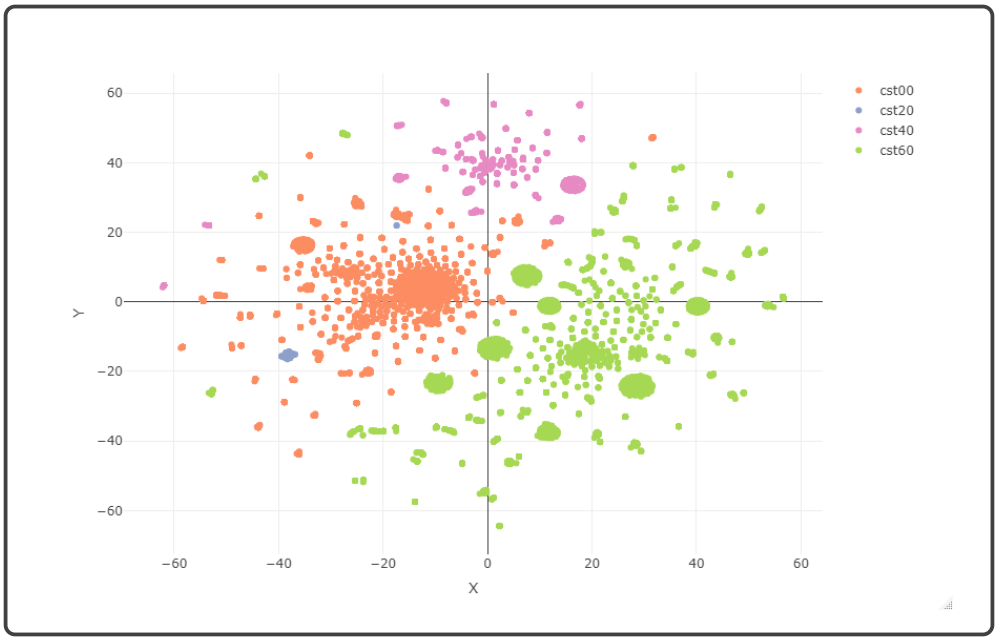


Figura 5.14: Processo de visualização exploratória, coloração atribuída pelo CST.

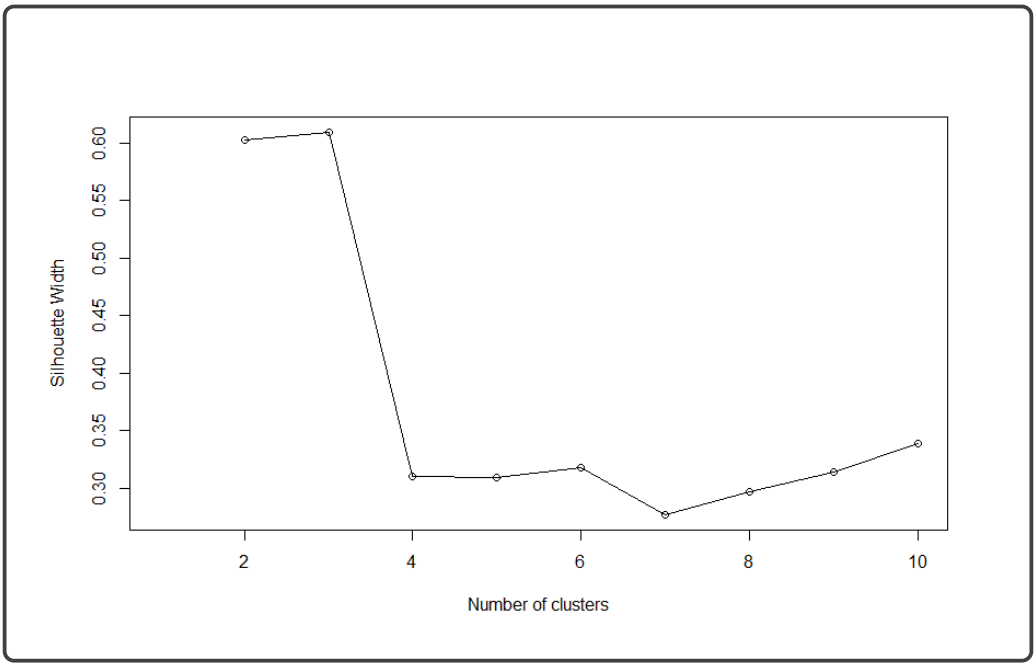


Figura 5.15: Gráfico de Silhueta, utilizado para escolher o melhor número de grupos para o algoritmo PAM.

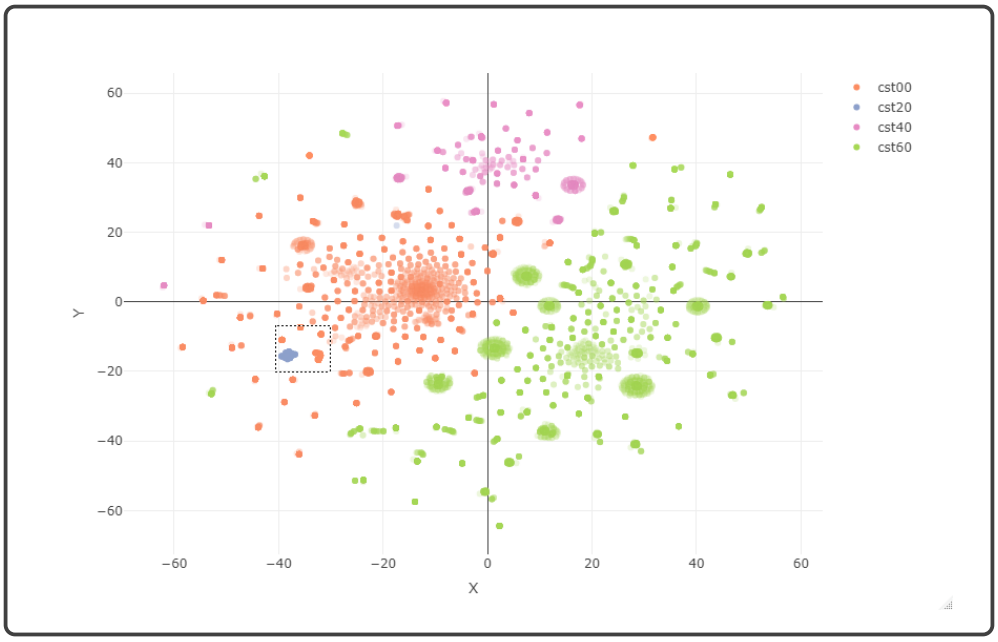


Figura 5.16: Processo de visualização exploratória utilizando a ferramenta de seleção.

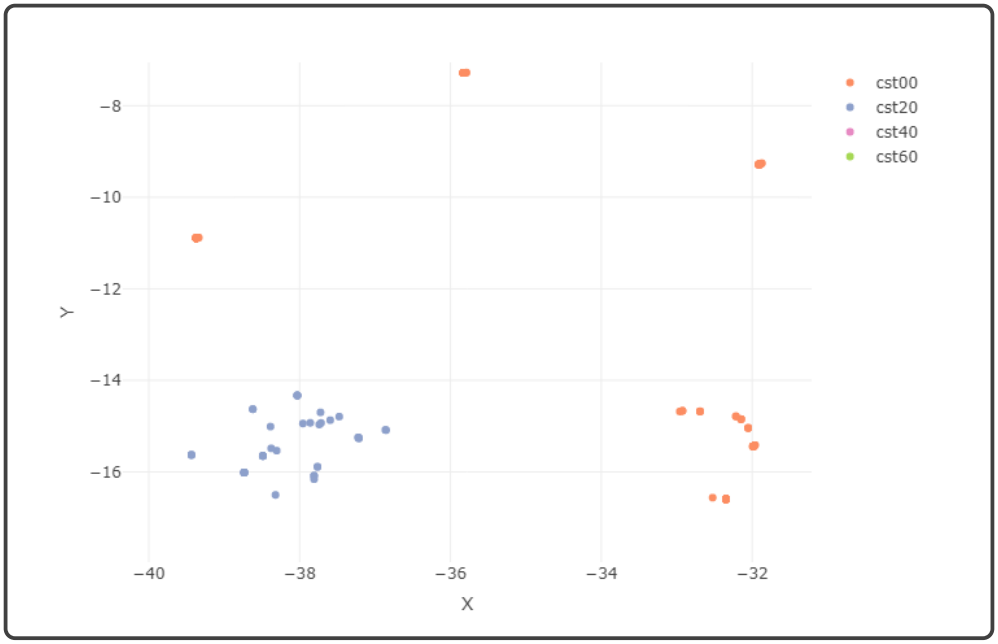


Figura 5.17: Processo de visualização exploratória utilizando a ferramenta de ampliação.

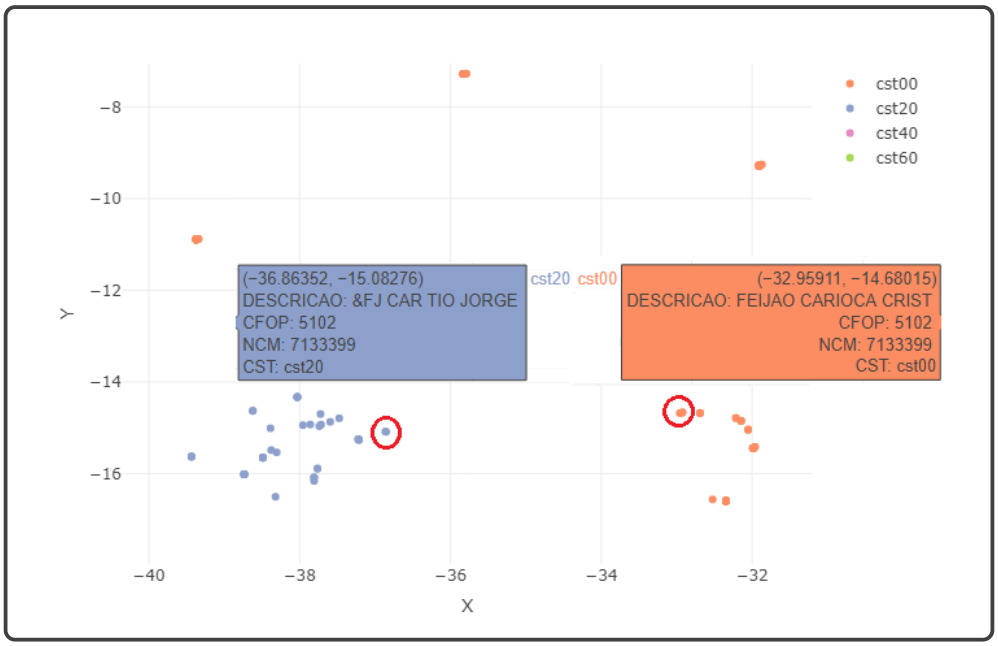


Figura 5.18: Processo de visualização exploratória observando-se detalhes-sob-demanda.

Capítulo 6

Conclusão

A tarefa de detectar fraudes em NFCes é extremamente complexa, porém se faz necessária uma vez que fraudes fiscais representam um grande prejuízo aos contribuintes. Dessa forma, essa monografia descreve um processo de visualização exploratória de dados visando auxiliar o especialista nesta tarefa. A abordagem aplicada no presente trabalho foi desenvolvida de modo a auxiliar a compreensão do especialista acerca do conjunto de dados em questão, a fim de aumentar a eficácia do processo de detecção de fraudes.

O método proposto consiste em pré-processamento de dados, seguido pela criação de uma matriz de dissimilaridade, com o objetivo de transformar um conjunto de dados que possui com atributos de diferentes tipos em uma estrutura de dados de atributos numéricos, que expressa a relação de distância entre instâncias do conjunto. Em seguida, é aplicado um algoritmo de agrupamento à matriz de dissimilaridade, de forma a revelar padrões implícitos nos dados e servir como uma ferramenta adicional para auxiliar o especialista na análise dos dados. Em sequência, a matriz de dissimilaridades é utilizada como entrada das técnicas de visualização baseadas projeções multidimensionais. Por fim, os *layouts* obtidos são empregados no processo de visualização exploratória, utilizando recursos de interação do usuário de forma a permitir que o especialista analise os padrões globais e locais dos pontos diferentes níveis de detalhe.

De forma a validar o processo proposto, foram realizados experimentos em um conjunto de dados de NFCes e em outro conjunto de dados relacionado à transações financeiras de cartões de crédito. Em ambos os casos, foram obtidos *layouts* que realçam as relações contidas no conjunto, potencialmente dando suporte ao analista a cargo da auditoria. Pôde-se verificar que os dados fraudulentos tendem a se agrupar no *layout* por serem semelhantes entre si, enquanto que os dados não-fraudulentos são posicionados de maneira mais "espalhada", ou seja, de forma mais natural em representações gráficas que armazenam informações de distância entre dados [15].

O processo proposto é capaz de gerar *layouts* representando as relações de similaridade

entre os dados de forma clara. A interatividade do especialista com a visualização também fornece versatilidade ao processo, permitindo que ele possa observar instâncias e modificar o *layout* de diversas formas. A desvantagem do processo proposto está na escalabilidade, mais especificamente, o gargalo imposto pela criação de uma matriz de dissimilaridade $n \times n$ que implica em uma complexidade de espaço $\mathcal{O}(n^2)$. Sendo assim, o processo de visualização exploratória proposto se mostra aplicável a conjuntos relativamente pequenos de dados.

Trabalhos futuros podem ampliar a gama de técnicas utilizadas, com ênfase em otimizações e variações de técnicas de projeção multidimensional, além de encontrar um meio de aproveitar as informações contidas em conjuntos de dados mistos sem o gargalo imposto por uma matriz de dissimilaridade. Algoritmos de agrupamento que são aplicáveis diretamente a dados mistos como o *k-prototypes* podem ser interessantes ao processo. Aplicações de redes neurais, tais como a utilização de *autoencoders* para melhorar os resultados obtidos pela projeção também mostram potencial. Por fim, trabalhos futuros devem buscar mais formas de integrar o elemento humano no processo, pois o problema é de grande complexidade e não parecem existir soluções triviais.

Referências

[1] Ward, Matthew O, Georges Grinstein e Daniel Keim: *Interactive data visualization: foundations, techniques, and applications*. CRC Press, 2015. ix, 9, 10, 11

[2] R Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. `https://www.R-project.org/`. ix, 12, 14, 16, 18, 35

[3] Sun, Jiao, Qixin Zhu, Zhifei Liu, Xin Liu, Jihae Lee, Zhigang Su, Lei Shi, Ling Huang e Wei Xu: *Fraudvis: understanding unsupervised fraud detection algorithms*. Em *2018 IEEE Pacific Visualization Symposium (PacificVis)*, páginas 170–174. IEEE, 2018. ix, 21, 22

[4] Deng, Ruoyu e Na Ruan: *Fraudjudger: Real-world data oriented fraud detection on digital payment platforms*. arXiv preprint arXiv:1909.02398, 2019. ix, 21, 22, 27

[5] Distrito Federal, Receita do: *Documentos fiscais eletrônicos*, 2020. `https://www.receita.fazenda.df.gov.br/aplicacoes/CartaServicos/servico.cfm?codTipoPessoa=7&codServico=772&codSubCategoria=218`. 1

[6] Wiesner, Rodrigo: *A influência do programa nota fiscal goiana no combate à sonegação fiscal em micros e pequenas empresas*. Revista Brasileira de Contabilidade, 2019. 1

[7] Heidemann, Maristela Gheller e Valmor Luiz Alievi: *Direito tributário*. Ed. Unijuí, 2012. 1

[8] Boudjeloud-Assala, Lydia, Philippe Pinheiro, Alexandre Blansché, Thomas Tamisier e Benoît Otjacques: *Interactive and iterative visual clustering*. Information Visualization, 15(3):181–197, 2016. 1, 2

[9] Witten, Ian H e Eibe Frank: *Data mining: practical machine learning tools and techniques with java implementations*. ACM Sigmod Record, 31(1):76–77, 2002. 2

[10] Piatetsky-Shapiro, Gregory: *Knowledge discovery in real databases: A report on the ijcai-89 workshop*. AI magazine, 11(4):68–68, 1990. 2

[11] Fayyad, Usama, Gregory Piatetsky-Shapiro e Padhraic Smyth: *From data mining to knowledge discovery in databases*. AI magazine, 17(3):37–37, 1996. 2

[12] Card, Mackinlay: *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999. 2

[13] Ankerst, Mihael: *Visual data mining with pixel-oriented visualization techniques.* Em *Proceedings of the ACM SIGKDD Workshop on Visual Data Mining*, página 23, 2001. 2

[14] Chen, Chun houh, Wolfgang Karl Härdle e Antony Unwin: *Handbook of data visualization.* Springer Science & Business Media, 2007. 2, 5

[15] Akoglu, Leman, Rishi Chandy e Christos Faloutsos: *Opinion fraud detection in online reviews by network effects.* Em *Seventh International AAAI Conference on Weblogs and Social Media*, páginas 1–10, 2013. 2, 51

[16] Keim, Daniel A: *Information visualization and visual data mining.* IEEE Transactions on Visualization and Computer Graphics, 8(1):1–8, 2002. 3, 11, 25, 28

[17] Nonato, Luis Gustavo e Michael Aupetit: *Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment.* IEEE Transactions on Visualization and Computer Graphics, 25(8):2650–2673, 2018. 3

[18] Bruce, Peter, Andrew Bruce e Peter Gedeck: *Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python.* O'Reilly Media, 2020. 5

[19] Tan, Pang Ning, Michael Steinbach e Vipin Kumar: *Introduction to data mining.* Pearson Education India, 2016. 6, 9

[20] Kaufman, Leonard e Peter J Rousseeuw: *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009. 6

[21] Gower, John C: *A general coefficient of similarity and some of its properties.* Biometrics, páginas 857–871, 1971. 7

[22] Gan, Guojun, Chaoqun Ma e Jianhong Wu: *Data clustering: theory, algorithms, and applications.* SIAM, 2007. 7

[23] Han, Jiawei, Jian Pei e Micheline Kamber: *Data mining: concepts and techniques.* Elsevier, 2011. 8

[24] Saket, Swarndeep e Sharnil Pandya: *An overview of partitioning algorithms in clustering techniques.* International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Vol, 5:2278–1232, 2016. 8

[25] MacQueen, James *et al.*: *Some methods for classification and analysis of multivariate observations.* Em *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, páginas 281–297. Oakland, CA, USA, 1967. 8

[26] Kassambara, Alboukadel: *Practical guide to cluster analysis in R: Unsupervised machine learning*, volume 1. Sthda, 2017. 8

[27] Hartigan, John A e Manchek A Wong: *Algorithm as 136: A k-means clustering algorithm.* Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1):100–108, 1979. 8

[28] Rousseeuw, Peter J: *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis*. Journal of Computational and Applied Mathematics, 20:53–65, 1987. 8

[29] Jin, Xin e Jiawei Han: *K-Medoids Clustering*, páginas 564–565. Springer US, Boston, MA, 2010, ISBN 978-0-387-30164-8. https://doi.org/10.1007/978-0-387-30164-8_426. 8

[30] Schubert, Erich e Peter J Rousseeuw: *Faster k-medoids clustering: improving the pam, clara, and clarans algorithms*. Em *International Conference on Similarity Search and Applications*, páginas 171–187. Springer, 2019. 9

[31] Baitaluk, Michael, Mayya Sedova, Animesh Ray e Amarnath Gupta: *Biologicalnetworks: visualization and analysis tool for systems biology*. Nucleic Acids Research, 34(suppl_2):W466–W471, 2006. 11

[32] Goodman, Alyssa A: *Principles of high-dimensional data visualization in astronomy*. Astronomische Nachrichten, 333(5-6):505–514, 2012. 11

[33] Zinovyev, Andrei: *Data visualization in political and social sciences*. arXiv preprint arXiv:1008.1188, 2010. 11

[34] Wilkinson, Leland e Michael Friendly: *The history of the cluster heat map*. The American Statistician, 63(2):179–184, 2009. 12

[35] Inselberg, Alfred e Bernard Dimsdale: *Parallel coordinates: a tool for visualizing multi-dimensional geometry*. Em *Proceedings of the First IEEE Conference on Visualization*, páginas 361–378. IEEE, 1990. 12

[36] Paulovich, Fernando Vieira: *Mapeamento de dados multi-dimensionais-integrando mineração e visualização*. Tese de Doutoramento, Universidade de São Paulo, 2008. 13

[37] Tejada, Eduardo, Rosane Minghim e Luis Gustavo Nonato: *On improved projection techniques to support visual exploration of multi-dimensional data sets*. Information Visualization, 2(4):218–231, 2003. 13

[38] Fruchterman, Thomas MJ e Edward M Reingold: *Graph drawing by force-directed placement*. Software: Practice and Experience, 21(11):1129–1164, 1991. 13

[39] Morrison, Alistair, Greg Ross e Matthew Chalmers: *Fast multidimensional scaling through sampling, springs and interpolation*. Information Visualization, 2(1):68–77, 2003. 13

[40] Sammon, John W: *A nonlinear mapping for data structure analysis*. IEEE Transactions on Computers, 100(5):401–409, 1969. 13

[41] Tenenbaum, Joshua: *Mapping a manifold of perceptual observations*. Advances in Neural Information Processing Systems, 10:682–688, 1997. 13

[42] Jolliffe, Ian T: *Principal components in regression analysis*. Em *Principal Component Analysis*, páginas 129–155. Springer, 1986. 13, 27

[43] Roweis, Sam T e Lawrence K Saul: *Nonlinear dimensionality reduction by locally linear embedding*. science, 290(5500):2323–2326, 2000. 13

[44] Maaten, Laurens van der e Geoffrey Hinton: *Visualizing data using t-sne*. Journal of Machine Learning Research, 9(Nov):2579–2605, 2008. 13, 16

[45] Mead, Al: *Review of the development of multidimensional scaling methods*. Journal of the Royal Statistical Society: Series D (The Statistician), 41(1):27–39, 1992. 14

[46] Yang, Tynia, Jinze Liu, Leonard McMillan e Wei Wang: *A fast approximation to multidimensional scaling*. Em *IEEE Workshop on Computation Intensive Methods for Computer Vision*, 2006. 14

[47] Tenenbaum, Joshua B, Vin De Silva e John C Langford: *A global geometric framework for nonlinear dimensionality reduction*. science, 290(5500):2319–2323, 2000. 15

[48] Van Der Maaten, Laurens, Eric Postma e Jaap Van den Herik: *Dimensionality reduction: a comparative review*. Journal of Machine Learning Research, 10(66-71):13, 2009. 15, 16

[49] Indyk, Piotr e Rajeev Motwani: *Approximate nearest neighbors: towards removing the curse of dimensionality*. Em *Proceedings of the 30-th Annual ACM Symposium on Theory of Computing*, páginas 604–613, 1998. 15

[50] Hinton, Geoffrey E e Sam T Roweis: *Stochastic neighbor embedding*. Em *Advances in Neural Information Processing Systems*, páginas 857–864, 2003. 16

[51] Krijthe, Jesse, Laurens van der Maaten e Maintainer Jesse Krijthe: *Package 'rtsne'*, 2018. 18, 35

[52] Ridzuan, Fakhitah e Wan Mohd Nazmee Wan Zainon: *A review on data cleansing methods for big data*. Procedia Computer Science, 161:731–738, 2019. 19

[53] Dilla, William N e Robyn L Raschke: *Data visualization for fraud detection: Practice implications and a call for future research*. International Journal of Accounting Information Systems, 16:1–22, 2015. 20

[54] Chang, Remco, Alvin Lee, Mohammad Ghoniem, Robert Kosara, William Ribarsky, Jing Yang, Evan Suma, Caroline Ziemkiewicz, Daniel Kern e Agus Sudjianto: *Scalable and interactive visual analysis of financial wire transactions for fraud detection*. Information Visualization, 7(1):63–76, 2008. 20

[55] Sarkar, Tirthajyoti: *Clustering metrics better than the elbow method*, 2019. 27

[56] Wattenberg, Martin, Fernanda Viégas e Ian Johnson: *How to use t-sne effectively*. Distill, 1(10):e2, 2016. 27, 30

[57] Shneiderman, Ben: *The eyes have it: A task by data type taxonomy for information visualizations*. Em *Proceedings of the IEEE Symposium on Visual Languages*, páginas 336–343. IEEE, 1996. 28

[58] Wickham, Hadley, Romain François, Lionel Henry e Kirill Müller: *dplyr: A Grammar of Data Manipulation*, 2020. `https://CRAN.R-project.org/package=dplyr`, R package version 1.0.2. 35

[59] Maechler, Martin, Peter Rousseeuw, Anja Struyf, Mia Hubert e Kurt Hornik: *cluster: Cluster Analysis Basics and Extensions*, 2019. R package version 2.1.0 — For new features, see the 'Changelog' file (in the package source). 35

[60] Sievert, Carson: *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman and Hall/CRC, 2020, ISBN 9781138331457. `https://plotly-r.com`. 35, 39

[61] Oksanen, Jari, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, Eduard Szoecs e Helene Wagner: *vegan: Community Ecology Package*, 2019. `https://CRAN.R-project.org/package=vegan`, R package version 2.5-6. 35