UNIVERSIDADE FEDERAL FLUMINENSE

ALEX SANDRO SILVA BATISTA DE SOUZA

JOSÉ RICARDO DA COSTA SAMPAIO

VISUALIZADOR PARA REGRAS DE ASSOCIAÇÃO: BASE DE   
DADOS DE MÚLTIPLA ESCOLHA COM MAPEAMENTO DO   
DUAL SCALING

Niterói

2019ALEX SANDRO SILVA BATISTA DE SOUZA

JOSÉ RICARDO DA COSTA SAMPAIO

VISUALIZADOR PARA REGRAS DE ASSOCIAÇÃO: BASE DE   
DADOS DE MÚLTIPLA ESCOLHA COM MAPEAMENTO DO   
DUAL SCALING

Trabalho de Conclusão de Curso submetido ao Curso de Tecnologia em Sistemas de Computação da Universidade Federal Fluminense como requisito parcial para obtenção do título de Tecnólogo em Sistemas de Computação.

Orientador:

ALTOBELLI DE BRITO MANTUAN

NITERÓI

2019

Folha reservada para a ficha catalográfica

ALEX SANDRO SILVA BATISTA DE SOUZA

JOSÉ RICARDO DA COSTA SAMPAIO

VISUALIZADOR PARA REGRAS DE ASSOCIAÇÃO: BASE DE   
DADOS DE MÚLTIPLA ESCOLHA COM MAPEAMENTO DO   
DUAL SCALING

Trabalho de Conclusão de Curso submetido ao Curso de Tecnologia em Sistemas de Computação da Universidade Federal Fluminense como requisito parcial para obtenção do título de Tecnólogo em Sistemas de Computação.

Niterói, 20 de junho de 2019.

Banca Examinadora:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Prof. Altobelli de Brito Mantuan, MSc. – Orientador

UFF – Universidade Federal Fluminense

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Prof. Jacó Júlio de Souza Costa, MSc. – Avaliador

IME – Instituto Militar de Engenharia

AGRADECIMENTOS

sampaio

A Deus, que sempre iluminou a minha caminhada.

A meu Orientador Fulano de Tal pelo estímulo e atenção que me concedeu durante o curso.

Aos Colegas de curso pelo incentivo e troca de experiências.

A todos os meus familiares e amigos pelo apoio e colaboração.

# RESUMO

Este trabalho tem como premissa a geração de visualizações que permitam o auxílio na tomada de decisões a partir da busca por relacionamentos ou padrões frequentes entre itens presentes em transações que compõem um conjunto de consultas realizadas (Base de Dados (BD)) em determinado campo ou área de interesse de cunho científico e/ou mercadológico. Para tal, serão empregadas as Regras de Associação (RA), pois além de apresentarem o conceito acima em seu escopo, possuem como bônus a facilidade de produzir as efeitos desejáveis, pois as pesquisas multidimensionais desenvolvidas por elas apresentam como métricas pilares: suporte e confiança mínimos fundamentais para o entendimento do comportamento da base de dados em análise. Além das RA, também será utilizado conjunto métodos dentre os quais, merecem menção, o algoritmo *Dual Scaling* que tem sua utilidade aplicada ao mapeamento da BD, e o cálculo de distância chi-quadrado que propiciará, como veremos detalhadamente, o cálculo de distâncias entre itens. Tais distâncias permitirão a confecção dos gráficos objeto das visualizações. Em função destas ações espera-se demonstrar claramente a aplicabilidade destes procedimentos nas tomadas de decisões dentro de qualquer cenário que se enquadre na proposta apresentada.

**Palavras-chaves:** Dual Scaling e Regras de Associação.

# ABSTRACT

This Course Completion Work {(CBT) ou (TCC)} has as premise the generation of visions that allow the aid in decision making from the search for relations or frequent patterns between items present in transactions that set up a set of queries (Data Base {(DB) ou (BD)}) in a particular field or area of ​​scientific and / or market interest. In order to do so, the Association Rules (RA) will be used, since besides presenting the above concept in their scope, they have as a bonus the ease of producing the desired ​​effects, since the multidimensional researches developed by them present as pillars metrics: support and confidence that are fundamental to the understanding of the behavior of the database under analysis. Besides the RA will also be used set methods among which, deserve mention, the Dual Scaling algorithm that has its utility applied to the mapping of BD, and the chi-square technique that will provide, as we will see in detail, the calculation of distances between items. Such distances will allow the creation of the graphs object of the visualizations. Due to these actions, it is expected to clearly demonstrate the applicability of these procedures in decision making within any scenario that falls within the proposal presented.

**Keywords**: Dual Scaling and Association Rules.

# LISTA DE ILUSTRAÇÕES

# LISTA DE TABELAS

# LISTA DE GRÁFICOS

# LISTA DE ABREVIATURAS E SIGLAS

BD – Base de Dados

CAD – *Computer Aided Design* (Desenho Assistido por Computador)

DS – *Dual Scaling*

DM – Data Mining

IDEA -- *Image Diagnosis Enhancement through Association rules*

KDD – Knowledge Discovery in Databases

RA – Regras de Associação

**SUMÁRIO**

[RESUMO 6](#_Toc11753375)

[ABSTRACT 7](#_Toc11753376)

[LISTA DE ILUSTRAÇÕES 8](#_Toc11753377)

[LISTA DE TABELAS 9](#_Toc11753378)

[LISTA DE GRÁFICOS 10](#_Toc11753379)

[LISTA DE ABREVIATURAS E SIGLAS 11](#_Toc11753380)

[1 INTRODUÇÃO 13](#_Toc11753381)

[2 TRABALHOS RELACIONADOS 15](#_Toc11753382)

[3 FUNDAMENTAÇÃO TEÓRICA 18](#_Toc11753383)

[3.1 REGRAS DE ASSOCIAÇÃO 18](#_Toc11753384)

[3.2 DUAL SCALING 20](#_Toc11753385)

[3.3 CÁLCULO DAS DISTANCIAS 24](#_Toc11753386)

[3.3.1 DISTÂNCIAS INTRA-GRUPO – ITENS 25](#_Toc11753387)

[3.3.2 DISTÂNCIAS INTRA-GRUPO - TRANSAÇÕES 26](#_Toc11753388)

[4 VISUALIZADOR 29](#_Toc11753389)

[4.1 SUBMATRIZ DE DISTÂNCIA DADA UMA RA 29](#_Toc11753390)

[4.2 VERIFICAÇÃO DAS RELAÇÕES DE DISTÂNCIAS 30](#_Toc11753391)

[4.3 PREPARAÇÃO PARA A PLOTAGEM GRÁFICA 35](#_Toc11753392)

[4.4 META INFORMAÇÃO 40](#_Toc11753393)

[4.5 41](#_Toc11753394)

[5 TESTES 42](#_Toc11753395)

[6 CONCLUSÕES 43](#_Toc11753396)

[7 BIBLIOGRAFIA 44](#_Toc11753397)

# INTRODUÇÃO

A velocidade da evolução tecnológica e a crescente utilização de bancos de dados para as mais diversas finalidades somadas ao incremento da necessidade de conhecimento mais abrangente e eficaz acerca das relações destas bases de dados transacionais torna-se imprescindível. Relacionada com esta necessidade a utilização dos conceitos de *data mining* – mineração de dados (DM), também conhecido por *Knowledge Discovery in Databases* (KDD), um conjunto de técnicas de pesquisas em bases de dados, é utilizado cada vez mais.

Dentre várias técnicas e processos utilizados para extração das correlações transacionais, e frequência de padrões em bases de dados transacionais, utilizaremos as regras de associação para detecção e extração destas correlações.

Por que utilizar as regras de associação? Por causa da facilidade de utilização dos padrões de transações em bases de dados fornecido por suas pesquisas multidimensionais baseados em suas métricas pilares: suporte, e confiança mínimos, estes conceitos e regras encontradas serão apresentados serão apresentados ao longo do projeto.

Para reorganizar e verificar como as regras de associação (RA) estão projetadas nestes espaços solução dimensionais utilizaremos o algoritmo *Dual Scaling* como ferramenta de visualização das RA, e foi utilizado o *Phyton* como linguagem de programação.

Resultados esperados através da pesquisa realizada neste projeto:

* A criação de submatrizes que representam a regra de associação;
* Implementação do cálculo da distância entre os pontos médios das regras;
* Relacionar as distâncias entre antecedente e consequente;
* Projetar as transações neste espaço solução.

Para maiores informações a respeito do projeto, o código fonte encontra-se disponível no repositório *github* através do link <https://github.com/altobellibm/CEDERJ_2019_ALEX_SOUZA_E_JOSE_SAMPAIO.git>

# TRABALHOS RELACIONADOS

Neste capítulo serão citados alguns trabalhos que utilizam este algoritmo de identificação de padrões chamado regras de associação (RARARARARA) para auxiliar na tomada de decisões em área de pesquisa a respeito dos possíveis problemas relacionados à respectiva base de dados pesquisada.

Neste primeiro exemplo de utilização das regras de associação foi pesquisada a base de dados da Secretaria de Saúde de Londrina que une características sócio-econômincas a respeito de dados de procedimentos realizados em internações hospitalares (SILVA, 2004), este estudo tem como objetivo melhorar o entendimento geral sobre as características do município, e teve como grandes obstáculos a descentralização das fontes de dados, e a inconsistência da base de dados, após a superação das inconsistências apresentadas pela base de dados foram destacados alguns resultados importantes, tais como:

* 88,85% das safececomias interna radical são realizadas em pessoas do sexo feminino que trabalhavam no lar com mais de 35 anos, procedimento realizado devido as dores que podem ser agravadas pelo tipo de atividade física(ocasionada pelo trabalho no lar) e também pela idade.
* 80,45% das herniorrafias inguinais(unilateral) múltiplas são realizadas em pessoas do sexo masculino em crianças de 0 a 4 anos, foi caracterizado um erro de nomenclatura nos procedimentos em crianças desta faixa etária diminuindo o custo de funcionamento dos hospitais visto que a herniorrafias inguinal pode levar a uma internação de urgência ou emergencia enquanto que o tratamento urológico da hidrocele comunicante é um procedimento eletivo.
* Verificou-se que em áreas menos favorecidas é alta a incidência de procedimentos de parto e pediátricos de urgência ou emergência.

No segundo exemplo foram estudadas a utilização das RA a respeito das forças de mercado que regem a comercialização de touros nelore com avaliação genética. O estudo foi realizado pelo programa Nelore Brasil (NOMELINI, REZENDE, *et al.*, 2010). A identificação das métricas das RA foi feita através do método da análise de Pareto. Esse estudo evidencia a eficácia da utilização das métricas das regras de associação para identificação de padrões de mercado mandatódios implícitos nas transações de grandes Bancos de dados, visto que a base estudada teve aproximadamente 20000 cabeças de gado comercializados por fazendas de todo o país. Foram utilizados como base do estudo das RA os 15 atributos mais desejados pelo mercado indicando as principais causas e efeitos da comercialização dos rebanhos de gado nelore no país . O estudo sugere que o mérito genético total, índice oficial do programa Nelore Brasil seja um índice fundamenta para a comercialização dos touros no país, indentificando combinações de atributos genétcos, geográfico, e temporais mandatórios nas segmentações de rebanhos de touros para comercialização pelo programa Nelore Brasil.

Neste terceiro exemplo as RA são utilizadas para dar suporte a dois tipos de sistemas médicos: o sistema de busca por conteúdo em imagens e os sistemas de auxílio ao diagnóstico (RIBEIRO, 2008). No sistema de buscas por conteúdo o emprego da RA tem por finalidade a redução dos vetores característicos de representação das imagens e reduzir as redundâncias existentes entre as característica de baixo nível das imagens e seu significado semântico com a ajuda do algoritmo *StARMiner*. Enquanto que no sistema de auxílio ao diagnóstico para dar suporte aos sistemas *CAD*, foi desenvolvido o método IDEA(*Image Diagnosis Enhancement through Association rules*), que utiliza as RA para sugerir uma segunda opinião automaticamente, ou um diagnóstico preliminar de uma nova imagem para acelerar o diagnóstico de um radiologista, ou para prover auxílio nos diagnósticos médicos baseados em RA. Os resultados mais relevantes apresentados por este estudo foram o desenvolvimento e validação de técnicas de segmentação e extração de características, e o aumento da precisão de consultas utilizando realimentação de relevância.

Os experimentos realizados referenciam a utilização das RA como ferramenta poderosa na descoberta de padrões em sistemas médicos, e como referencial na busca por conteúdo e diagnóstico de imagens médicas.

Os estudos citados ao longo deste capítulo do trabalho evidenciam, a contribuição agregada nas tomadas de decisões administrativas proporcionadas pela utilização do algoritmo de regras de associação. Visto que RA é uma ferramenta de identificação de padrões complexos e multidimensionais entre atributos de bases de dados naturezas distintas.

# FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão apresentados os conceitos aplicados ao estudo das RA, e seu comportamento sobre as relações estatísticas complexas entre atributos de uma base de dados, estes conceitos podem ser obtidos através de várias técnicas, porém o objetivo deste estudo é evidenciar estas relações através do estudo das regras de associação que estão implícitas nas relações multidimensionais de uma base de dados, estas relações são evidenciadas através de cálculos específicos. Cabe ressaltar, que a obtenção destas regras visa ajudar as pessoas responsáveis pela base de dados na tomada de decisões, visto que está ferramenta proporciona uma visão mais abrangente acerca de variáveis totalmente desconhecidas do ponto de vista administrativo.

## REGRAS DE ASSOCIAÇÃO

Formada por milhares de itens armazenados uma grande base de dados é crescente necessidade de conhecimento a respeito das associações das transações entre estes dados não categóricos, ou seja, não aplicável a dados numéricos, é o objetivo deste algoritmo chamado de regras de associação.

Ferramentas primordiais para a verificação e a validação das RA os algoritmos de *data mining* têm por objetivo, então, encontrar todas as associações relevantes entre itens nas relações do tipo X (antecedente da regra) ⇒ Y (consequente da regra), e no modelo matemático proposto (AGRAWAL, IMIELINSKI e SWAMI, 1993), as RA devem atender as métricas de suporte e confiança mínimos propostos na pesquisa feita a base de dados. Seja I ={i1,...,in} um conjunto de literais, denominados itens. Qualquer conjunto é chamado de *itemset*. Logo um *itemset* X com k elementos é chamado de *itemset-k*. Seja R uma tabela com tuplas t que envolvem elementos que são subconjuntos de I. A tupla t suporta um *itemset* X, se Seja |Z| o número total de ocorrências do *itemset* Z na tuplas da tabela T. As métricas de suporte *sup* e confiança conf são apresentadas a seguir:

|  |  |  |
| --- | --- | --- |
|  |  | () |

|  |  |  |
| --- | --- | --- |
|  |  | () |
|  |  |  |

O problema da obtenção das RA, foi como foi estabelecido inicialmente, envolve o descobrimento de regras que satisfaçam as restrições de suporte mínimo (*minsup*) e confiança mínima (*minconf*) especificadas pelo usuário.

O suporte de um *itemset* X é a razão entre o número de tuplas em T que suportam X e o número total de tuplas de R. O suporte é utilizado como restrição para a obtenção das regras. Um *itemset* X é chamado de *itemset frequente* se o suporte de X for maior ou igual ao suporte mínimo especificado pelo usuário. Também podemos traduzir uma regra de associação , onde , pode ser traduzida por “se X então Y”, a qual indica que quando ocorre X tende a ocorrer Y, enquanto que a confiança de uma regra é a razão entre o número de tuplas que contém X e Y, e o número de tuplas que contém X, também chamada de medida de força de uma regra. Dentre as técnicas usadas para mineração das RA destaca-se o algoritmo Apriori como principal ferramenta de verificação das RA. Descrito por Tan, Steinbach e Kumar (2009), parte do princípio que, se a frequência de um conjunto de itens é relevante, implica na relevância dos seus subconjuntos também, ou seja, se {x, y, z} é um conjunto de itens frequentes, qualquer transação que contenha {x, y, z} deve conter seus subconjuntos {x,y}, {x, z}, {y,z}, {z}, {x}, {y}. Cabe ressaltar que este algoritmo suporta um grande número de atributos, fornecendo inúmeras alternativas combinatórias entre os atributos através de buscas sucessivas a base de dados.

## DUAL SCALING

Método versátil para análise de dados, o *Dual Scaling* (DS)foi desenvolvido por Nishisato para ser uma *ferramenta para inspeção visual* de indivíduos e suas preferências para estímulos coletados através de questionários de opinião. O mapeamento resultante do DS, transforma cada atributo ou associação através de um ponto no espaço-solução resultante. Os comportamentos e preferências de grupos de indivíduos que tem opiniões similares emergem da distribuição de pontos porque indivíduos e estímulos relacionados são mapeados pertos um dos outros, enquanto dados não relacionados aparecem apartados no espaço-solução (FORTES, 2019).

Apesar de ter sido desenvolvido para análise de preferências de indivíduos, segundo Nishisato o *Dual Scaling* pode descobrir estilos de respostas em praticamente todos os tipos de bases de dados.

Os dados resultantes da análise utilizando o DS são expressados em função do padrão de resposta escolhido, e as unidades de análise são as opções de respostas. O DSprocura as combinações ponderadas mais informativas de categorias de itens, e gera uma matriz de correlação entre os itens para cada dimensão.

Combinações não-lineares de categorias de itens estão envolvidas em cada dimensão. No DS, a correlação linear é maximizada pela transformação das categorias de forma linear ou não linear, dependendo dos dados.

Para o estudo a respeito da importância das RA pesquisaremos bases de dados com dados do tipo múltipla escolha que será representada como , e será nossa matriz de padrão de respostas, baseada na tabela de padrão de respostas de 0s e 1s, de tamanho , onde cada transação é um indivíduo (linhas da matriz), e os itens ficam organizados como possíveis estímulos ou respostas de múltipla escolha (colunas da matriz).

Inicialmente para o cálculo do DStem por objetivo descobrir a quantidade de dimensões do espaço-solução (, através da seguinte equação:

|  |  |  |
| --- | --- | --- |
|  | , | () |

onde é o número de tuplas de nossa matriz de padrão de respostas , e é o número de categorias dos itens de resposta (questões).

Em seguida, definimos o vetor através do somatório das linhas da matriz , e o vetor como o somatório das colunas da matriz . Esses vetores são conhecidos como vetores de frequência de linhas e colunas de .

|  |  |  |
| --- | --- | --- |
|  |  | () |

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  | () |

Dados os vetores de frequência, vamos gerar para cada um deles uma matriz diagonal. A matriz diagonal de linhas é gerada através da diagonalização do vetor de frequência de linhas , que significa gerar uma matriz quadrada de tamanho , onde os valores do vetor serão os valores da diagonal principal da matriz; do mesmo modo, a matriz diagonal de colunas é gerada através da diagonalização do vetor de frequência de colunas , que segue o mesmo modo de operação explicado acima:

|  |  |  |
| --- | --- | --- |
|  |  | () |

|  |  |  |
| --- | --- | --- |
|  |  | () |
|  |  |  |

O próximo passo é definir as correlações entre colunas da matriz , cujo resultado chamaremos de matriz , dada pelo resultado da equação:

|  |  |  |
| --- | --- | --- |
|  |  | () |

A transposição de matriz é representada por , enquanto a inversão de matriz é representada por *.*

Representado por e , respectivamente, o vetor de autovalores e a matriz de autovetores de . O vetor de autovalores deve ser ordenado, de tal forma que . As colunas da matriz de autovetores devem acompanhar a ordenação de seus respectivos autovalores. Uma vez ordenados, o primeiro item do vetor de autovalores , bem como a primeira coluna da matriz de autovetores , devem ser descartados; e o número máximo de elementos em e colunas em devem ser iguais ao . Logo, temos como o vetor final de autovalores, e como a matriz final de autovetores.

Em seguida, vamos calcular a matriz , dada pela equação:

|  |  |  |
| --- | --- | --- |
|  |  | () |

Cabe ressaltar que a operação entre as matrizes da equação que define , representada pelo símbolo , é o produto de *Hadamard* . Obtida a matriz , deve-se calcular o seu vetor de frequência de colunas , através do somatório de todos os valores das colunas da matriz .

|  |  |  |
| --- | --- | --- |
|  |  | () |

A seguir no cálculo do *DS*, é calculado o vetor , cujos valores representam os multiplicadores das colunas da matriz final de autovetores para chegarmos à matriz de pesos padrão dos itens (*x-normed weights*). O vetor pode ser definido pela equação:

|  |  |  |
| --- | --- | --- |
|  |  | () |

onde representa o somatório de todos os valores da matriz de padrão de respostas . Uma vez conhecido o vetor de multiplicadores , calculamos então a matriz de pesos padrão dos itens, representada por , e dada pela equação:

|  |  |  |
| --- | --- | --- |
|  | . | () |

As coordenadas finais de cada um dos itens no espaço-solução são dadas pela matriz de pesos projetados dos itens (*x-projected weights*), representada por e obtida pela equação:

|  |  |  |
| --- | --- | --- |
|  | , | () |

onde é um vetor que contém os multiplicadores para as colunas da matriz de pesos padrão , definido pela equação:

|  |  |  |
| --- | --- | --- |
|  | . | () |

De posse das coordenadas dos itens no espaço-solução, o próximo passo é calcular as coordenadas das transações. A primeira etapa deste cálculo consiste na multiplicação da matriz de padrão de respostas pela matriz de pesos padrão de itens . A matriz resultante deste produto é representada pela seguinte equação:

|  |  |  |
| --- | --- | --- |
|  | . | () |

Dados os valores da matriz , calculamos a matriz de pesos padrão das transações (*y-normed weights*), representada por , utilizando a fórmula abaixo:

|  |  |  |
| --- | --- | --- |
|  |  | () |

As coordenadas finais de cada uma das transações no espaço-solução são dadas pelos valores da matriz de pesos projetados das transações (*y-projected weights*), representada por e obtida pela equação:

|  |  |  |
| --- | --- | --- |
|  |  | () |

O último passo do *Dual Scaling* é calcular , vetor com os valores do percentual de representatividade de cada uma das dimensões do espaço-solução na solução como um todo. O cálculo de acontece através da equação:

|  |  |  |
| --- | --- | --- |
|  |  | () |

Os resultados obtidos através dos cálculos dos algoritmos do DS permitem verificação das características apresentadas pelas RA verificadas nas pesquisas feitas a base tais como:

* Aproximação gráfica da representação dos atributos que tem suas associações verificadas nos algoritmos;
* Simplificação da quantidade de dimensões do espaço solução para visualização linear das RA;

## CÁLCULO DAS DISTANCIAS

As distâncias podem ser classificadas de duas formas: (1) distância intra-grupo (*within-set distance*), que são as distâncias de um item (coluna) para os demais itens e as distâncias de uma transação (linha) para as demais transações; e (2) distância inter-grupo (*between-set distance*), que são as distâncias dos itens para as transações, e vice-versa.

### DISTÂNCIAS INTRA-GRUPO – ITENS

Para descobrir a distância quadrada entre os itens em um espaço-solução de n dimensões utilizando a métrica chi-quadrado, empregaremos a equação abaixo:

|  |  |  |
| --- | --- | --- |
|  | , | () |

onde e são as -ésimas coordenadas dos itens indexados por e , respectivamente, e são os k-ésimos índices do vetor de frequência de colunas e é o número de linhas da matriz de padrão de respostas .

Através das coordenadas de cada item, é possível então calcular a matriz de distância quadrada entre eles. Essa matriz é de extrema importância para a análise, pois quanto menor a distância ente os itens, mais relacionados eles estão. Utilizando a equação **Erro! Fonte de referência não encontrada.**, vamos exemplificar o cálculo da distância dos itens 3 (pressão alta) e 9 (idoso). Temos então:

.

Quebrando a equação nos valores de , demostraremos os cálculos de forma detalhada para , e apenas os resultados para os demais valores de k. Com isso, temos:

A distância quadrada final então entre os itens 3 e 9 é . A matriz completa da distância quadrada entre os itens pode ser visualizada na **Erro! Fonte de referência não encontrada.**.

Se representarmos essa distância entre os itens utilizando gráficos de dispersão, podemos facilmente visualizar os itens que tem maior relação entre si, que são aqueles que estão mais pertos do eixo x e representados por um círculo maior. Podemos visualizar os gráficos dos resultados para cada conjunto de itens nos gráficos 3, 4, 5, 6, 7 e 8, que representam, respectivamente, a distância de cada um dos itens para os itens das categorias pressão, enxaquecas, idade, ansiedade, peso e altura.

### DISTÂNCIAS INTRA-GRUPO - TRANSAÇÕES

Para descobrir a distância quadrada entre as transações em um espaço-solução de n dimensões utilizando a métrica chi-quadrado, utilizaremos a equação abaixo:

|  |  |  |
| --- | --- | --- |
|  |  | () |

onde e são as -ésimas coordenadas das transações indexadas por e , respectivamente, e são os k-ésimos índices do vetor de frequência de linhas e é o número de colunas da matriz de padrão de respostas .

Através das coordenadas de cada transação, é possível então calcular a matriz de distância quadrada entre elas. Utilizando a equação **Erro! Fonte de referência não encontrada.**, vamos exemplificar o cálculo da distância das transações 2 e 5. Temos então

.

Quebrando a equação nos valores de , demostraremos os cálculos de forma detalhada para , e apenas os resultados para os demais valores de k. Com isso, temos:

A distância final então entre as transações 2 e 5 é . A matriz completa da distância quadrada entre as transações pode ser visualizada na {**Erro! Fonte de referência não encontrada.**.} (o que ocorre aqui)

Se representarmos essa distância entre as transações utilizando gráficos de dispersão, podemos facilmente visualizar as transações que tem maior relação entre si, que são aquelas que estão mais pertos do eixo x e representados por um círculo maior. Podemos visualizar os gráficos dos resultados de cada transação nos gráficos 9, 10, 11, 12 e 13, que representam, respectivamente, a distância de cada uma das transações para as demais transações, agrupadas de 3 em 3.

# VISUALIZADOR

Nesta seção serão abordados métodos diferenciados de apresentação dos resultados para determinado conjunto de RA de modo a facilitar sobremaneira a análise dos dados categóricos de múltipla escolha obtidos em determinada pesquisa. Cabe salientar que não existe a preocupação primeira e específica com a informação em si, mas sim no como, a partir dos dados, pode-se criar formas simplificadas de análise em função de regra de associação específica.

Atualmente, devido ao enorme número de informações geradas de inúmeras formas, é imprescindível que o observador seja capaz de visualizar os resultados de suas pesquisas de maneira simples e eficiente, de modo que tal análise visual se dê de maneira amigável e satisfatória levando a resultados importantes à cerca dos dados de que disponha, fazendo-o aprender com os acertos e possíveis erros detectados nas visualizações, auxiliando-o assim na tomada de decisões. Para tal, propomos a adoção do Dual Scaling mencionado na Seção [3.2](#_dUAL_sCALING).

## SUBMATRIZ DE DISTÂNCIA DADA UMA RA

Em um primeiro momento serão selecionadas as submatrizes de distância dos itens a partir de uma lista de RA previamente produzida. Neste processo serão empregados:

* o vetor de frequência de itens que nos informa o número de indivíduos, no grupo pesquisado, que apresentam determinada característica, i.e., respondem positivamente àquele item;
* o vetor de multiplicadores dos pesos padrão dos itens referentes aos quadrados dos autovalores de cada dimensão;
* a matriz de pesos projetados responsável por fornecer as coordenadas finais de cada um dos itens no espaço solução.

De posse dessas informações mapeiam-se as diferentes submatrizes a partir do arquivo de regras, que se encontra no formato: A ==> B #SUP: C #CONF: D, onde os dados de interesse são:

A – Conjunto de itens antecedentes (mínimo de 1 item);

B – Conjunto de itens consequentes (mínimo de 1 item);

C – Suporte, visto no detalhamento das RA;

D – Confiança, vista no detalhamento das RA;

Separam-se estes dados regra por regra selecionando-se as submatrizes, tanto para antecedentes quanto para consequentes, utilizando-se iterações sucessivas e armazenando-as em um arquivo no formato .csv a partir do qual gerar-se-ão as visualizações relevantes.

## VERIFICAÇÃO DAS RELAÇÕES DE DISTÂNCIAS

É chegado o momento trabalharmos com as submatrizes obtidas segundo descrito na subseção anterior com o intuito de encontramos os pontos médios de antecedentes e consequentes, para tal será aplicada a fórmula (15) onde é o ponto desejado, é o número de itens de determinada RA e é o vetor de coordenadas para cada item presente na mesma.

|  |  |  |
| --- | --- | --- |
|  |  | () |

Tomemos como exemplo a regra da linha 272 do arquivo bloodpressurenishisatobook.d18.n15.txt indicada por: 9, 16 ==> 6, 12, onde 9 (idoso) e 16 (baixo) são os itens antecedentes e 6 (enxaqueca frequente) e 12 (ansiedade alta) os consequentes.

Aplicando os dados na equação (21) teremos:

Onde:

Sendo assim, obtemos o ponto médio dos itens antecedentes ():

Cálculo semelhante se dará com os itens consequentes de modo a obtermos o ponto médio desejado (P).

De posse destes valores já é possível calcular as distâncias entre os pontos médios de antecedentes e consequentes () bem como as distâncias de cada um dos pontos médios de antecedentes () e consequentes () até a origem, o que se dará através da métrica chi-quadrado, aplicada aqui por ser uma das distribuições mais utilizadas em estatística inferencial e permitir a avaliação quantitativamente em relação entre o resultado de um experimento e a distribuição esperada para um fenômeno. Este método encontra-se muito bem e didaticamente pormenorizado em (FORTES, 2018).

Ocorre que, como vimos na Seção [3.2](#_dUAL_sCALING), faz-se necessário o conhecimento do suporte referente a cada item da base de dados, entretanto não existe suporte que corresponda a um ponto médio e nem mesmo a origem, o que nos remete a imposição de criarmos uma pequena adaptação neste tocante. Esta ocorrerá de duas maneiras distintas.

Na determinação da distância entre os pontos médios de antecedentes e consequentes empregar-se-á método semelhante àquele utilizado para o suporte a partir da Matriz Padrão de Repostas (MPR) (F em) tomando-se apenas os itens indicados em cada uma das RA, enquanto que no do cálculo das distâncias dos pontos médios em relação a origem os suportes usados na determinação anterior serão reutilizados ocorrendo porém a simplificação natural da origem que, indicada por 0 (zero) fará o termo assim como a expressão, assumirem valor idêntico à origem, deixando a fórmula número ... com a seguinte grafia:

Onde, dirá respeito ao ponto médio antecedente ou consequente conforme o caso.

Para dar maior clareza façamos uso do exemplo acima, onde já são conhecidos os valores dos pontos médios de antecedentes e consequentes. A base de dados se dá a partir de um questionário médico composto por seis perguntas com o objetivo de avaliar a pressão arterial de pacientes (material fornecido por Nishisato). Vejamos:

1. Como você avalia a sua pressão sanguínea? (Baixa, Normal, Alta)

Itens: 1, 2, 3

2. Você tem enxaquecas com que frequência? (Raramente, Algumas Vezes, Sempre)

Itens: 4, 5, 6

3. Qual a sua idade? (20-34, 35-49, 50-65)

Itens: 7, 8, 9

4. Como você avalia seu nível diário de ansiedade? (Baixa, Normal, Alta)

Itens: 10, 11, 12

5. Como você avalia o seu peso? (Abaixo do Peso, Normal, Acima do Peso)

Itens: 13, 14, 15

6. Como você avalia a sua altura? (Baixo, Mediano, Alto)

Itens: 16, 17, 18

O questionário realizado com 15 indivíduos, foi tabulado no padrão 0 e 1, onde 0 corresponde à resposta negativa e 1 à positiva para cada um dos itens. O resultado pode ser observado na Tabela 1. Nesta, cada indivíduo é tratado como uma transação enquanto que as respostas o são como itens. Os vetores padrão de respostas para cada um dos itens tomados em nosso exemplo em estão destacados em negrito com o exclusivo intuito de facilitar a percepção pelo leitor.

Tabela 1 – Matriz Padrão de Respostas.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Transações | | | | | | | | | | | | | | |
| Itens | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 2 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 |
| 3 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 4 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| **6** | **1** | **1** | **1** | **1** | **0** | **0** | **0** | **1** | **0** | **1** | **0** | **0** | **1** | **1** | **1** |
| 7 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| 8 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| **9** | **1** | **0** | **1** | **1** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **1** | **1** | **0** | **1** |
| 10 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 11 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| **12** | **1** | **1** | **1** | **1** | **0** | **1** | **0** | **1** | **0** | **0** | **1** | **1** | **1** | **0** | **1** |
| 13 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 |
| 14 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| 15 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| **16** | **1** | **0** | **0** | **1** | **0** | **1** | **0** | **0** | **0** | **0** | **0** | **0** | **1** | **1** | **0** |
| 17 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 |
| 18 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |

Estes vetores, 9 e 16 para antecedentes e 6 e 12 para consequentes, demonstram claramente que o numerador do suporte dos antecedentes () é 3 enquanto que o dos consequentes () é 7. Como assim? Simples! Este valor (numerador do suporte) se dá quando, para determinada transação, ocorre resposta positiva (1) para os itens, tanto antecedentes como consequentes. Observe que comparando transação por transação (coluna por coluna) dos itens antecedentes é possível notar que as transações 1, 4, 13 estão assinaladas com ‘1’ o que impõe numerador igual a 3. Fazendo o mesmo estudo para os vetores dos itens consequentes vê-se claramente o valor 7 como resultado. Estes numeradores serão empregados na fórmula ... donde obteremos a .

Pormenorizando o cálculo de para o exemplo tomado, veremos que:

Efetuando o cálculo aproximado para cada valor de teremos:

E assim vem que para

Repetindo o cálculo para cada valor de teremos como resultado final:

Exemplificando agora o cálculo da distância do ponto médio antecedente até a origem (que se dará através da equação (16) e tomando uma vez mais a RA acima escolhida constataremos:

Para

Replicando o procedimento para os demais valores de o valor final para a distância assim calculada será .

Usar-se-ão estas distâncias para, em função de critério predefinido, montarmos as visualizações objeto deste trabalho.

## PREPARAÇÃO PARA A PLOTAGEM GRÁFICA

Entramos no estágio de finalização dos cálculos visando a produção final das visualizações e para tal faremos uso do cálculo do lift e da all-confidence

Na mineração de dados e no aprendizado de regras de associação, o **lift** é uma medida do desempenho de um modelo de segmentação (regra de associação) na previsão ou classificação de casos como tendo uma resposta aprimorada (com relação à população como um todo), medida contra uma segmentação aleatória modelo. Um modelo de segmentação está fazendo um bom trabalho se a resposta dentro do alvo for muito melhor do que a média da população como um todo. O aumento é simplesmente a proporção desses valores: resposta ao alvo dividida pela resposta média.

Por exemplo, suponha que uma população tenha uma taxa de resposta média de 5%, mas um determinado modelo (ou regra) identificou um segmento com uma taxa de resposta de 20%. Então esse segmento teria um aumento de 4,0 (20% / 5%).

Tipicamente, o modelador procura dividir a população em quantis e classificar os quantis por sustentação. As organizações podem então considerar cada quantil e, pesando a taxa de resposta prevista (e o benefício financeiro associado) em relação ao custo, eles podem decidir se vão ou não comercializar para esse quantil. O aumento é análogo à métrica de precisão média da recuperação de informações, se considerarmos a precisão (fração dos positivos que são verdadeiros positivos) como a probabilidade de resposta desejada.

A curva de elevação também pode ser considerada uma variação da curva característica de operação do receptor (ROC), e também é conhecida em econometria como a curva de potência ou Lorenz. [1]

Suponha que o conjunto de dados que está sendo extraído seja:

Antecedente Consequente

Um 0

Um 0

Um 1

Um 0

B 1

B 0

B 1

onde o antecedente é a variável de entrada que podemos controlar, e a consequente é a variável que estamos tentando prever. Os problemas reais de mineração normalmente teriam antecedentes mais complexos, mas geralmente focam em conseqüentes de valor único.

A maioria dos algoritmos de mineração determinaria as seguintes regras (modelos de segmentação):

Regra 1: A implica 0

Regra 2: B implica 1

porque estes são simplesmente os padrões mais comuns encontrados nos dados. Uma simples revisão da tabela acima deve tornar essas regras óbvias.

O suporte para a Regra 1 é 3/7 porque é o número de itens no conjunto de dados em que o antecedente é A e o consequente 0. O suporte para a Regra 2 é 2/7, porque dois dos sete registros atendem ao antecedente de B e o conseqüente de 1. Os suportes podem ser escritos como:

A confiança da Regra 1 é 3/4, porque três dos quatro registros que atendem ao antecedente de A atendem ao consequente de 0. A confiança da Regra 2 é 2/3, porque dois dos três registros que atendem ao antecedente de B atendem ao consequente de 1. As confidências podem ser escritas como:

O lift pode ser encontrado dividindo a confiança pela probabilidade incondicional do conseqüente, ou dividindo o suporte pela probabilidade do antecedente multiplicar a probabilidade do conseqüente, assim:

O lift da Regra 1 é

O lift da Regra 2 é

Se alguma regra tivesse um lift de 1, isso implicaria que a probabilidade de ocorrência do antecedente e a do consequente são independentes uma da outra. Quando dois eventos são independentes um do outro, nenhuma regra pode ser desenhada envolvendo esses dois eventos.

Observe que, embora a Regra 1 tenha maior confiança, ela tem menor elevação. Intuitivamente, parece que a Regra 1 é mais valiosa por causa de sua maior confiança - parece mais precisa (melhor suportada). Mas a precisão da regra independente do conjunto de dados pode ser enganosa. O valor do aumento é que ele considera a confiança da regra e o conjunto geral de dados.

{\displaystyle lift={\frac {P(A\cap B)}{P(A)\*P(B)}}}

A *H-confidence* (H-confiança) também conhecida como *all-confidence* (confiança total), é uma medida da associação de itens aplicada quase que exclusivamente a atributos binários. Se um conjunto de itens tem um h-confiança mais do que um limite especificado pelo usuário, hc, então o conjunto de itens (*itemset*) é chamado de hyperclique. Definição 1 define hipercliques e h-confiança mais formalmente. As quantidades de "apoio" e "confiança" são as definidas na análise de associação padrão

[1, 29].

DEFINIÇÃO 1. Hyperclique Um conjunto de itens (atributos binários),

X, forma um hyperclique com um nível particular de h-confiança,

onde h-confiança é definida como

hconf (X) = min

iX

{confiança ({i}! {X - {i}})} (2)

= suporte (X) / max

iX

{suporte ({i})} (3)

H-confiança está entre 0 e 1, com um valor de 0 indicando

nenhuma associação e um valor de 1 indicando a associação mais forte

entre um grupo de itens, isto é, os itens sempre ocorrem juntos.

Assim, a confiança h pode ser usada como uma medida de similaridade entre

os atributos em uma matriz de dados binários.

Especificamente, a matriz de adjacência A de uma rede de interação de proteínas

é considerado como uma matriz de dados binários, tratando suas linhas como

transações e colunas como itens. (Note que ambos correspondem a

proteína). Em seguida, uma matriz de adjacência ponderada A ′ das mesmas dimensões

como o original pode ser gerado usando A ′ (i, j) = hconfidence

(eu j). Informalmente, a confiança h de um par de proteínas,

p1 e p2, serão altos se p1 tende a ser vizinho de uma proteína

sempre que p2 é e vice-versa. Usando a Equação 3 e a terminologia

introduzido para a transformação do valor p, h-confiança para

um par de proteínas é dado pela seguinte equação:

hconf ({p1, p2}) = min m

n1

,

m

n2 «(4)

Usando a confiança h de duas proteínas como o peso da borda

entre as proteínas, um novo gráfico pode ser criado. No entanto, em

Além de um limiar de confiança h, também é necessário

em conta o número absoluto de vezes que duas proteínas aparecem juntas

(m) bem como a fração de vezes que a ocorrência de

uma proteína como vizinha implica a ocorrência da outra proteína

como vizinho (min (n1 / m, n2 / m)). Por exemplo, se as proteínas

p1 e p2 ambos têm apenas uma borda, que é a proteína p3, então

sua confiança h será uma. Por outro lado, as proteínas p1 e

p2, pode ter bordas com outras 10 proteínas, das quais 8 compartilham.

Eles terão uma confiança h de 0,8 e, assim, parece não ser tão

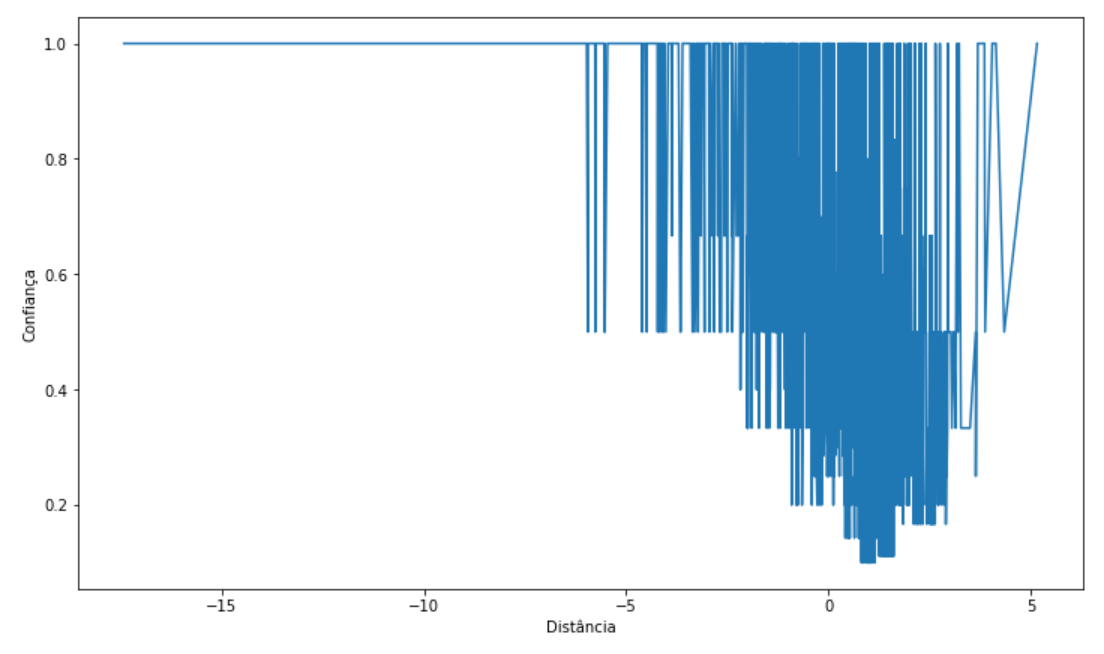
fortemente conectado como o primeiro par. Para lidar com esses problemas,

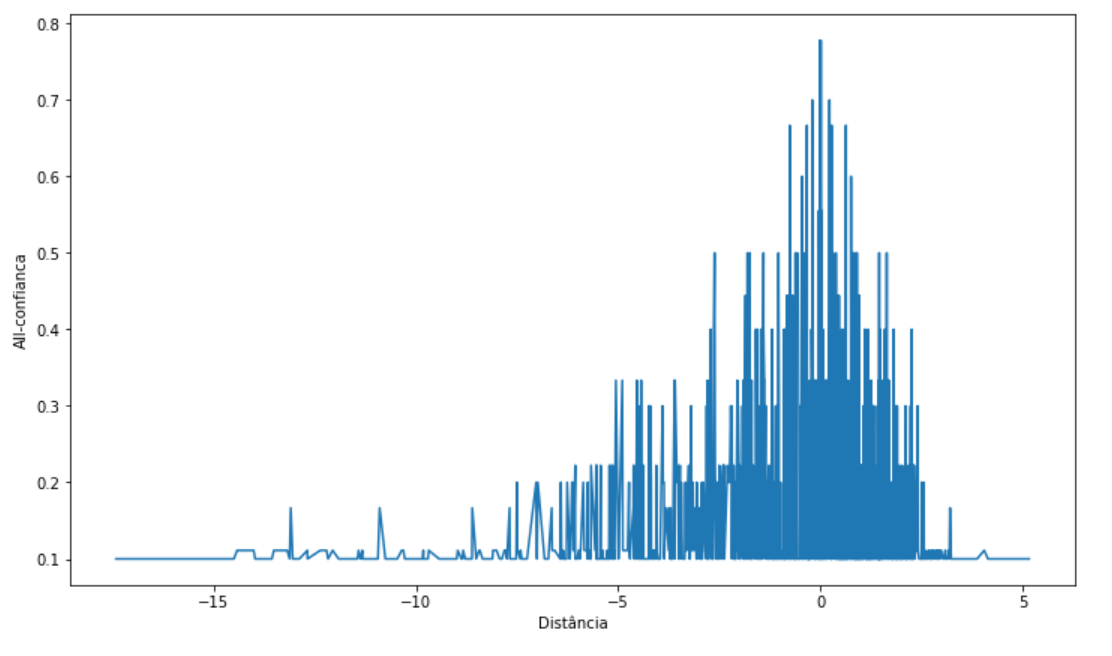
É comum definir um limite de suporte, ou seja, exigir que m tenha

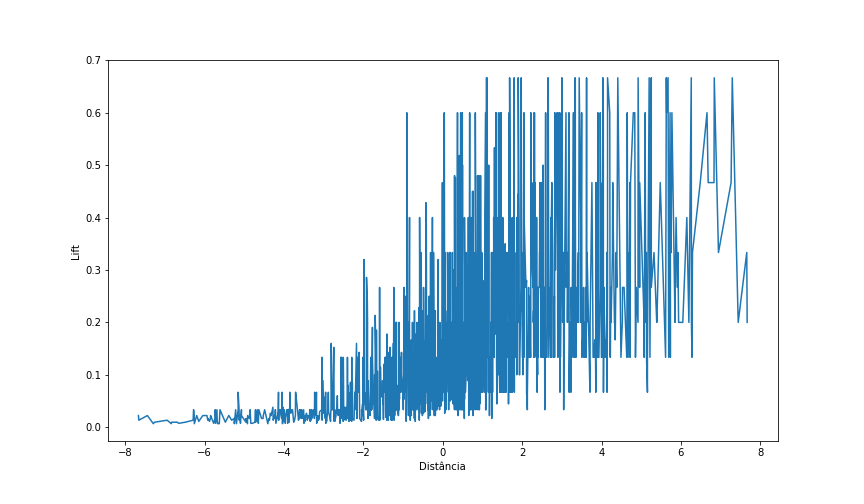
menos algum valor especificado.

## META INFORMAÇÃO

Assim que concluímos os diferentes cálculos de distância passamos a ter condições de trabalhar com esses metadados de tal modo, por fim, sermos capazes de prever comportamentos







# TESTES

Neste capítulo aplicaremos o projeto em bases de dados reais (1) Led 7 Data, () Page B data, retiradas do repositório da UCI<> para ratificar a eficácia da ferramenta.

## LED7 DATA

## pAGE b DATA

# CONCLUSÕES

Neste projeto apresentamos a importância das principais métricas das regras de associação na verificação de padrões relacionais complexos, a criação de submatrizes representativas das principais métricas das regras de associação, a implementação do cálculo da distancia entre os pontos médios das regras e suas origens, o estudo a respeito do relacionamento das distancias entre os antecedentes e os consequentes das regras abordadas, e foi feita a projeção das transações no espaço solução através do algoritmo *Dual Scaling*.

Seguem como desafios futuros o estudo dos pontos médios entre as transações e suas origens, a geração dos gráficos destas transações, e as conjecturas sobre as projeções destas transações no espaço-solução.

# BIBLIOGRAFIA

AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A. **Mining Association Rules between Sets of Items in Large Databases**. ACM SIGMOD Conference. Washington DC: [s.n.]. may 1993. p. 10.

LÍVIA MARIA ROCHA DE VASCONCELOS, C. L. D. C. Fractal Telemática. **Telemática Fractal.com.br**, 29 abril 2018. Disponivel em: <https://telematicafractal.com.br/revista/index.php/telfract/article/view/8/5>. Acesso em: 28 abril 2019.

NOMELINI, J. et al. Emprego de regras de associação para extração de padrões mercadológicos de touros Nelore com avaliação genética. **Revista Brasileira de Zootecnia**, São Paulo, v. 39, n. 12, p. 8, dezembro 2010. ISSN 1806-9290. Disponivel em: <http://dx.doi.org/10.1590/S1516-35982010001200011>. Acesso em: 9 maio 2019.

RIBEIRO, M. X. Digital Library Usp. **Suporte a sistemas de auxílio ao diagnóstico e de recuperação de imagens por conteúdo usando mineração de regraas de associação**, 17 novembro 2008. Disponivel em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-16022009-144432/en.php>. Acesso em: 09 maio 2019.

SILVA, G. C. Repositório Digital. **Mineração de regras de associação aplicada a dados da Secretaria Municipal de Saúde de Londrina PR**, 2004. Disponivel em: <https://www.lume.ufrgs.br/handle/10183/8696>. Acesso em: 05 maio 2019.