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
2019

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.

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 tem iluminado nossa caminhada, que tem nos carregado no colo nos momentos de maior dor e aflição. Obrigado Senhor!

A família primeira, especialmente à minha mãe, Acidinéa da Costa Sampaio (*in memoriam*), por ter sido, durante sua maravilhosa vida, àquela que carinhosa e incansavelmente colocou-me no caminho acadêmico, abraçando-me a cada vitória e confortando-me a cada decepção, que Deus a tenha em Sua Glória. À minha irmã Rosana, por tudo pelo que tem feito pela nossa família, pelo apoio incondicional e amor infinito. Ao meu pai, Telmo, que mesmo acamado, sempre me incentiva com seu sorriso “maroto”.

Ao nosso orientador, Altobelli de Brito Mantuan, pelo apoio incondicional, pelo profissionalismo fantástico, pelos inúmeros estímulos e a grande atenção sem os quais não teríamos atingido o objetivo fim deste trabalho.

Ao grande amigo e parceiro Alex Sandro pela a amizade fiel, pelo incentivo e troca de experiências ricas e constantes.

A família que constituí, primeiramente à minha esposa, por apoiar-me e fortalecer-me a cada momento de angústia, sendo a companheira de todas as horas. Ao meu grande amigo e filho, Pedro Henrique, pela amizade incondicional, pelo companheirismo de sempre, pela ajuda nos inúmeros cálculos e textos que com ele ao longo da caminhada tive o prazer de realizar.

(**Alex**)
A Deus por ter me dado forças para esta caminhada.

Ao nosso orientador Altobelli de Brito Mantuan, pelo apoio e dedicação nas orientações a respeito trabalho.

Aos meus pais João Batista de Souza (*in memoriam)*, e minha mãe Celina da Silva de Souza, por todo apoio, dedicação e incentivo incondicionais em minhas jornadas pessoais, culturais e profissionais.

Ao Amigo José Ricardo Sampaio pela parceria e trocas de conhecimento durante essa jornada do curso.

Ao CEDERJ por me proporcionar a amizade do José Ricardo Sampaio.

Ao mais do que amigo de todas, ao irmão mais velho que a vida me presenteou José Américo Barroso Soares, aquele que me deu suporte incondicional no momento mais difícil da minha vida.

RESUMO

Com a chegada das novas tecnologias e o aumento consequente do volume de informações geradas diariamente, à realidade contemporânea impõem-se a necessidade de cuidar e tratar desses dados. Tal demanda requer não só um maior número de profissionais capacitados, como também a produção de ferramentas eficazes no fazer exigido. Ferramentas que propiciem a tomada de decisões são cada vez mais procuradas e vimos nisto uma grande oportunidade. Este trabalho, que não se encerra em si, 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 (DB)), 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 os 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 de métodos dentre os quais, merecem menção, o algoritmo *Dual Scaling* que tem sua utilidade aplicada ao mapeamento da BD, e a técnica do 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, e dentro de qualquer cenário que se enquadre na proposta apresentada.

Palavras-chaves: Dual Scaling e Regras de Associação.

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

1 INTRODUÇÃO

A velocidade da evolução tecnológica e a crescente utilização de bancos de dados para as mais diversas finalidades, somadas à necessidade de conhecimento mais abrangente e eficaz acerca das relações desses dados transacionais, torna-se imprescindível a utilização dos conceitos de *data mining* – mineração de dados (DM), também conhecido por *Knowledge Discovery in Databases* (KDD), esse processo, que vem sendo cada vez mais utilizado, envolve encontrar e interpretar padrões nos dados através da execução de algoritmos e da análise de seus resultados. As principais etapas do KDD encontram-se ilustradas na Figura 1 (RIBEIRO, 2008, p. 53).

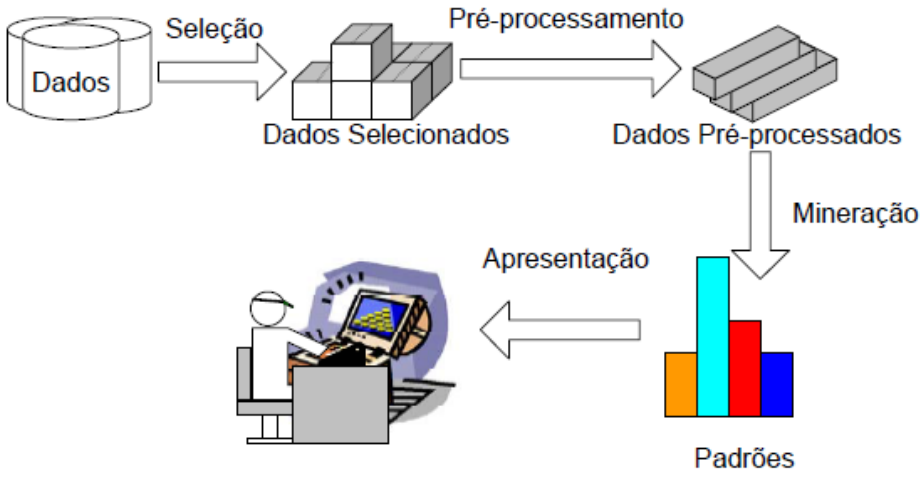

Figura 1: Etapas do Processo de KDD

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 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/CE-DERJ_2019_ALEX_SOUZA_E_JOSE_SAMPAIO.git

2 TRABALHOS RELACIONADOS

Neste capítulo serão citados alguns trabalhos que utilizam este algoritmo de identificação de padrões chamado regras de associação (RA), 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, p. 72), 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, p. 8). 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 fundamental para a comercialização dos touros no país, identificando combinações de atributos genéticos, 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, p. 102). 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 de naturezas distintas.

3 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão apresentados os conceitos aplicados ao estudo das RA, e o comportamento dos padrões relacionais complexos verificados entre atributos de uma base de dados estudada. Estes padrões são obtidos através das técnicas contidas no processo do KDD.

As RA, que são o foco do nosso projeto, encontram-se na principal etapa do processo de KDD. – a mineração dos dados com objetivo de descrever padrões de relacionamentos complexos entre os itens da base de dados estudada.

Cabe ressaltar, que a obtenção destas RA visa ajudar as pessoas responsáveis pela administração da base de dados na tomada de decisões, visto que esta ferramenta proporciona uma visão mais abrangente acerca destes relacionamentos totalmente desconhecidos do ponto de vista administrativo.

3.1 REGRAS DE ASSOCIAÇÃO

Formada por milhares de itens armazenados em uma grande base de dados, é crescente a 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 (RAKESH, AGRAWAL, *et al.*, 1993), as RA devem atender as métricas de suporte e confiança mínimos propostos na pesquisa feita na base de dados. Seja I ={i_1,...,i_n} um conjunto de literais, denominados itens. Qualquer conjunto $X \in I$ 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 $X \in t$. 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:

$$sup(X \rightarrow Y) = \frac{|X \cup Y|}{|R|} \quad (1)$$

$$conf(X \rightarrow Y) = \frac{|X \cup Y|}{|X|} \quad (2)$$

O problema da obtenção das RA, 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 $X \rightarrow Y$, onde $X \cap Y = \emptyset$, 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 $X \rightarrow Y$ 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 (TAN, N, *et al.*, 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 na base de dados.

3.2 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, 2018, p. 21)

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 DS procura 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 DB, e Erro! Fonte de referência não encontrada. será nossa matriz de padrão de respostas, baseada na tabela de padrão de respostas de 0s e 1s, de tamanho $n \times m$, 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 DS tem por objetivo descobrir a quantidade de dimensões do espaço-solução (n_s), através da seguinte equação:

$$n_s = m - q - 1, \tag{3}$$

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

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

$$fr_i = \sum_{k}^{m} F_{i,k} \tag{4}$$

$$fc_j = \sum_{k}^{n} F_{k,j} \tag{5}$$

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

$$\mathrm{D}r_{i,q} = \begin{cases} fr_i & i = q \\ 0 & i \neq q \end{cases} \tag{6}$$

$$\mathrm{D}c_{r,j} = \begin{cases} fc_r & r = j \\ 0 & r \neq j \end{cases} \tag{7}$$

O próximo passo é definir as correlações entre colunas da matriz F, cujo resultado chamaremos de matriz $M_{m,m}$, dada pelo resultado da equação:

$$M = F^T \; {D_r}^{-1} \; F \; {D_c}^{-1} \tag{8}$$

A transposição de matriz é representada por F^T, enquanto a inversão de matriz é representada por ${D_c}^{-1}$.

Representado por $\lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_m\}$ e $V_{m,m}$, respectivamente, o vetor de autovalores e a matriz de autovetores de M. O vetor de autovalores λ deve ser ordenado, de tal forma que $1 = \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$. As colunas da matriz de autovetores

V devem acompanhar a ordenação de seus respectivos autovalores. Uma vez ordenados, o primeiro item do vetor de autovalores λ, bem como a primeira coluna da matriz de autovetores V, devem ser descartados; e o número máximo de elementos em λ e colunas em V devem ser iguais ao n_s. Logo, temos $\lambda f = \{\lambda_2, \lambda_3, \dots, \lambda_{n_s+1}\}$ como o vetor final de autovalores, e Vf_{m,n_s} como a matriz final de autovetores.

Em seguida, vamos calcular a matriz T_{m,n_s}, dada pela equação:

$$T = Dc(Vf \circ Vf) \tag{9}$$

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

$$tc_o = \sum_k^m T_{k,o} \tag{10}$$

A seguir no cálculo do *DS*, é calculado o vetor Cc_{n_s}, 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 Cc pode ser definido pela equação:

$$Cc_p = \sqrt{\frac{ft}{tc_p}}, \tag{11}$$

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

$$Nx_{i,p} = Vf_{i,p}\ Cc_p. \tag{12}$$

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 Px e obtida pela equação:

$$Px_{i,p} = Nx_{i,p} \ \rho_p \ , \tag{13}$$

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

$$\rho_i = \sqrt{\lambda f_i} \ . \tag{14}$$

Os resultados obtidos através dos cálculos dos algoritmos do DS, permitem verificação das características apresentadas pelas RA verificadas nas pesquisas feitas na 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;

3.3 CÁLCULO DAS DISTÂNCIAS

As distâncias podem ser classificadas de duas formas:

- Distância intragrupo (*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;
- Distância intergrupo (*between-set distance*), que são as distâncias dos itens para as transações, e vice-versa.

A verificação da distância quadrada entre os itens em um espaço-solução de n dimensões é alcançada através da métrica chi-quadrado (15).

$$d_{i,i'}^2 = \sum_{k=1}^{n_s} \rho_k \left(\left(\frac{Px_{i,k}}{\sqrt{\frac{fc_i}{n}}} - \frac{Px_{i',k}}{\sqrt{\frac{fc_{i'}}{n}}} \right)^2 \right), \quad (15)$$

onde $Px_{i,k}$ e $Px_{i',k}$ são as k-ésimas coordenadas dos itens indexados por i e i', respectivamente, fc_i e $fc_{i'}$ são os k-ésimos índices do vetor de frequência de colunas e n é o número de linhas da matriz de padrão de respostas F.

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 (15), vamos exemplificar o cálculo da distância dos itens 3 (pressão alta) e 9 (idoso). Temos então:

$$d_{3,9}^2 = \sum_{k=1}^{11} \rho_k \left(\left(\frac{Px_{3,k}}{\sqrt{\frac{fc_3}{n}}} - \frac{Px_{9,k}}{\sqrt{\frac{fc_9}{n}}} \right)^2 \right).$$

Quebrando a equação nos valores de k, demostraremos os cálculos de forma detalhada para $k = 1$, e apenas os resultados para os demais valores de k. Com isso, temos:

$$k = 1 \rightarrow \rho_1 \left(\left(\frac{Px_{3,1}}{\sqrt{\frac{fc_3}{n}}} - \frac{Px_{9,1}}{\sqrt{\frac{fc_9}{n}}} \right)^2 \right) = 0{,}7376 \left(\left(\frac{-0{,}6268}{\sqrt{\frac{4}{15}}} - \frac{-0{,}6602}{\sqrt{\frac{6}{15}}} \right)^2 \right)$$

$$= 0{,}7376 \left(\left(\frac{-0{,}6268}{0{,}5164} - \frac{-0{,}6602}{0{,}6324} \right)^2 \right) = 0{,}7376((-1{,}2138 - (-1{,}0439))^2)$$

$$= 0{,}7376((-0{,}1699)^2) = 0{,}7376 \times 0{,}0289 \cong 0{,}212$$

$k = 2 \cong 0{,}0058$

$k = 3 \cong 0{,}0009$

$k = 4 \cong 0{,}1874$

$k = 5 \cong 0{,}0003$

$k = 6 \cong 0{,}3049$

$k = 7 \cong 0{,}3160$

$k = 8 \cong 0{,}1474$

$k = 9 \cong 0{,}0109$

$$k = 10 \cong 0{,}0208$$

$$k = 11 \cong 0{,}0120$$

$$\sum k = \{0{,}0212 + 0{,}0058 + 0{,}0009 + \ 0{,}1874 + 0{,}0003 + 0{,}3049 + 0{,}3160 + 0{,}1474 + 0{,}0109 + 0{,}0208 + 0{,}0120\} \cong 1.0276$$

A distância quadrada final então entre os itens 3 e 9 é $d_{3,9}^2 \cong 1.03$.

4 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 acerca 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.

4.1 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 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.

4.2 VERIFICAÇÃO DAS RELAÇÕES DE DISTÂNCIAS

Tendo 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 (16), onde Pm é o ponto desejado, u é o número de itens de determinada RA e $vetor[k]$ é o vetor de coordenadas para cada item presente na mesma.

$$Pm = \left(\sum_{k=1}^{u} vetor[k]\right)/u \qquad (16)$$

Tomemos como exemplo a regra da linha 272 do arquivo bloodpressurenishisato-book.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 (16) teremos:

$$\left(\sum_{k=1}^{2} vetor[k]\right)/u = \frac{(vetor[1] + vetor[2])}{2}$$

Onde:

$vetor[1] = 0.66 \;\; -0.69 \;\; 1.11 \;\; 0.26 \;\; -0.26 \;\; 0.09 \;\; -0.37 \;\; -0.15 \;\; -0.04 \;\; 0.13 \;\; -0.18$

$vetor[2] = 0.51 \;\; -0.65 \;\; -0.05 \;\; -0.59 \;\; 0.53 \;\; 0.86 \;\; -0.01 \;\; 0.03 \;\; -0.23 \;\; -0.07 \;\; 0.03$

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

$$Pm\_a = 0.58 \ -0.67 \ 0.53 \ -0.17 \ 0.13 \ 0.48 \ -0.19 \ -0.06 \ -0.13 \ 0.03 \ -0.07$$

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

De posse destes valores, já é possível calcular as distâncias entre os pontos médios de antecedentes e consequentes ($DpmA\_C$), bem como as distâncias de cada um dos pontos médios de antecedentes ($DpmA\_O$) e consequentes ($DpmC\_O$) 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, p. 40).

Ocorre que, como vimos na Seção 3.2, 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 $Px_{i',k}$ assim como a expressão$\left(\frac{Px_{i',k}}{\sqrt{\frac{fc_{i'}}{n}}}\right)$, assumirem valor idêntico à origem, deixando a fórmula número (15) com a seguinte grafia:

$$d_{i,0}^2 = \sum_{k=1}^{n_s} \rho_k \left[\left(\frac{Px_{i,k}}{\sqrt{\frac{fc_i}{n}}}\right)^2\right]$$

Onde, $\left(\frac{Px_{i,k}}{\sqrt{\frac{fc_i}{n}}}\right)$ 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 (fc_{Pm_a}) é 3 enquanto que o dos consequentes (fc_{Pm_c}) 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 $DpmA\_C$.

Pormenorizando o cálculo de $\mathrm{DpmA\_C}$ para o exemplo tomado, veremos que:

$$DpmA\_C = d^2_{Pm_a,Pm_c} = \sum_{k=1}^{11} \rho_k \left[\left(\frac{Px_{Pm_a,k}}{\sqrt{\frac{fc_{Pm_a}}{n}}} - \frac{Px_{Pm_c,k}}{\sqrt{\frac{fc_{Pm_c}}{n}}} \right)^2 \right]$$

Efetuando o cálculo aproximado para cada valor de k teremos:

$$k = 1 \rightarrow DpmA\_C_1 = 0{,}7376 \times \left[\left(\frac{0{,}5839}{\sqrt{\frac{3}{15}}} - \frac{0{,}9533}{\sqrt{\frac{7}{15}}} \right)^2 \right] = 0{,}7376 \left[\left(\frac{0{,}5839}{0{,}4472} - \frac{0{,}9533}{0{,}6831} \right)^2 \right]$$

$$= 0{,}7376 \times \left[\left(\frac{0{,}5839}{0{,}4472} - \frac{0{,}9533}{0{,}6831} \right)^2 \right] = 0{,}7376[(1{,}3057 - 1{,}3954)^2] = 0{,}7376 \times 0{,}0081$$

E assim vem que para $k = 1 \rightarrow DpmA\_C_1 \cong 0{,}0059$

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

$$DpmA\_C = 2.5024$$

Exemplificando agora o cálculo da distância do ponto médio antecedente até a origem ($DpmA\_O$), que se dará através da equação (15), e tomando uma vez mais a RA acima escolhida constataremos:

Para $k = 1 \rightarrow DpmA\_O_1 = \rho_1 \left[\left(\frac{Px_{Pm_a,1}}{\sqrt{\frac{fc_{Pm_a}}{n}}} \right)^2 \right]$

$$k = 1 \rightarrow DpmA\_O_1 = 0{,}7376 \times \left[\left(\frac{0{,}5839}{\sqrt{\frac{3}{15}}} \right)^2 \right] = 0{,}7376 \times \left[\left(\frac{0{,}5839}{0{,}4472} \right)^2 \right]$$

$$k = 1 \rightarrow DpmA\_O_1 \cong 1{,}2574$$

Replicando o procedimento para os demais valores de k, o valor final para a distância assim calculada será $DpmA\_O \cong 4{,}0251$.
Usar-se-ão estas distâncias para, em função de critério predefinido, montarmos as visualizações objeto deste trabalho.

4.3 MÉTRICAS DE RELEVÂNCIA EM REGRAS DE ASSOCIAÇÃO

Entramos no estágio de finalização dos cálculos visando a produção das visualizações, e para tal faremos uso da métrica *Lift*, outro índice estatístico empregado para definir o grau de interesse de uma regra de associação. O Lift tem como tratamento matemático a equação (17), mostrando que seu valor indica, para a RA aplicada, o quão mais frequente torna-se o antecedente (X), quando o consequente (Y) ocorre.

$$lift\ (X\ \rightarrow Y) = \frac{\sup(\mathrm{X}\ /\ \mathrm{Y})}{\sup(X) * \sup(Y)}\ = \frac{P(X \cup Y)}{P(X) * P(X)} \qquad (17)$$

Para esta métrica, valores iguais a 1 indicarão que a RA não é interessante, de tal modo que quanto maior o valor do *Lift* mais interessante será a regra.

A *H-confidence* também conhecida como *all-confidence* (H_{conf}), é uma medida que reflete a correlação global entre os itens dentro de um determinado itemset, aplicada quase que exclusivamente a atributos binários, sendo calculada através da fórmula:

$$H_{conf} = \frac{sup\ (i_1, i_2, \ldots i_k)}{max[sup(i_1)\,, sup(i_2)\,, \ldots, sup(i_k)]} \qquad (18)$$

Onde i_k é o *k*-ésimo item do itemset, *sup* é o suporte e *max* retorna o maior dentre os valores informados. O resultado de H_{conf} encontra-se entre 0 e 1, e valores próximos a 0 indicam que os itens do *itemset* escolhido são pouco correlacionados ao passo que valores próximos a 1 mostram um forte correlacionamento.

4.4 COPORTAMENTO DAS RA EM RELAÇÃO AS DISTÂNCIAS

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 da base de dados observada. Passemos então a análise gráfica.

Podemos observar no Gráfico 1 de acordo com os algoritmos de medição aplicados DS na seção 3.3, que os pontos médios mais distantes são fortes regras e possuem alto valor de $minConf$, enquanto que os valores mais próximos são regras mais fracas tem valor de $minConf\ \ mais$ baixo.

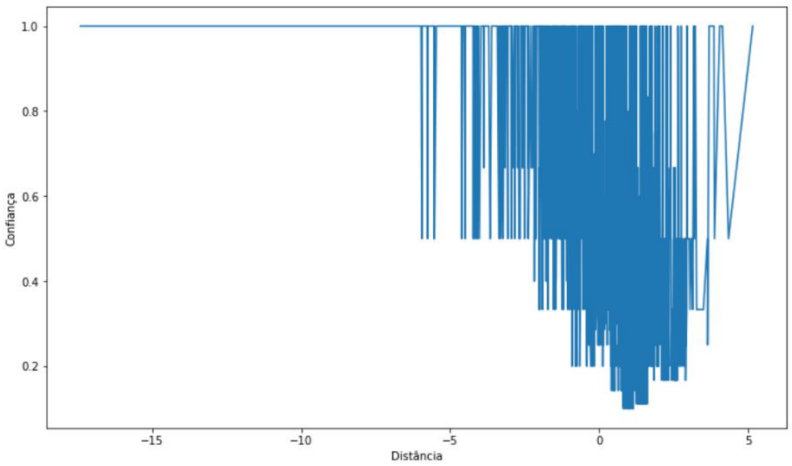

Gráfico 1: Confiança da Base Blood Pressure

A métrica H_{conf} apresentada no Gráfico 2 indica que as RA fortes são as regras de mais próximas da base.

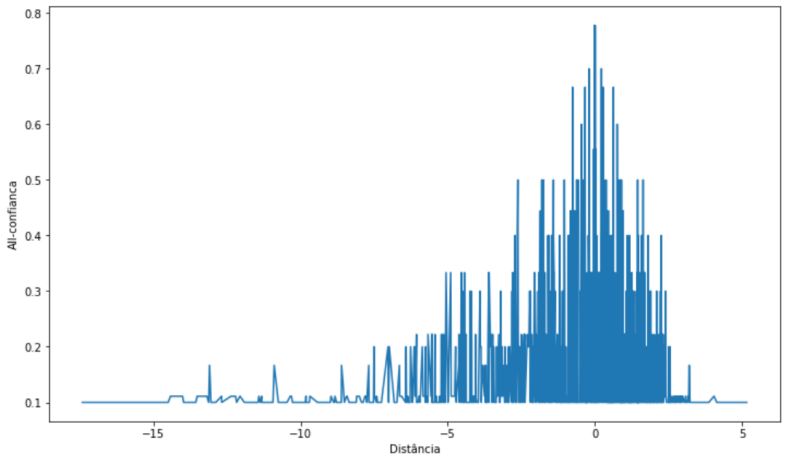

Gráfico 2: H_{conf} da Base Blood Pressure

Enquanto que na métrica *Lift* (Gráfico 3) à medida que aumenta a distância dos pontos médios para origem a força da regra é maior, e os pontos médios muito próximos possuem valor de Lift muito baixo.

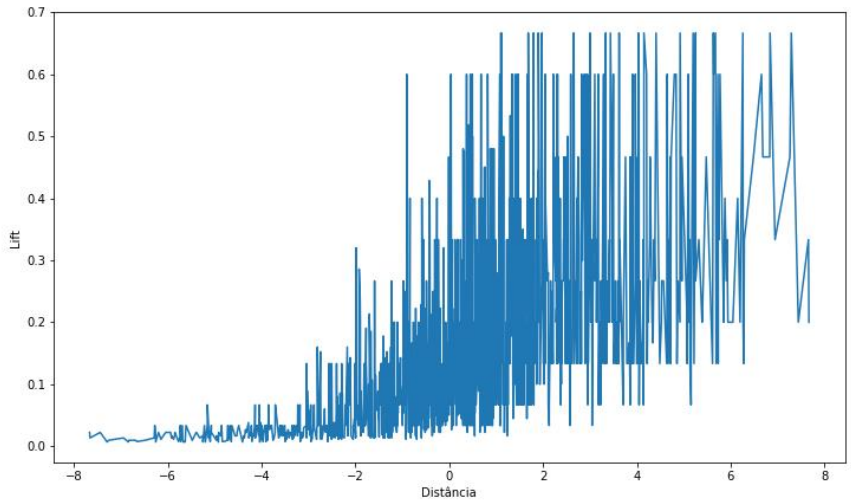

Gráfico 3: *Lift* da Base Blood Pressure

5 TESTES

Neste capítulo aplicaremos o projeto em bases de dados conhecidas por *Led7 Display Domain* e *Page Blocks*, que representam cenários distintos para análise gráfica da ferramenta, e foram obtidas a partir do repositório da UCI [1].

Tabela 2: Bases de Dados usadas

Bases de Dados	Quantidade de itens	Nº de transações
Led7 Display Domain	7	3200
Page Blocks Classification	10	5473

Nesta sessão são apresentados experimentos do algoritmo proposto implementado em *Phyton*. Estes experimentos foram executados em máquina com CPU Core I7 4.0GHz e 16GB de memória RAM rodando o sistema operacional Windows 8 64-bits.

5.1 LED7 DISPLAY DOMAIN

Apresentada como uma base de dados de conjunto de dígitos decimais, a *Led7 displays domain* contém a representação dos 7 diodos emissores de luz, que seria um problema de fácil resolução se não fosse pela interferência de ruídos.

Os valores de atributo são 0 ou 1 de acordo com a luz correspondente ser representada ou não pelo dígito decimal, logo, cada valor de atributo tem probabilidade de 10% de ter seu valor invertido, exceto o atributo *class* que é representado por um inteiro entre 0 e 9 inclusive. Cabe ressaltar que a taxa de classificação incorreta é de 6% levando em consideração a taxa ideal de Bayes, ou seja, precisão de classificação de 74%.

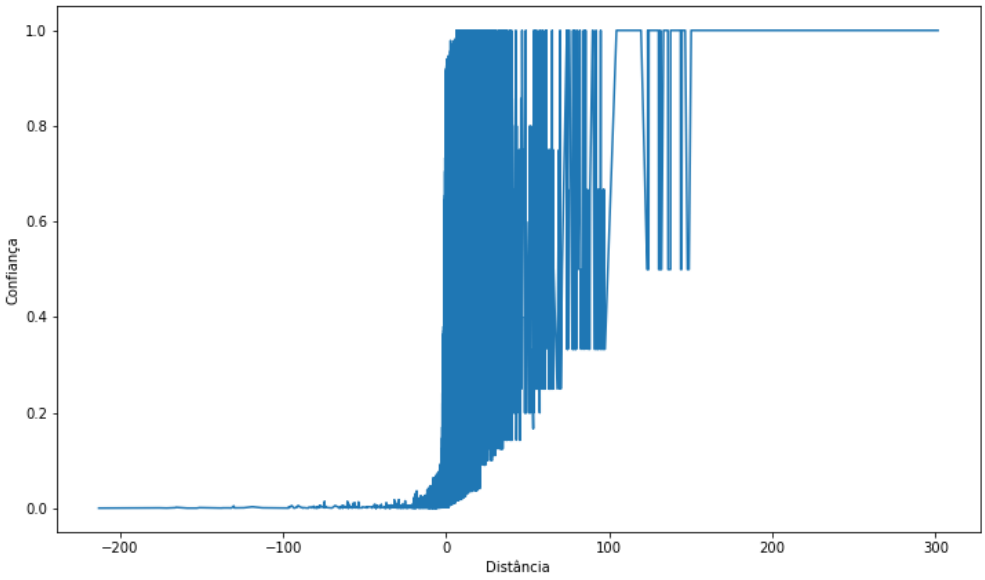

Gráfico 4: Confiança da Base Led7 Display Domain

Como podemos observado na seção 4.4, o Gráfico 4 apresenta maior grau de confiança para pontos distantes da origem da base *LED7 DISPLAY DOMAIN .*

O Gráfico 5 da métrica H_{conf} apresenta apenas pontos próximos a origem como regras fortes, ou de interesse.

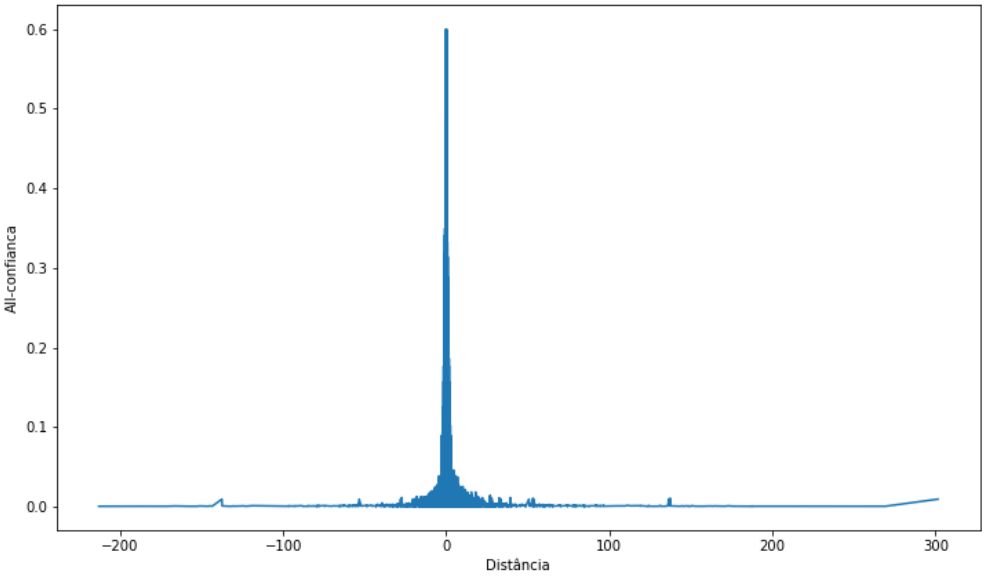

Gráfico 5: H_{conf} LED7 DISPLAY DOMAIN

Enquanto que na métrica *Lift*, assim como no estudo apresentado na seção 4.4, os itens de maior força são os mais distantes da origem conforme Gráfico 6.

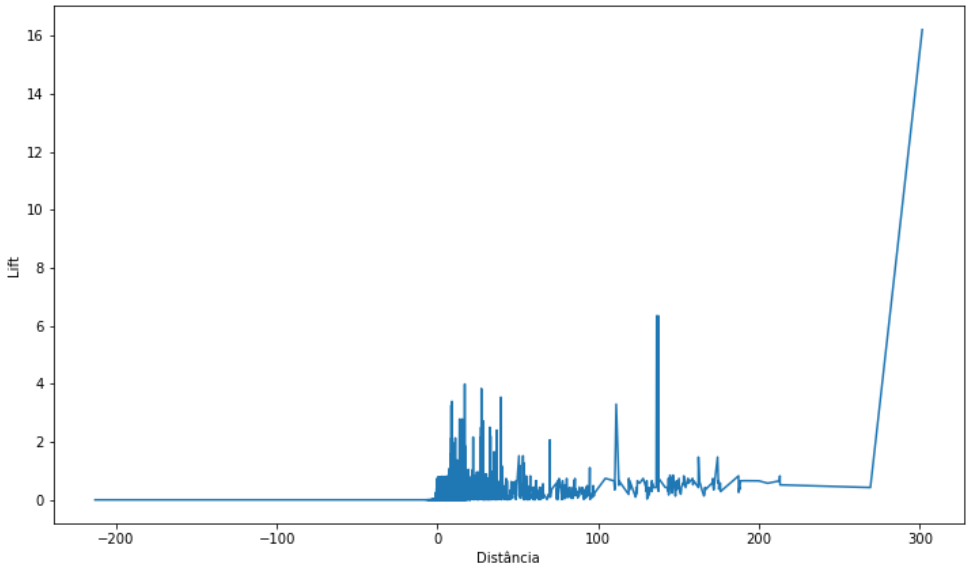

Gráfico 6: *Lift* da Base LED7 DISPLAY DOMAIN

5.2 PAGE BLOCKS CLASSIFICATION

Esta base de dados é uma classificação de blocos de layout de paginas de algum documento detectado por um processo de segmentação, tem 10 atributos e 5473 transações vindas de 54 documentos distintos e representantes de um bloco.

Como apresentado no Gráfico 7 a força das RA mais próximas da origem são as regras de $minconf$ e de maior força de acordo com as técnicas apresentadas na seção 4.4.

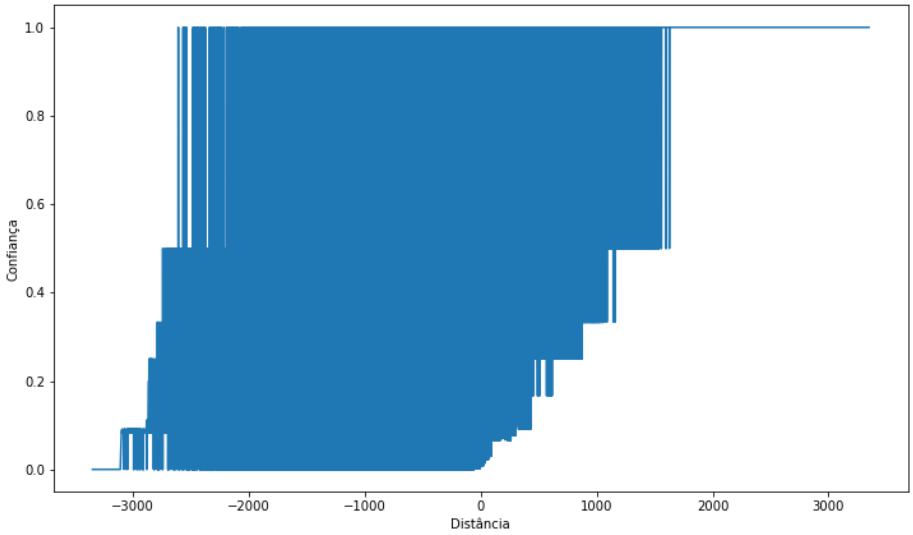

Gráfico 7:Confiança da base PAGE BLOCKS CLASSIFICATION

No Gráfico 8 é facilmente verificado pela métrica H_{conf} que os itens mais próximos representam as regras mais interessantes, e quanto maior a distância sua força tende a 0.

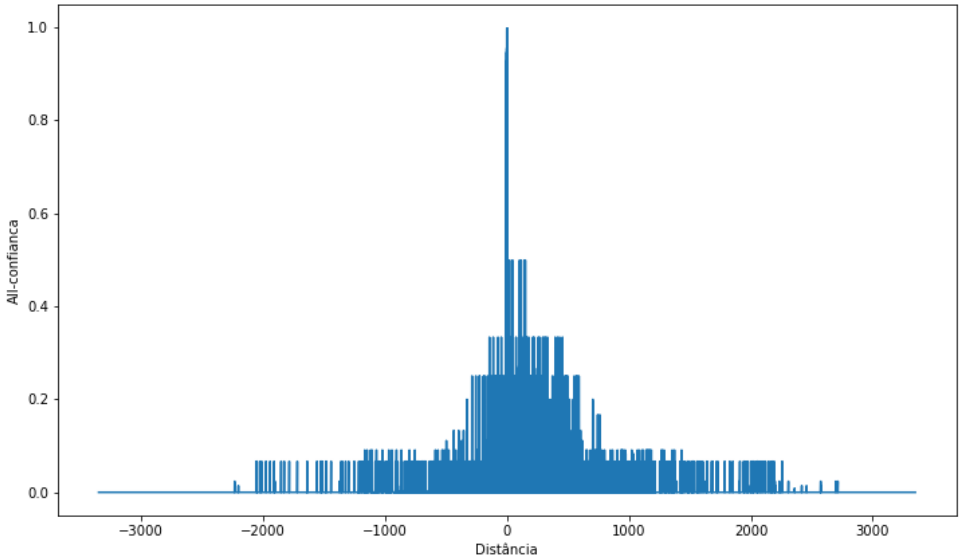

Gráfico 8: H_{conf} PAGE BLOCKS CLASSIFICATION

A métrica de *Lift* apresentada no Gráfico 9 apresenta as regras de maior força entre os pontos mais distantes da origem.

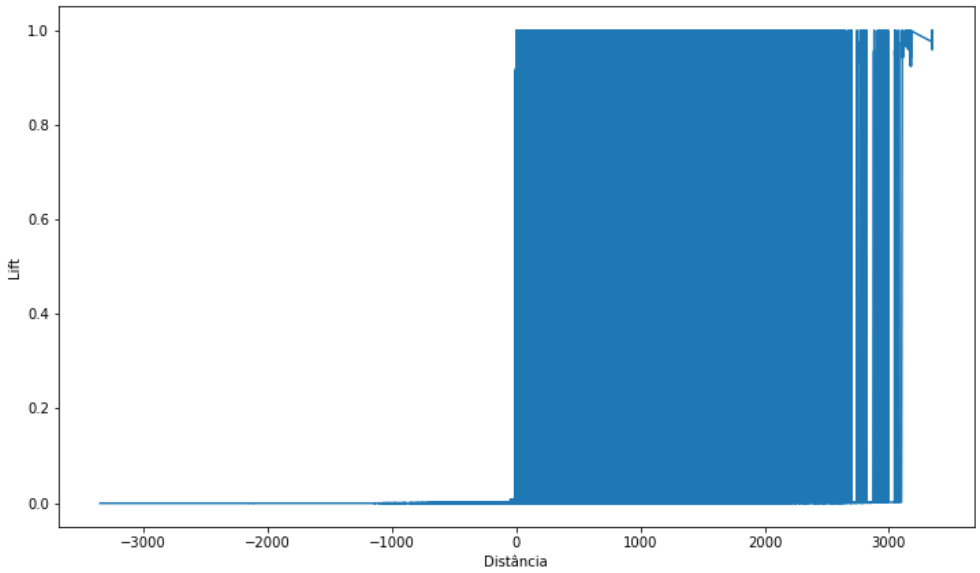

Gráfico 9: *Lift* da Base PAGE BLOCKS CLASSIFICATION

6 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 distância entre os pontos médios das regras e suas origens, o estudo a respeito do relacionamento das distâncias entre os antecedentes e os consequentes das regras abordadas, e foi feita a projeção das transações no espaço solução através do algoritmo *Dual Scaling*.

A utilização da ferramenta desenvolvida em bases reais possibilitou a verificação da eficácia da ferramenta na representação das características intragrupos das bases de dados estudadas.

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.

7 BIBLIOGRAFIA

1. DUA, D.; GRAFF, C. UCI Machine Learning Repository. **University of California, School of Information and Computer Science**, 2019.
2. EVANGELISTA, D. F. M. UFS. **Repositório Institucional UFS**, 25 setembro 2017. Disponivel em: <https://ri.ufs.br/jspui/handle/riufs/7163>. Acesso em: 05 maio 2019.
3. FORTES, J. L. S. Dual scaling viewer: uma ferramenta visual para interpretação comportamental de uma base de dados. **RIUFF - Repósitório Institucional da UFF**, 2018. ISSN [1]. Disponivel em: <https://app.uff.br/riuff/handle/1/8895>. Acesso em: 15 maio 2019.
4. MARCO AURÉLIO DOMINGUES, S. O. R. InfoComp. **InfoComp**, 1 junho 2005. Disponivel em: <http://www.dcc.ufla.br/infocomp/index.php/INFOCOMP/article/view/89>. Acesso em: 28 abril 2019.
5. NISHISATO, S. Multidimensional Nonlinear Descripitive Analysis. **https:** //taylorandfrancis.com/, 2007. Disponivel em: <https://taylorandfrancis.com/>. Acesso em: 10 mar. 2019.
6. 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.
7. RAKESH et al. **Mining Association Rules between Sets of Items in Large Databases**. ACM SIGMOD Conference. Washington DC: [s.n.]. may 1993. p. 10.
8. 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.

9. 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.
10. TAN et al. **Introdução ao DATA MINING Mineração de Dados.** Rio de Janeiro: Editora Ciência Moderna Ltda, 2009., 2009.