UNIVERSIDADE FEDERAL FLUMINENSE

CAIO XAVIER CABRAL

RICARDO ALCÂNTARA ANDRADE

IMPLEMENTAÇÃO E ESTUDO DE MÉTRICAS DE SIGNIFICÂNCIA EM REGRAS DE ASSOCIAÇÕES PARA BASES DE DADOS TRANSACIONAIS

Niterói

2019

CAIO XAVIER CABRAL
RICARDO ALCÂNTARA ANDRADE

IMPLEMENTAÇÃO E ESTUDO DE MÉTRICAS DE SIGNIFICÂNCIA EM REGRAS DE ASSOCIAÇÕES PARA BASE DE DADOS TRANSACIONAIS

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

Ficha catalográfica

Ficha catalográfica automática - SDC/BEE
Gerada com informações fornecidas pelo autor

A553i Andrade, Ricardo Alcântra
IMPLEMENTAÇÃO E ESTUDO DE MÉTRICAS DE SIGNIFICÂNCIA EM REGRAS DE ASSOCIAÇÕES PARA BASE DE DADOS TRANSACIONAIS / Ricardo Alcântra Andrade, Caio Xavier Cabral ; Altobelli de Brito Mantuan, orientador. Niterói, 2019.
83 f. : il.

Trabalho de Conclusão de Curso (Graduação em Tecnologia de Sistemas de Computação)-Universidade Federal Fluminense, Instituto de Computação, Niterói, 2019.

1. Regras de associação relacionais. 2. Mineração de dados (Computação). 3. Produção intelectual. I. Cabral, Caio Xavier. II. Mantuan, Altobelli de Brito, orientador. III. Universidade Federal Fluminense. Instituto de Computação. IV. Titulo.

CDD -

Bibliotecária responsável: Fabiana Menezes Santos da Silva - CRB7/5274

CAIO XAVIER CABRAL
RICARDO ALCÂNTARA ANDRADE

IMPLEMENTAÇÃO E ESTUDO DE MÉTRICAS DE SIGNIFICÂNCIA EM REGRAS DE ASSOCIAÇÕES PARA BASE DE DADOS TRANSACIONAIS

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, 06 de Dezembro de 2019.

Banca Examinadora:

_____

Prof. Altobelli de Brito Mantuan, Msc. – Orientador
UFF –Universidade Federal Fluminense

_____

Prof. Leandro Botelho Alves de Miranda, Msc. – Avaliador
UFF –Universidade Federal Fluminense

Dedico este trabalho a minha família.

AGRADECIMENTOS

A Deus, que sempre iluminou a minha caminhada.

A meu Orientador Altobelli de brito Mantuan 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.

“A Escola é uma arena onde grupos sociais lutam por legitimidade e poder”.
Dinair Leal da Hora

RESUMO

A maior parte das empresas utiliza sistemas computação para apoiar suas atividades e com o passar do tempo fica acumulado uma grande massa de dados em suas bases de dados. Esses dados representam o histórico dessas atividades e possuem implicitamente uma variedade de informações ocultas relevantes para essas entidades. Mas como revelar essas informações? Uma das ferramentas de mineração de dados utilizadas para esse propósito são as regras de associação. Com o passar do tempo a quantidade de dados acumulados crescem exponencialmente o que torna necessário uma evolução contínua dos processos de mineração para torná-lo cada vez mais eficiente. Ao longo desse processo de evolução um dos estudos que se encontra na literatura com esse propósito é o desenvolvimento de métricas de associação. O objetivo deste trabalho é realizar um estudo sobre métricas de associação, descrevendo suas propriedades e características, implementando um grupo de métricas na linguagem de programação *PYTHON* para posteriormente elaborar uma análise inicial de similaridade entre elas através da "correlação de Pearson" e então agrupá-las em *clusters* de acordo com sua similaridade de comportamentos.

Palavras-chaves: regras de associação, mineração de dados, medidas de interesse, *PYTHON*, correlação de Pearson.

ABSTRACT

Most companies use computing systems to support their activities and over time a large mass of data accumulates in their databases. This data represents the history of these activities and implicitly has a variety of hidden information relevant to these entities. But how to reveal this information? One of the data mining tools used for this purpose is association rules. Over time the amount of accumulated data grows exponentially which necessitates a continual evolution of mining processes to make it increasingly efficient. Throughout this process of evolution one of the studies in the literature for this purpose is the development of association metrics. The aim of this paper is to conduct a study on association metrics, describing their properties and characteristics, implementing a group of metrics in the PYTHON programming language to further elaborate an initial similarity analysis between them using the “Pearson correlation” and then grouping them together. there in clusters according to their similarity of behaviors.

Keywords: association rules, data mining, measures of interest, PYTHON, Pearson correlation.

LISTA DE ILUSTRAÇÕES

Figura 1: Operações em uma tabela de contingência [17].....64

LISTA DE TABELAS

Tabela 1: Base de dados transacional.....18
Tabela 2: Possíveis formações de *Itemsets* através da combinação de itens.....20
Tabela 3: Valor acumulado de possíveis formações de *Itemsets*.....21
Tabela 4: Possíveis candidatos a regras de associação com total de itens m = 4….23
Tabela 5: Relação de valores acumulados de itens, *Itemsets* e candidatos a regras de associação.....24
Tabela 6: Relação de itens presentes nas transações.....26
Tabela 7: Identificação de *Itemsets* mais frequentes em 2-itemset na BD.....26
Tabela 8: Identificação de *Itemsets* mais frequentes em 1-itemset na BD.....27
Tabela 9: Identificação de *Itemsets* mais frequentes em 3-itemset na BD.....27
Tabela 10: Identificação de *Itemsets* mais frequentes em 4-itemset na BD.....28
Tabela 11: Identificação de *Itemsets* mais frequentes em 5-itemset na BD.....29
Tabela 12: Identificação de *Itemsets* mais frequentes em 6-itemset na BD.....29
Tabela 13: Relação de *Itemsets* candidatos a regras de associação.....31
Tabela 14: Regras de associação.....32
Tabela 15: Tabela de Contingência 2x2 para $A \Rightarrow C$.....33
Tabela 16: Propriedades de métricas objetivas para regras de associação.....65

LISTA DE GRÁFICOS

Gráfico 1: Evolução de itens x Candidatos a regras de associação.....24
Gráfico 2: Matriz de correlação.....67
Gráfico 3: Matriz de correlação Clusterizada.....68
Gráfico 4: Correlação entre Suporte da regra e outras métricas.....69

SUMÁRIO

RESUMO....8
ABSTRACT....9
LISTA DE ILUSTRAÇÕES....10
LISTA DE TABELAS....11
LISTA DE GRÁFICOS....12
SUMÁRIO....13
1 INTRODUÇÃO....15
2 FUNDAMENTAÇÃO TEÓRICA....17
2.1 BASE DE DADOS....17
2.2 ITEMSET....19
2.3 REGRA DE ASSOCIAÇÃO....22
2.4 SUPORTE....25
2.5 CONFIANÇA....30
3 MÉTRICAS DE REGRAS DE ASSOCIAÇÃO....33
3.1 ADDED VALUE....35
3.2 ALL-CONFIDENCE....35
3.3 CAUSAL SUPPORT....36
3.4 CAUSAL CONFIDENCE....36
3.5 CERTAINTY FACTOR....37
3.6 CHI-SQUARED....38
3.7 CROSS-SUPPORT RATIO....38
3.8 COLLECTIVE STRENGTH....39
3.9 CONVICTION....40
3.10 COSINE....41
3.11 COVERAGE....41
3.12 DESCRIPTIVE CONFIRMED CONFIDENCE....42
3.13 DIFFERENCE OF CONFIDENCE....42

3.14 EXAMPLE AND COUNTER-EXAMPLE RATE....43
3.15 FISHER'S EXACT TEST....43
3.16 GINI INDEX....44
3.17 HYPER CONFIDENCE....44
3.18 HYPER LIFT....45
3.19 IMBALANCE RATIO....45
3.20 IMPORTANCE....46
3.21 IMPROVEMENT....46
3.22 JACCARD COEFFICIENT....47
3.23 J-MEASURE....47
3.24 KAPPA....48
3.25 KLOSGEN....48
3.26 KULCZYNSKI....49
3.27 GOODMAN-KRUSKAL ( )....50
3.28 LAPLACE CORRECTED CONFIDENCE....50
3.29 LEAST CONTRADICTION....51
3.30 LERMAN SIMILARITY....51
3.31 LEVERAGE....51
3.32 LIFT....52
3.33 MAXCONFIDENCE....53
3.34 MUTUAL INFORMATION....54
3.35 ODDS RATIO....54
3.36 Φ CORRELATION COEFFICIENT....55
3.37 RALAMBONDRAINY MEASURE....56
3.38 RELATIVE LINKAGE DISEQUILIBRIUM....56
3.39 RULE POWER FACTOR....57
3.40 SEBAG-SCHOENAUER MEASURE....57
3.41 VARYING RATES LIAISON....58
3.42 YULE'S Q AND YULE'S Y....58
4 ANÁLISE EXPERIMENTAL DAS MÉTRICAS CALCULADAS....60
5 CONCLUSÕES E TRABALHOS FUTUROS....70
REFERÊNCIAS BIBLIOGRÁFICAS....71
ANEXOS A: RESUMO DAS MÉTRICAS APRESENTADAS....77

1 INTRODUÇÃO

A maioria das entidades públicas ou privadas se utiliza de computadores para apoiar suas atividades realizando assim vários registros em seu banco de dados. Esses registros de dados representam históricos das atividades dessas organizações e implicitamente contém informações de muita relevância para que essas entidades possam evoluir seus processos de forma a torná-los mais eficientes ou simplesmente identificar algum conhecimento presente nessa base de dados. Ao longo do tempo o conjunto de dados registrados fica muito extenso tanto pela quantidade de transações realizadas quanto pela variedade de atributos envolvidos nessas transações, então tentar extrair informações desse histórico torna-se uma tarefa bastante complexa para realizar através de métodos manuais, planilhas ou pesquisas SQL por exemplo.

Sendo assim surgiu a necessidade de se desenvolver métodos de pesquisa para extrair essas informações dando origem então a área de pesquisa de descoberta de conhecimento a partir de base de dados (*KDD – Knowlegde Discovery in Databases*). A descoberta de conhecimento a partir de base de dados é basicamente realizada através de processos automatizados para analisar grandes bancos de dados com o objetivo de determinar regras e padrões revelando comportamentos frequentes, potencialmente úteis e de interesse relevante para as organizações. Diferentes tipos de informações podem ser obtidos através das técnicas de *KDD* como: hierarquias de classificação, padrões sequenciais e outros como as regras de associação.

As regras de associação são a base deste trabalho e são definidas por revelar todos os padrões de relacionamento entre itens ou atributos, que ocorrem com uma determinada frequência entre eventos correlacionados, pertencentes a uma determinada base de dados. Uma aplicação para a descoberta de um padrão de comportamento em que clientes que compram (evento 1) um determinado produto (item 1) tendem a comprar (evento 2) outro específico (item 2). Isto pode auxiliar determinada empresa a potencializar suas vendas através de ações voltadas a esse comportamento por exemplo.

No entanto, devido ao crescimento exponencial relacionado as bases transacionais, a quantidade de regras de associação mineradas permanece muito extensa, dificultando o trabalho dos especialistas em selecionar ou até identificar quais regras são realmente interessantes dentro do contexto observado. O que identificamos na literatura, ao longo de décadas de desenvolvimento e pesquisa nesta área, é uma diversidade grande de métricas com diferentes formas de desenvolvimento e características, elaboradas com o objetivo de auxiliar na seleção de quais regras são mais interessantes aos seus propósitos.

Neste trabalho temos o objetivo de abordar suas características e propriedades, desenvolvendo o cálculo de um grupo dessas métricas a fim de elucidar e compreender melhor seu funcionamento. Dentre as contribuições deste trabalho podemos listar:

- Seleção de um grupo de métricas de interesse que possuem uma diversidade grande de características e propriedades, demonstrando assim as várias formas de abordar a questão que envolve a obtenção de regras de associação mais interessantes.
- Implementação, na linguagem *PYTHON*, das métricas estudadas.
- Realização de validação das métricas selecionadas através de seu uso em uma base de dados real utilizada na literatura.
- Realização de uma análise inicial de comportamento das métricas selecionadas, utilizado Plot`s com o objetivo de obter conhecimento de correlação entre as métricas selecionadas.

As informações pertinentes a codificação em *PYTHON,* a base de dados testada, as regras mineradas e aos Plot`s realizados durante o trabalho, se encontram no repositório público no *GitHub*, https://github.com/altobellibm/CEDERJ_2019_CAIO_CABRAL_E_RICARDO_ANDRADE.git .

2 FUNDAMENTAÇÃO TEÓRICA

O objetivo desta seção é demonstrar com clareza os aspectos que envolvem o *KDD* através de regras de associação em bancos de dados, definindo seus aspectos matemáticos e exemplificando através de modelos.

A identificação de regras de associação em bancos de dados pode ser subdividida em 3 etapas: (1) A geração de *itemsets* candidatos, (2) a análise e identificação de *itemsets* frequentes, (3) A definição das regras de associação identificadas.

2.1 BASE DE DADOS

Banco de dados são sistemas de persistência de informação (armazenamento de informações de forma permanente) e tem por objetivo armazenar conjuntos de dados (bases de dados) inter-relacionados e organizados de forma a permitir a recuperação desses dados com o objetivo de obter uma informação histórica específica [39]. Bases de dados são os conjuntos de dados que são armazenados nos bancos de dados que não possuem, a princípio, nenhuma visão de correlação associada. São todas as informações relativas as atividades específicas das entidades que as utilizam, mas somente através da análise dessas atividades registradas (estado dos atributos ou variáveis relacionadas) e armazenadas nesses bancos de dados que o *KDD* revelará padrões de comportamentos frequentes que representam um conhecimento relevante e de grande importância para essas organizações.

Os dados armazenados nesses bancos de dados podem ser de vários tipos e representar uma informação qualitativa (expressa uma qualidade da variável ou atributo, ou seja, uma característica não mensurável numericamente, ex: cor → azul) ou quantitativa que ainda pode ser discreta (expressa uma quantidade finita ou enumerável relacionada ao atributo, ex: pneus → 4) ou contínua (expressa uma quanti-

dade infinita de possibilidades dentro de um intervalo ou conjunto união de números Reais relacionada ao atributo, ex: peso → 1 kg).

Os diferentes tipos de informação representadas por esses dados determinam as técnicas que serão utilizadas para a obtenção do conhecimento pretendido na base de dados. O objeto de estudo desse trabalho são regras de associação, sendo assim definimos como exemplo para essa base de dados um exemplo clássico em regras de associação, a análise de cestas de mercado (*Market Basket Analisys*).

Essa base de dados é então assim definida como um conjunto de atributos ou itens $I = \{i_1, i_2, i_3, \ldots, i_m\}$ que podem assumir um valor binário (0 → não está contido, 1 → está contido) e $D = \{T_1, T_2, T_3, \ldots, T_n\}$ um conjunto de transações ou registros presentes nesta base de dados que representam compras desses itens. A **tabela 1** demonstra um exemplo de uma base de dados com um conjunto total de 10 registros de transações ou simplesmente transações e um conjunto de 6 itens.

Tabela 1: Base de dados transacional

Transações	**Lista de Itens *I***					
	Lápis	Caneta	Caderno	Borracha	Apontador	Estojo
1ª	0	1	0	1	0	1
2[2]	1	1	0	0	0	0
3[3]	0	1	1	1	1	1
4ª	0	0	0	0	1	1
5ª	1	1	1	0	0	0
6ª	0	0	1	0	0	1
7ª	1	0	0	1	1	0
8ª	0	0	1	0	0	0
9ª	1	1	1	1	1	1
10ª	0	0	1	1	1	0

2.2 ITEMSET

Em regras de associação existe um conceito que é a base da identificação dos padrões de comportamento buscados em uma base de dados, este conceito é chamado de *itemset.* Um *itemset* é um subconjunto do conjunto de itens pertencentes a base de dados através da qual avaliaremos seu estado e correlações com outros itens nas diversas transações existentes na base de dados. Nossa base de dados exemplo descrita na **tabela 1,** possui um total de 6 atributos, ou itens diferentes (lápis, borracha, caneta, etc…).

Um *itemset* pode possuir de diversas dimensões, limitado ao número total de itens diferentes existentes em nossa base de dados. O total de itens em nossa base é m = 6 itens, então possíveis dimensões de *itemset* em nossa base de dados seriam: 1-*Itemset*, 2-*Itemset*, 3-*Itemset*, 4-*Itemset,* 5-I*temset*, 6-*Itemset*. Generalizando, um *itemset* contendo p dimensões para testes, onde $p \leq m$ é um *p-Itemset*. O total de *itemsets* possível para realizar busca na base de dados é igual ao somatório de todas as possíveis dimensões de *itemsets* existentes em nossa base de dados, ou seja, o somatório de total as possíveis combinações de itens.

Fórmula de combinações de itens → $$C(m,p)=\binom{m}{p}=\frac{m!}{p!(m-p)!} \qquad \textbf{(1)}$$

1-I*temset* = *C* (6,1) = 6! / 1! (6-1)! = 6 x ~~5!~~ / 1 x ~~5!~~ = 6/1 = 6
2-*Itemset* = *C* (6,2) = 6! / 2! (6-2)! = 6 x 5 x~~4!~~ / 2 x ~~4!~~ = 30/2 = 15
3-*Itemset* = *C* (6,3) = 6! / 3! (6-3)! = 6 x 5 x 4 x ~~3!~~ / 3 x 2 x ~~3!~~ = ~~6~~ x 20 / ~~6~~ = 20
4-I*temset* = *C* (6,4) = 6! / 4! (6-4)! = 6 x 5 x~~4!~~ / ~~4!~~ x 2! = 30/2 = 15
5-*Itemset* = *C* (6,5) = 6! / 5! (6-5)! = 6 x ~~5!~~ / ~~5!~~ x 1! = 6/1 = 6
6-*Itemset* = *C* (6,6) = 6! / 6! (6-6)! = ~~6!~~ / ~~6!~~ x 0! = 1/1 = 1

Na **tabela 2** descrevemos todas as possíveis combinações de itens existentes em nossa base de dados de acordo com nosso conjunto *I.*

Tabela 2: Possíveis formações de *Itemsets* através da combinação de itens

Nº	**1-*Itemset***	**2-*Itemset***	**3-*Itemset***	**4-*Itemset***	**5-*Itemset***	**6-*Itemset***
1	Lápis-L	(L,CN)	(L,CN,CD)	(L,CN,CD,B)	(L,CN,CD,B,A)	(L,CN,CD, B,A,E)
2	Caneta-CN	(L,CD)	(L,CN,B)	(L,CN,CD,A)	(L,CN,CD,B,E)	
3	Caderno-CD	(L,B)	(L,CN,A)	(L,CN,CD,E)	(L,CN,CD,A,E)	
4	Borracha-B	(L,A)	(L,CN,E)	(L,CN,B,A)	(L,CN,B,A,E)	
5	Apontador-A	(L,E)	(L,CD,B)	(L,CN,B,E)	(L,CD,B,A,E)	
6	Estojo-E	(CN,CD)	(L,CD,A)	(L,CN,A,E)	(CN,CD,B,A,E)	
7		(CN,B)	(L,CD,E)	(L,CD,B,A)		
8		(CN,A)	(L,B,A)	(L,CD,B,E)		
9		(CN,E)	(L,B,E)	(L,CD,A,E)		
10		(CD,B)	(L,A,E)	(L,B,A,E)		
11		(CD,A)	(CN,CD,B)	(CN,CD,B,A)		
12		(CD,E)	(CN,CD,A)	(CN,CD,B,E)		
13		(B,A)	(CN,CD,E)	(CN,CD,A,E)		
14		(B,E)	(A,B,CN)	(CN,B,A,E)		
15		(A,E)	(E,B,CN)	(CD,B,A,E)		
16			(A,E,CN)			
17			(CD,B,A)			
18			(CD,B,E)			
19			(A,E,CD)			
20			(B,A,E)			
T	6	15	20	15	6	1

O total de *itemsets* possíveis descritos na **tabela 3** é = $\sum_{p=0}^{m} \binom{m}{p} = 2^m$ (**2**)

Tabela 3: Valor acumulado de possíveis formações de *Itemsets*

m	***Itemsets***
1	2
2	4
3	8
4	16
5	32
6	64

Então o total de *itemsets* possíveis em nossa base de dados é igual a 2^6= 64, mas somando todos nossos agrupamentos obtidos na tabela dá um valor de 63. Isto se dá porque não temos interesse no agrupamento (*itemset*) que não possui itens, ou seja, o 1º *itemset*.

0-Itemset = *C* (6,0) = 6! / 0! (6-0)! = ~~6!~~ / 0! x ~~6!~~ = 1/1 = 1 → *itemset* (Ø)

Então a **fórmula 2** evolui para *itemsets* totais = $\sum_{p=1}^{m} \binom{m}{p} = 2^m - 1$ **(3)**

Será através dos *Itemsets* definidos na base de dados que o algoritmo de mineração em regras de associação realizará o processo de revelação dos relacionamentos nas transações registradas na base de dados. O primeiro algoritmo desenvolvido baseado em *Itemsets* foi o Algoritmo APRIORI que foi desenvolvido por *Agrawal et al* [2].

2.3 REGRA DE ASSOCIAÇÃO

Regras de associação são basicamente padrões de relacionamento frequentes identificados em um conjunto de dados. O objetivo principal é encontrar elementos ou grupos de elementos (*itemsets*) que implicam em outros elementos dentro de uma transação, em um grupo de transações registradas ou armazenadas em um banco de dados, ou seja, encontrar relacionamentos ou padrões frequentes em um conjunto de dados.

Um dos maiores desafios em definição de regras de associação é sua utilização em grandes bancos de dados. Os bancos de dados das organizações que possuem interesse em regras de associação não param de crescer [40][41][42], o que torna o processo de descoberta das regras de associação cada vez mais custoso computacionalmente. Como demonstrado no **gráfico 1,** o crescimento de itens na base de dados gera um aumento exponencial de *itemsets* (**fórmula 3**), e em consequência o número de regras de associações em potencial e o tempo de processamento se eleva na mesma proporção, o gerenciamento de memória se torna um grande problema, o que obriga o desenvolvimento de algoritmos cada vez mais eficientes.

Em uma base de dados que possui *p-itemsets, o* número total de *itemsets* é definido pela equação 2.3 (*itemsets* totais = $2^m - 1$), sendo m o total de itens na base de dados, cada *p-itemset* gera 2^p-1 regras em potencial. Sendo assim o total de regras de associação em potencial (ou seja, todas as possíveis combinações de *p-itemsets* com itens ou grupo de itens contidos neste *itemset*) é:

Total de regras de associação possíveis é = $$\sum_{p=1}^{m} \binom{m}{p} \times (2^p - 1) = 3^m - 2^m \qquad \textbf{(4)}$$

Como em regras de associação não é possível a implicação de um item por ele mesmo, ou seja, o item implicado não pode estar contido no conjunto de itens que o implica. A fórmula de contagem de possíveis regras evolui para:

Total de regras de associação em potencial é = $$\sum_{p=1}^{m-1} \binom{m}{p} \times 2^p - 1 \qquad \textbf{(5)}$$

Em uma base de dados com m = 4 itens, I = {Lápis, Apontador, Estojo, Borracha}
Possui *itemsets* totais = $2^m - 1 = 2^4 - 1 = 16 - 1 = 15$

Total de possíveis regras de associação:

Tot = { [4 x (2^1-1)] + [6 x (2^2-1)] + [4 x (2^3-1)] } = { [4 x 1] + [6 x 3] + [4 x 7] }

Tot = { 4 + 18 + 28 } = 50

Tabela 4: Possíveis candidatos a regras de associação com total de itens m = 4

m	***Itemsets***	**Regras de associação**
1 - (**L**ápis)	1 - (**L**ápis)	Ø – (não é possível obter com m = 1)
2 – (**L**ápis, **A**ponta-dor)	3 – (**L**ápis), (**A**pontador), (**L**ápis, **A**pontador)	2 - (**L**ápis) ⇒ (**A**pontador) (**A**pontador) ⇒ (**L**ápis)
3 – (**L**ápis, **A**pontador, **E**stojo)	7 – (**L**), (**A**), (**E**), (**L**, **A**), (**L**, **E**), (**A**, **E**), (**L**, **A**, **E**)	12 - (**L**) ⇒ (**A**), (**A**) ⇒ (**L**), (**L**) ⇒ (**E**) (**E**) ⇒ (**L**), (**A**) ⇒ (**E**), (**E**) ⇒ (**A**) (**L**) ⇒ (**A**, **E**), (**A**, **E**) ⇒ (**L**), (**A**) ⇒ (**L**, **E**) (**L**, **E**) ⇒ (**A**), (**L**, **A**) ⇒ (**E**), (**E**) ⇒ (**L**, **A**)
4 – (**L**ápis, **A**pontador, **E**stojo, **B**orracha)	15 - (**L**), (**A**), (**E**), (**B**), (**L**, **A**), (**L**, **E**), (**L**, **B**),(**A**, **E**), (**A**, **B**), (**E**, **B**) (**L**, **A**, **E**), (**L**, **A**, **B**), (**A**, **E**, **B**) , (**L**, **E**, **B**), (**L**, **A**, **E**, **B**)	50 - (**L**) ⇒ (**A**), (**A**) ⇒ (**L**), (**L**) ⇒ (**E**), (**E**) ⇒ (**L**), (**L**) ⇒ (**B**), (**B**) ⇒ (**L**), (**A**) ⇒ (**E**), (**E**) ⇒ (**A**), (**A**) ⇒ (**B**), (**B**) ⇒ (**A**), (**E**) ⇒ (**B**), (**B**) ⇒ (**E**), (**L**) ⇒ (**A**, **E**), (**A**, **E**) ⇒ (**L**), (**L**) ⇒ (**A**, **B**), (**A**, **B**) ⇒ (**L**), (**L**) ⇒ (**E**, **B**), (**E**, **B**) ⇒ (**L**), (**A**) ⇒ (**L**, **E**), (**L**, **E**) ⇒ (**A**), (**A**) ⇒ (**L**, **B**), (**L**, **B**) ⇒ (**A**), (**A**) ⇒ (**E**, **B**), (**E**, **B**) ⇒ (**A**), (**E**) ⇒ (**L**, **A**), (**L**, **A**) ⇒ (**E**), (**E**) ⇒ (**L**, **B**), (**L**, **B**) ⇒ (**E**), (**E**) ⇒ (**A**, **B**), (**A**, **B**) ⇒ (**E**), (**B**) ⇒ (**L**, **A**), (**L**, **A**) ⇒ (**B**), (**B**) ⇒ (**L**, **E**), (**L**, **E**) ⇒ (**B**), (**B**) ⇒ (**A**, **E**), (**A**, **E**) ⇒ (**B**), (**L**, **A**) ⇒ (**E**, **B**), (**E**, **B**) ⇒ (**L**, **A**), (**L**, **E**) ⇒ (**A**, **B**), (**A**, **B**) ⇒ (**L**, **E**), (**L**, **B**) ⇒ (**A**, **E**), (**A**, **E**) ⇒ (**L**, **B**), (**L**) ⇒ (**A**, **E**, **B**), (**A**, **E**, **B**) ⇒ (**L**), (**A**) ⇒ (**L**, **E**, **B**), (**L**, **E**, **B**) ⇒ (**A**), (**E**) ⇒ (**L**, **A**, **B**), (**L**, **A**, **B**) ⇒ (**E**), (**B**) ⇒ (**L**, **A**, **E**), (**L**, **A**, **E**) ⇒ (**B**)

Tabela 5: Relação de valores acumulados de itens, *Itemsets* e candidatos a regras de associação

m	***Itemsets***	**Regras de associação**
1	1	Ø
2	3	2
3	7	12
4	15	50
5	31	180
6	63	602

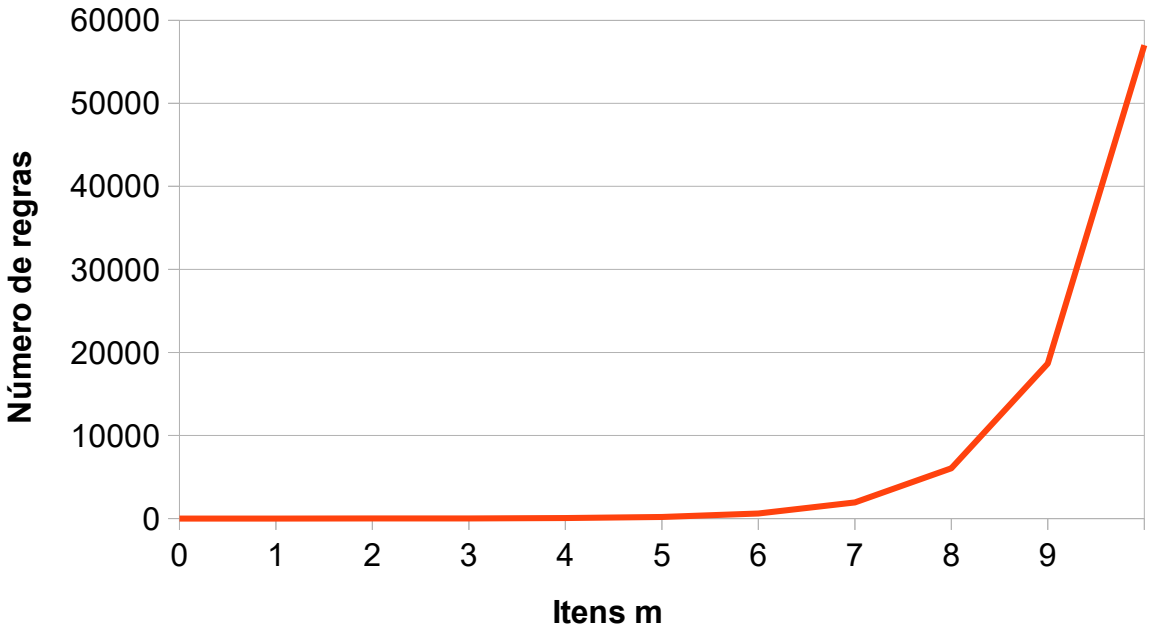

Gráfico 1: Evolução de itens x Candidatos a regras de associação

O conceito introduzido em 1993 por *Agrawal et al.*[1] descreve as relações entre *itemsets*, e consiste em declarações da seguinte forma, $A \Rightarrow C$ onde A (antecedente) e C (consequente) representam conjuntos disjuntos de itens de uma mesma

transação. Considerando que *A* é subconjunto de *I*, uma transação *T* contém *A* se para todos os itens i_k em *A*, t[k] = 1.

Uma regra de associação é uma implicação da forma $A \Rightarrow C$, onde $A \subset I$, $C \subset I$, e $A \cap C = \emptyset$, ou seja, *A* é subconjunto de *I*, e *C* é subconjunto de *I* que não está presente em *A*. A regra $A \Rightarrow C$ é válida no conjunto de transações *T*, com o grau de confiança $0 \leq \text{Conf} \leq 1$, se no mínimo Conf das transações em *T* que contêm *A* também contêm *C.* A regra $A \Rightarrow C$ tem suporte Supp em *T*, se Supp das transações em *T* contêm *A* U *C.* Se as condições forem satisfeitas, Conf representará o fator de confiabilidade e Supp o fator de suporte.

2.4 SUPORTE

Um dos maiores desafios em regras de associação é o tamanho do banco de dados a ser minerado. Como a quantidade de possíveis regras de associação cresce exponencialmente de acordo com o número de itens presentes no banco de dados, se faz necessário a construção de métodos que evidenciem e tornem mais eficientes os processos de descoberta de regras de associação pois o custo computacional cresce na mesma proporção. Sendo assim se fez necessário a introdução de métodos que separem de todas as regras existentes, as que possuem maior interesse.

Com o objetivo de reduzir o número de regras evitando gerar regras que cubram poucos eventos, ou seja, *itemsets* pouco frequentes, utiliza-se o conceito de “Suporte”. O suporte considera a frequência que os *itemsets* candidatos estão presentes nas transações do banco de dados. Através de uma restrição definida como “Suporte mínimo” revelam-se os *itemsets* que possuam um valor de frequência ou suporte maior ou igual a este. O valor de suporte mínimo pode ser descrito como valor absoluto (valor ≥ 1) ou valor relativo ($0 \leq$ valor ≤ 1). Pegando nossa base de dados (**tabela 1**), vamos estabelecer um suporte mínimo = 3.

Tabela 6: Relação de itens presentes nas transações

Nº	**Itens presentes na transação**
1ª	Caneta-CN, Borracha-B, Estojo-E
2ª	Lápis-L, Caneta-CN
3ª	Caneta-CN, Caderno-CD, Borracha-B, Apontador-A, Estojo-E
4ª	Apontador-A, Estojo-E
5ª	Lápis-L, Caneta-CN, Caderno-CD
6ª	Caderno-CD, Estojo-E
7ª	Lápis-L, Borracha-B, Apontador-A
8ª	Caderno-CD
9ª	Lápis-L, Caneta-CN, Caderno-CD, Borracha-B, Apontador-A, Estojo-E
10ª	Caderno-CD, Borracha-B, Apontador-A

Tabela 7: Identificação de *Itemsets* mais frequentes em 2-itemset na BD

transações	**2-*Itemset***	**Suporte *Itemset***	**Suporte Mínimo**	**Aprovados**
1ª → 10ª	(L,CN)	3	3	√
1ª → 10ª	(L,CD)	2	3	*X* - não aprovado
1ª → 10ª	(L,B)	2	3	*X* - não aprovado
1ª → 10ª	(L,A)	2	3	*X* - não aprovado
1ª → 10ª	(L,E)	1	3	*X* - não aprovado
1ª → 10ª	(CN,CD)	3	3	√
1ª → 10ª	(CN,B)	3	3	√
1ª → 10ª	(CN,A)	2	3	*X* - não aprovado
1ª → 10ª	(CN,E)	3	3	√
1ª → 10ª	(CD,B)	3	3	√
1ª → 10ª	(CD,A)	3	3	√
1ª → 10ª	(CD,E)	3	3	√
1ª → 10ª	(B,A)	4	3	√
1ª → 10ª	(B,E)	3	3	√
1ª → 10ª	(A,E)	3	3	√

Tabela 8: Identificação de *Itemsets* mais frequentes em 1-itemset na BD

transações	1-*Itemset*	Suporte *Itemset*	Suporte Mínimo	Aprovados
1ª → 10ª	(L)	4	3	√
1ª → 10ª	(CN)	5	3	√
1ª → 10ª	(CD)	6	3	√
1ª → 10ª	(B)	5	3	√
1ª → 10ª	(A)	5	3	√
1ª → 10ª	(E)	5	3	√

Tabela 9: Identificação de *Itemsets* mais frequentes em 3-itemset na BD

transações	3-*Itemset*	Suporte *Itemset*	Suporte Mínimo	Aprovados
1ª → 10ª	(L,CN,CD)	2	3	*X* - não aprovado
1ª → 10ª	(L,CN,B)	1	3	*X* - não aprovado
1ª → 10ª	(L,CN,A)	1	3	*X* - não aprovado
1ª → 10ª	(L,CN,E)	1	3	*X* - não aprovado
1ª → 10ª	(L,CD,B)	1	3	*X* - não aprovado
1ª → 10ª	(L,CD,A)	1	3	*X* - não aprovado
1ª → 10ª	(L,CD,E)	1	3	*X* - não aprovado
1ª → 10ª	(L,B,A)	2	3	*X* - não aprovado
1ª → 10ª	(L,B,E)	1	3	*X* - não aprovado
1ª → 10ª	(L,A,E)	1	3	*X* - não aprovado
1ª → 10ª	(CN,CD,B)	2	3	*X* - não aprovado
1ª → 10ª	(CN,CD,A)	2	3	*X* - não aprovado
1ª → 10ª	(CN,CD,E)	2	3	*X* - não aprovado
1ª → 10ª	(A,B,CN)	2	3	*X* - não aprovado
1ª → 10ª	(E,B,CN)	3	3	√
1ª → 10ª	(A,E,CN)	2	3	*X* - não aprovado
1ª → 10ª	(CD,B,A)	3	3	√
1ª → 10ª	(CD,B,E)	2	3	*X* - não aprovado
1ª → 10ª	(A,E,CD)	2	3	*X* - não aprovado
1ª → 10ª	(B,A,E)	2	3	*X* - não aprovado

Tabela 10: Identificação de *Itemsets* mais frequentes em 4-itemset na BD

transações	**4-*Itemset***	**Suporte *Itemset***	**Suporte Mínimo**	**Aprovados**
1ª → 10ª	(L,CN,CD,B)	1	3	*X* - não aprovado
1ª → 10ª	(L,CN,CD,A)	1	3	*X* - não aprovado
1ª → 10ª	(L,CN,CD,E)	1	3	*X* - não aprovado
1ª → 10ª	(L,CN,B,A)	1	3	*X* - não aprovado
1ª → 10ª	(L,CN,B,E)	1	3	*X* - não aprovado
1ª → 10ª	(L,CN,A,E)	1	3	*X* - não aprovado
1ª → 10ª	(L,CD,B,A)	1	3	*X* - não aprovado
1ª → 10ª	(L,CD,B,E)	1	3	*X* - não aprovado
1ª → 10ª	(L,CD,A,E)	1	3	*X* - não aprovado
1ª → 10ª	(L,B,A,E)	1	3	*X* - não aprovado
1ª → 10ª	(CN,CD,B,A)	2	3	*X* - não aprovado
1ª → 10ª	(CN,CD,B,E)	2	3	*X* - não aprovado
1ª → 10ª	(CN,CD,A,E)	2	3	*X* - não aprovado
1ª → 10ª	(CN,B,A,E)	2	3	*X* - não aprovado
1ª → 10ª	(CD,B,A,E)	2	3	*X* - não aprovado

Tabela 11: Identificação de *Itemsets* mais frequentes em 5-itemset na BD

transações	**1-*Itemset***	**Suporte *Itemset***	**Suporte Mínimo**	**Aprovados**
1ª → 10ª	(L,CN,CD,B,A)	1	3	*X* - não aprovado
1ª → 10ª	(L,CN,CD,B,E)	1	3	*X* - não aprovado
1ª → 10ª	(L,CN,CD,A,E)	1	3	*X* - não aprovado
1ª → 10ª	(L,CN,B,A,E)	1	3	*X* - não aprovado
1ª → 10ª	(L,CD,B,A,E)	1	3	*X* - não aprovado
1ª → 10ª	(CN,CD,B,A,E)	2	3	*X* - não aprovado

Tabela 12: Identificação de *Itemsets* mais frequentes em 6-itemset na BD

transações	**6-*Itemset***	**Suporte *Itemset***	**Suporte Mínimo**	**Aprovados**
1ª → 10ª	(L,CN,CD,B,A,E)	1	3	*X* - não aprovado

Baseado no exemplo e utilizando um suporte mínimo de valor 3, diminuímos significativamente os *itemsets* de base de 63 (**tabela 5**) para 18**.**

O conceito de “Suporte” introduzido em 1993 por *Agrawal et al.*[1] representa a quantidade de *itemsets* na base de dados que apoiam uma regra. É a probabilidade de um *itemset X* estar contido em uma transação *t* de uma base de dados *D*.

Pode ser representado por: $$Supp(X)=\frac{|\{t\in D; X\subseteq t\}|}{|D|}=P(X) \qquad \textbf{(6)}$$

O suporte utiliza o princípio dos *Itemsets* frequentes, que diz que:

- Qualquer subconjunto de um *itemset* frequente também é um *itemset* frequente.

  Este princípio é válido devido a seguinte propriedade da métrica suporte:

- O suporte de um *itemset X* nunca excede o suporte de seus sub-*itemsets X'* (propriedade anti-monotônica do suporte).
- $\forall X \forall Y : X' \subseteq X \rightarrow suporte(X') \geq suporte(X)$

2.5 CONFIANÇA

Após a identificação de todos os *itemsets* frequentes o próximo passo seria a transformação de todos os *itemsets* frequentes em regras de associação. Com a métrica de suporte e a restrição de suporte mínimo identifica-se então uma quantidade de *itemsets* candidatos a regras de associação com uma relação de importância de frequências associadas a base de dados mas ainda não possuem uma característica de correlação entre os itens dentro de um determinado *itemset.*

Foi então introduzida uma outra métrica definida como "Confiança", que é uma medida de força que a regra escolhida possui dentro da base de dados. Através de uma restrição definida como "Confiança mínima" estabelece-se uma precisão para *itemsets* que possuam um valor de correlação maior ou igual a este identificando assim nossas regras de associação de interesse.

O valor de confiança mínima pode ser descrito como um valor relativo ($0 \leq$ valor ≤ 1) ou um valor percentual (valor relativo $\times$ 100). Pegando nossos *itemsets* aprovados, com suporte mínimo = 3 (**tabela 13**), vamos estabelecer um valor de confiança mínima = 0,7 e avaliar as regras de associação (**tabela 14**).

Introduzido por *Agrawal et al* [1] em 1993, ele pode ser descrito como:

Dada uma associação $i = A \Rightarrow C$

Dado um *itemset t* tal que $t \supset A$;

Se $A \not\subset t$, podemos afirmar que a regra *i* não se aplica a *t*, já que a existência de *A*, segundo *i,* implica na existência de *C*.

A confiança pode ser definida então como a probabilidade de que *t* contenha também o termo consequente (seja verdadeira).

É dada por:

$$Conf(A \Rightarrow C) = \frac{Supp(A \Rightarrow C)}{Supp(A)} = \frac{Supp(A \cup C)}{Supp(A)} = \frac{P(A \cap C)}{P(A)} = P(C \mid A) \qquad \textbf{(7)}$$

Tabela 13: Relação de *Itemsets* candidatos a regras de associação

***Itemsets* aprovados**	**Valor de Suporte relativo**	**Candidatos a regras**
(L)	4/10 = 0,4	*X* – não aprovado
(CN)	5/10 = 0,5	*X* – não aprovado
(CD)	6/10 = 0,6	*X* – não aprovado
(B)	5/10 = 0,5	*X* – não aprovado
(A)	5/10 = 0,5	*X* – não aprovado
(E)	5/10 = 0,5	*X* – não aprovado
(L,CN)	3/10 = 0,3	√
(CN,CD)	3/10 = 0,3	√
(CN,B)	3/10 = 0,3	√
(CN,E)	3/10 = 0,3	√
(CD,B)	3/10 = 0,3	√
(CD,A)	3/10 = 0,3	√
(CD,E)	3/10 = 0,3	√
(B,A)	4/10 = 0,4	√
(B,E)	3/10 = 0,3	√
(A,E)	3/10 = 0,3	√
(E,B,CN)	3/10 = 0,3	√
(CD,B,A)	3/10 = 0,3	√

Tabela 14: Regras de associação

Regras candidatas	Suporte	Confiança	Conf. Mínima	Aprovadas
L ⇒ CN	0.3	0,3 / 0,4 = 0,75	0,7	√
CN ⇒ L	0.3	0,3 / 0,5 = 0,6	0,7	Não aprovadas
CN ⇒ CD	0.3	0,3 / 0,5 = 0,6	0,7	Não aprovadas
CD ⇒ CN	0.3	0,3 / 0,6 = 0,5	0,7	Não aprovadas
CN ⇒ B	0.3	0,3 / 0,5 = 0,6	0,7	Não aprovadas
B ⇒ CN	0.3	0,3 / 0,5 = 0,6	0,7	Não aprovadas
CN ⇒ E	0.3	0,3 / 0,5 = 0,6	0,7	Não aprovadas
E⇒ CN	0.3	0,3 / 0,5 = 0,6	0,7	Não aprovadas
CD ⇒ B	0.3	0,3 / 0,6 = 0,5	0,7	Não aprovadas
B ⇒ CD	0.3	0,3 / 0,5 =0,6	0,7	Não aprovadas
CD ⇒ A	0.3	0.3 / 0,6 = 0,5	0,7	Não aprovadas
A ⇒ CD	0.3	0.3 / 0,5 = 0,6	0,7	Não aprovadas
CD ⇒ E	0.3	0.3 / 0,6 = 0,5	0,7	Não aprovadas
E ⇒ CD	0.3	0.3 / 0,5 = 0,6	0,7	Não aprovadas
B ⇒ A	0.4	0.4 / 0,5 = 0,8	0,7	√
A ⇒ B	0.4	0.4 / 0,5 = 0,8	0,7	√
B ⇒ E	0.3	0.3 / 0,5 = 0,6	0,7	Não aprovadas
E ⇒ B	0.3	0.3 / 0,5 = 0,6	0,7	Não aprovadas
A ⇒ E	0.3	0.3 / 0,5 = 0,6	0,7	Não aprovadas
E ⇒ A	0.3	0.3 / 0,5 = 0,6	0,7	Não aprovadas
E ⇒ (B,CN)	0.3	0.3 / 0,5 = 0,6	0,7	Não aprovadas
B ⇒ (E,CN)	0.3	0.3 / 0,5 = 0,6	0,7	Não aprovadas
CN ⇒ (B,E)	0.3	0.3 / 0,5 = 0,6	0,7	Não aprovadas
(B,CN) ⇒ E	0.3	0.3 / 0,3 = 1	0,7	√
(E,CN) ⇒ B	0.3	0.3 / 0,3 = 1	0,7	√
(B,E) ⇒ CN	0.3	0.3 / 0,3 = 1	0,7	√
CD ⇒ (B,A)	0.3	0.3 / 0,6 =0,5	0,7	Não aprovadas
B ⇒ (CD,A)	0.3	0.3 / 0,5 = 0,6	0,7	Não aprovadas
A ⇒ (CD,B)	0.3	0.3 / 0,5 = 0,6	0,7	Não aprovadas
(B,A) ⇒ CD	0.3	0.3 / 0,4 = 0,75	0,7	√
(CD,A) ⇒ B	0.3	0.3 / 0,3 = 1	0,7	√
(CD,B) ⇒ A	0.3	0.3 / 0,3 = 1	0,7	√

3 MÉTRICAS DE REGRAS DE ASSOCIAÇÃO

O modelo de obtenção de regras de associação através das métricas de “Suporte” e “Confiança” é o universal e o mais utilizado. Embora suporte e confiança sejam métricas apropriadas para a construção de um modelo forte na maior parte dos casos, elas não são consideradas por vários autores [5] [3] [4] as medidas ideais ou suficientes e nem conseguem abranger todos os casos interessantes ou relevantes, pois na maioria dos casos é gerado um conjunto muito grande de regras (muitas destas regras redundantes ou não interessantes) que podem prejudicar a eficácia ou nem atingir alguns casos para os quais se tenha interesse. Em função dessas necessidades foram desenvolvidas métricas alternativas para estender ou modificar o modelo básico de suporte e confiança para alcançar casos em que o modelo básico não se mostra o mais eficiente ou eficaz.

As métricas objetivas são definidas em termos das contagens de frequência tabuladas em uma tabela de contingência 2x2, conforme mostrado na **tabela 15**.

Tabela 15: Tabela de Contingência 2x2 para $A \Rightarrow C$

	C	$\overline{C}$	
A	c_{AC}	$c_{A\overline{C}}$	c_A
$\overline{A}$	$c_{\overline{A}C}$	$c_{\overline{A}\overline{C}}$	$c_{\overline{A}}$
	c_C	$c_{\overline{C}}$	N

Batanero et al. [37] em seu trabalho descreve o uso de tabelas de contingência para a análise de associação ou dependência estatística entre variáveis. Tabelas de contingência são tabelas que apresentam os dados entre múltiplas variáveis categóricas (ou qualitativas) com o objetivo de investigar uma possível associação entre elas. Uma tabela de contingência 2x2, é uma tabela que demonstra as ocorrências de 2 variáveis categóricas, em nosso caso especificamente, dicotômicas (que podem assumir 2 valores). Suas células possuem o seguinte significado:

- $A \rightarrow$ Variável descrita como termo antecedente em uma determinada regra de associação.
- $C \rightarrow$ Variável descrita como termo consequente em uma determinada regra de associação.
- $\overline{A} \rightarrow$ Variável descrita como a complementar do termo antecedente em uma determinada regra de associação, ou seja, em transações (na base transacional) onde A não ocorre.
- $\overline{C} \rightarrow$ Variável descrita como a complementar do termo consequente em uma determinada regra de associação, ou seja, em transações (na base transacional) onde C não ocorre.
- $c_{AC} \rightarrow$ Variável descrita como o número de ocorrências do itemset AC na base transacional.
- $c_{A\overline{C}} \rightarrow$ Variável descrita como o número de ocorrências do itemset $A\overline{C}$ na base transacional.
- $c_{\overline{A}C} \rightarrow$ Variável descrita como o número de ocorrências do itemset $\overline{A}C$ na base transacional.
- $c_{\overline{A}\overline{C}} \rightarrow$ Variável descrita como o número de ocorrências itemset $\overline{A}\overline{C}$ na base transacional.
- $c_A \rightarrow$ Variável descrita como o número de ocorrências do itemset A na base transacional.
- $c_C \rightarrow$ Variável descrita como o número de ocorrências do itemset C na base transacional.
- $c_{\overline{A}} \rightarrow$ Variável descrita como o número de ocorrências do itemset $\overline{A}$ (complementar do termo antecedente) na base transacional.
- $c_{\overline{C}} \rightarrow$ Variável descrita como o número de ocorrências do itemset $\overline{C}$ (complementar do termo antecedente) na base transacional.
- $N \rightarrow$ Variável descrita como o número total de ocorrências na base transacional, ou seja, $(c_A + c_C + c_{\overline{A}} + c_{\overline{C}})$.

3.1 ADDED VALUE

Descrito por *Merceron et al.* [34], A métrica *Added Value* da regra $A \Rightarrow C$ mede se a proporção de transações contendo *C* entre as transações que contêm *A* é maior que a proporção de transações que contêm *C* entre todas as transações. Então, somente se a probabilidade de encontrar o item *C* quando o item *A* for encontrado for maior que a probabilidade de encontrar o item *C*, podemos dizer que *A* e *C* são associados e que *A* implica *C*. Um número positivo indica que *A* e *C* estão relacionados, enquanto um número negativo significa que a ocorrência de *A* impede que *C* ocorra e um valor igual a 0 significa independência, em termos de probabilidade, isso significa que a ocorrência de *A* e a ocorrência de *C* na mesma transação são eventos independentes, portanto, *A* e *C* não estão correlacionados.

Sua escala de valores está presente no intervalo [-1, 1] e sua fórmula é definida por:

$$Added-Value(A \Rightarrow C) = Conf(A \Rightarrow C) - Supp(C) \qquad \textbf{(8)}$$

A métrica *Added Value* está intimamente relacionada a outra medida de interesse conhecida como *Lift*.

3.2 ALL-CONFIDENCE

Omiecinski [3] afirma que a métrica *All-confidence* é uma métrica alternativa ao modelo suporte e confiança para a revelação de regras de associação em casos específicos pois possui a característica de anti-monoticidade (fechamento descendente) e consegue revelar padrões associados ao problema de itens raros (itens raros são itens que possuem baixa frequência na base transacional, ou seja, que integram *itemsets* com baixo suporte).

Essa medida é definida como a menor confiança de todas as regras que podem ser produzidas a partir de um conjunto de itens, ou seja, todas as regras produzidas a partir de um conjunto de itens terão uma confiança maior ou igual ao seu valor de *All-confidence*.

Assume valores entre 0 e 1, próximos de 0 indica que os itens estão poucos relacionados e próximos de 1 indica que os itens desse itemset estão muito relacionados, implicando em melhores regras.

Sua fórmula é definida por:

$$All-Confidence(A\cup C)=\frac{Supp(A\cup C)}{max(i)_{i\in(A\cup C)}(Supp(i))}=\frac{P(A\cup C)}{max(i)_{i\in(A\cup C)}(P(i))}$$

(9)

$$All-Confidence(A\cup C)=min\{P(A|C),P(C|A)\}$$

3.3 CAUSAL SUPPORT

Causal Support é uma métrica introduzida em 1999 por *Yves Kodratoff* [4]. De acordo com o autor o objetivo é reforçar os valores de suporte associados aos *itemsets* de acordo com sua contraposição (confirmação de hipótese), potencializando a intensidade da frequência do itemset em relação a base de dados transacional, baseado na observação do "paradoxo de *Hempel* e a teoria da confirmação"[4 p.2,3].

Seja uma regra de associação $A \Rightarrow C$. A métrica *Causal support* assume valores entre 0 e 2 levando em consideração também as transações cujos itens não estão contidos em *A U C*. Valor próximo de 0 indica que os itens estão poucos relacionados e próximo de 2 indica que os itens desse itemset estão muito relacionados, implicando em melhores regras.

Sua fórmula é dada por:

$$Causal\ Support(A\cup C)=Supp(A\cup C)+Supp(\overline{A}\cup\overline{C})=P(A\cap C)+P(\overline{A}\cap\overline{C}) \qquad \textbf{(10)}$$

3.4 CAUSAL CONFIDENCE

Causal Confidence é uma métrica introduzida por *Yves Kodratoff* [4]. De acordo com o autor o objetivo da métrica é adicionar a confiança trazida pelas ins-

tâncias diretas da implicação e a confiança trazida pela sua contraposição. O autor afirma que esta não é ainda uma medida ideal pois não considera o valor da descon-firmação de hipótese em seus cálculos [4 p.8].

Assume valores entre 0 e 1, sendo o valor próximo de 0 indica que os itens estão menos correlacionados e próximo de 1 indica que os itens desse itemset estão mais correlacionados, implicando em melhores regras.

Sua fórmula é dada por:

$$Causal-Confidence(A\Rightarrow C)=1/2\cdot[Conf(A\Rightarrow C)+Conf(\overline{A}\Rightarrow\overline{C})]=1/2\cdot[P(C\mid A)+P(\overline{C}\mid\overline{A})] \quad \textbf{(11)}$$

3.5 CERTAINTY FACTOR

Também chamado de *Loevinger*, foi definido por *Berzal et tal.* [5]. Descrito pelos autores como uma medida da variação da probabilidade de *C* estar em uma transação quando consideramos apenas aquelas transações em que *A* está. Mais especificamente, um *CF* positivo mede a diminuição da probabilidade de que C não esteja em uma transação, dado que *A* está. Uma interpretação semelhante pode ser feita para *CFs* negativos.

Pode assumir valores entre -1 e 1, onde valores mais altos implicam em uma baixa probabilidade de encontrarmos *C* em uma transação sem *A*, e valores mais baixos, em uma tendência a encontrarmos *C* em uma transação sem *A*, valores próximos a 0 indicam maior tendência a independência entre os termos.

Sua fórmula é dada por:

$$Certainty-Factor(A\Rightarrow C)=\frac{Conf(A\Rightarrow C)-Supp(C)}{Supp(\overline{C})}=\frac{P(C\mid A)-P(C)}{P(\overline{C})} \quad \textbf{(12)}$$

3.6 CHI-SQUARED

O teste *Chi-squared* (qui-quadrado) [6] é um teste de hipóteses frequentemente utilizado para encontrar valores de dispersão entre duas variáveis categóricas nominais para grandes amostras, alternativamente ao *Fisher's exact test*.

Em regras de associação pode ser usado para se estimar a independência estatística entre dois termos da regra de associação (antecedente e consequente) [7], e retorna valores entre 0 e +∞.

Sua fórmula é dada por:

$$X^2 = \frac{\sum_{i=0}^{1}\sum_{j=0}^{1}(N_{ij} - E_{ij})^2}{E_{ij}} \qquad \textbf{(13)}$$

Onde N_{ij} são as frequências observadas, *T* o valor total das amostras e E_{ij} (frequências esperadas) é dado por:

$$E_{ij} = \frac{N_i * N_j}{E_{ij}} \qquad \textbf{(14)}$$

3.7 CROSS-SUPPORT RATIO

Definido por *Xiong et tal.* [8]. De acordo com os autores, estratégias de remoção baseadas em suporte não são eficazes para conjuntos de dados com distribuições de suporte inclinadas. Estes propõem o conceito de padrão "hiperclique", que usa uma medida objetiva chamada *h-confidence* (igual à métrica *All-Confidence*) para identificar padrões de afinidade fortes.

Os autores apresentam o conceito de "propriedade de suporte cruzado". Essa propriedade é útil para evitar a geração de padrões de suporte cruzado, que são padrões que contêm itens de níveis de suporte substancialmente diferentes. A geração dos chamados "padrões de suporte cruzado" (padrões com itens com su-

porte substancialmente distante entre eles) é evitada pela propriedade de suporte cruzado da medida *h-confidence* sob *itemsets.* Os autores descrevem esta métrica como uma métrica alternativa ao suporte (restrição de suporte mínimo), pois possuí a propriedade de anti-monoticidade, utilizada para reduzir o espaço de pesquisa combinatória, e também trata do problema relacionado a itens raros, através da utilização do padrão hiperclique. Retorna valores entre 0 e 1, sendo os mais próximos a 0 tendendo a serem padrões espúrios (ou ruído) e os mais próximos a 1 sendo mais relacionados ou associados.

Sua fórmula é dada por:

$$Cross-Support\,Ratio(X)=\frac{min_{i\in X}(Supp(i))}{max_{i\in X}(Supp(i))} \qquad \textbf{(15)}$$

3.8 COLLECTIVE STRENGTH

Definido por *Aggarwal et al.* [9], É descrito como um método diferente para avaliar e encontrar conjuntos de itens referidos como “conjuntos de itens fortemente coletivos”. A força coletiva de um conjunto de itens (*Itemset*) é definida como um número entre 0 e +∞. Um valor 0 indica correlação negativa perfeita, enquanto um valor de ∞ indica correlação perfeitamente positiva. Um valor 1 indica o “ponto de equilíbrio”, se os itens coocorrerem conforme o esperado sob independência, correspondente a um *Itemset* presente no valor esperado.

É descrito na forma qual, um conjunto de itens *I* (*Itemset*) viola uma transação, se alguns dos itens estiverem presentes na transação e outros não, por exemplo, o conceito de violação indica quantas vezes um cliente pode comprar pelo menos alguns dos itens do *Itemset*, mas pode não comprar o restante dos itens. A taxa de violação de um *Itemset I* é denotada por v (*I*) e é a fração de violações do *Itemset I* em todas as transações. Isso também é igual à fração de transações que contêm um subconjunto não nulo adequado de *I*.

É definido da seguinte maneira, seja *B* uma base de dados de transações *t* tal que *B* = {t_1, t_2, …, t_n} e *S* = {s_1, s_2, …, s_m} o conjunto de todos os subconjuntos *s* possíveis de um *itemset I* = { i_1, i_2, …, i_p};

Dado um conjunto *K* tal que $\forall\ x \in \mathbb{N} : 1 \leqslant x \geqslant n;\ K_x = t_x \cap S$ dizemos ter ocorrido uma violação se e somente se $K_x \neq S_x$ **e** $K_x \neq \emptyset$.

Sua fórmula é dada por:

$$Collective-Strength(X)=\frac{1-v(X)}{1-E[v(x)]}\times\frac{E[v(X)]}{v(X)}=\frac{P(A\cap C)+P(\overline{C}\mid\overline{A})}{P(A)P(C)+P(\overline{A})P(\overline{C})} \qquad \textbf{(16)}$$

Onde *E* [v(*I*)] é o valor esperado para itens independentes.

3.9 CONVICTION

Definida por *Brin et al* [10], é descrita como uma métrica alternativa à *Confidence*, que não capta a direção das associações adequadamente. A métrica *Conviction* compara a probabilidade de *A* aparecer sem *C* se eles dependessem da frequência real da aparência de *A* sem *C*. Nesse aspecto, é semelhante a métrica *Lift*, no entanto, em contraste com o *Lift* é uma medida direcionada, pois também usa as informações da ausência do consequente. Um fato interessante é que a convicção é monótona as métricas *Confidence* e *Lift*.

Assume valores entre 0 e +∞, onde 1 implica em uma independência estatística, e +∞ implica em regras que são sempre verdadeiras. Sua fórmula é dada por:

$$Conviction(A\Rightarrow C)=\frac{1-Supp(C)}{1-Conf(A\Rightarrow C)}=\frac{P(A)P(\overline{C})}{P(A\cap\overline{C})} \qquad \textbf{(17)}$$

3.10 COSINE

Descrito por *Tan et al.* [43], como uma métrica nulo-invariante (propriedade definida em tabelas de contingência 2x2, **Fig. 1(f)**, que representa a não alteração de valor resultante, quando for realizada adição de registros que não contém as duas variáveis em observação), essa propriedade é útil quando a consideração de presença de itens é mais importante do que a consideração de ausência de itens. *Tan et al.* [12] a descreve como a média geométrica entre o fator de interesse, métrica descrita como *Interest Factor* ou simplesmente *Interest* (posteriormente chamado de métrica *Lift*) e a métrica de suporte. *Tan et al. o* descreve como uma medida de similaridade amplamente usada para modelos de espaço vetorial, e que pode ser derivada da métrica Φ *Correlation Coefficient*.

Sua escala de valores é definida no intervalo [0, 1] e sua fórmula é descrita por:

$$Cosine(A \Rightarrow C) = \frac{Supp(A \cup C)}{\sqrt{Supp(A)\,Supp(C)}} = \frac{P(A \cap C)}{\sqrt{P(A)P(C)}} = \sqrt{P(A|C) \times P(C|A)} \qquad \textbf{(18)}$$

3.11 COVERAGE

Descrito por *Ochin et al.* [20], como a quantidade de banco de dados coberta por uma regra $A \Rightarrow C$, *Coverage* é o número de transações que satisfazem o antecedente de uma regra. *Coverage* é também chamada suporte antecedente, *Coverage* (*A*) = Supp (*A*). Ele mede com que frequência uma regra $A \Rightarrow C$ é aplicável em um banco de dados.

Sua escala de valores é definida no intervalo [0, 1], significando que quanto mais próximo de 1 maior a relação desta regra com o banco de dados e sua fórmula é descrita por:

$$Coverage(A \Rightarrow C) = Supp(A) = P(A) \qquad \textbf{(19)}$$

3.12 DESCRIPTIVE CONFIRMED CONFIDENCE

Definido por *Kodratoff* [4], o autor descreve a métrica como uma evolução da métrica *Causal Confidence* pois esta não considera em seus cálculos a desconfirmação da regra. Ele afirma que a definição de *Confidence* (definida como P (*A*, *C*) / P (*A*) = P (*C* | *A*)) e de dependência (definida como P (*C* | *A*) - P (*C*)), utilizando o valor de *Confidence* para medir a força da dependência não é adequada. Sendo assim ele define a métrica *Descriptive Confirmed Confidence* com sua escala de valores no intervalo de [-1, 1] sendo os valores próximos a -1 significando que os dados não suportam a hipótese de implicação de $A \Rightarrow C$ e valores próximos a 1 significando que a hipótese de implicação de $A \Rightarrow C$ é desconfirmada através de sua contraposição e fortemente interessante, valores próximos a 0 indicam que os valores de confirmação e desconfirmação da regra estão alinhados tendendo a serem ruído na base de dados.

Sua fórmula é descrita por:

$$Descriptive-Confirmed\ Confidence = Conf(A \Rightarrow C) - Conf(A \Rightarrow \overline{C}) = P(C \mid A) - P(\overline{C} \mid A)$$

(20)

3.13 DIFFERENCE OF CONFIDENCE

Descrito por *Hofmann et al.*[33]. Sua escala de valores é definida no intervalo [-1, 1] e sua fórmula é descrita por:

$$Difference\ of\ Confidence(A \Rightarrow C) = Conf(A \Rightarrow C) - Conf(\overline{A} \Rightarrow C) = P(C \mid A) - P(C \mid \overline{A})$$

(21)

3.14 EXAMPLE AND COUNTER-EXAMPLE RATE

Sua escala de valores é definida no intervalo [0, 1] e sua fórmula é descrita por:

$$ECR(A \Rightarrow C) = \frac{P(A \cap C) - P(A \cap \overline{C})}{P(A \cap C)} \qquad \textbf{(22)}$$

3.15 FISHER'S EXACT TEST

O teste exato de Fisher (criado por Ronald Fisher) é um teste de significância estatística utilizado na análise de tabelas de contingência. Pertence a uma classe de testes exatos, assim chamados por conta da significância do desvio de uma hipótese nula (p-valor) que é calculada de forma exata.

Sua fórmula é definida por:

$$p(A \Rightarrow C) = \frac{(c_{AC})! \times (c_{A\overline{C}})! \times (c_{\overline{A}C})! \times (c_{\overline{A}\overline{C}})!}{(c_A)! \times (c_{\overline{A}})! \times (c_C)! \times (c_{\overline{C}})! \times N!} \qquad \textbf{(23)}$$

O valor de p varia de 0 a 1, onde no caso de regras de associação para hipótese de intensidade de correlação entre antecedente e consequente, o *Fisher's exact test* é realizado baseado em uma comparação com o valor de hipótese nula (ou seja, para a hipótese de não se ter correlação).

A métrica *Fisher's exact test* é utilizada para a análise de tabelas de contingência em que o tamanho da amostra é pequeno.

3.16 GINI INDEX

O Coeficiente Gini é uma medida de desigualdade desenvolvida pelo italiano *Corrado gini* e publicada no documento *"Variabilità e mutabilità"* (*"Variabilidade e mutabilidade"* em 1912). A métrica *Gini index* é um desenvolvimento de *Breiman et al.* [32] para o caso em regras de associação e consiste em um valor entre 0 e 1, onde 0 corresponde à completa igualdade (independência estatística entre antecedente e consequente) e 1 corresponde à completa desigualdade (probabilidade de ocorrência do consequente dado o antecedente é fortemente ligada).

Mede a entropia quadrática através de sua fórmula definida por:

$$Gini(A \Rightarrow C) = P(A)[P(C \mid A)^2 + P(\overline{C} \mid A)^2] + P(\overline{A})[P(C \mid \overline{A})^2 + P(\overline{C} \mid \overline{A})^2] - P(C)^2 - P(\overline{C})^2$$

(24)

3.17 HYPER CONFIDENCE

Definido por *Hahsler et al.*[32], descreve a métrica como um nível de confiança para observação de contagens muito altas ou muito baixas para as regras $A \Rightarrow C$ usando o modelo hipergeométrico [32 p.439,440]. Como as contagens são extraídas de uma distribuição hipergeométrica (representada pela variável aleatória C_{AC} com parâmetros conhecidos fornecidos pelas contagens c_A e c_C, podemos calcular um intervalo de confiança para as contagens observadas c_{AC} decorrentes da distribuição. Sua escala de valores é definida no intervalo [0, 1], em um nível de confiança de, por exemplo,> 0,95 indica que há apenas 5% de chance de a contagem da regra ser gerada aleatoriamente.

Sua fórmula é definida por:

$$Hyper - Confidence(A \Rightarrow C) = 1 - P[C_{AC} \geq c_{AC} \mid c_A, c_C] \qquad \textbf{(25)}$$

3.18 HYPER LIFT

Definido por *Hahsler et al.*[32], descreve a métrica *Hyper-Lift* como uma adaptação da métrica *Lift* mais robusta para contagens baixas usando um modelo de contagem hipergeométrica [32 p. 444, 445] para aplicações onde regras falsas são problemáticas. Seu valor é definido no intervalo de $[0, +\infty)$ (sendo o valor 1 caracterizado como independência estatística entre as variáveis).

Sua fórmula é definida por:

$$Hyper-Lift(A \Rightarrow C) = \frac{c_{AC}}{Q_{\delta}[C_{AC}]} \qquad \textbf{(26)}$$

onde c_{AC} é o número de transações contendo *A* e *C* e Q_{δ} $[C_{AC}]$ é o quantil da distribuição hipergeométrica com os parâmetros c_A e c_C dados por δ (geralmente o quantil de 99 ou 95%).

3.19 IMBALANCE RATIO

A métrica *Imbalance Ratio* foi proposta por *Wu et tal.*[25] e mede o grau de desequilíbrio entre dois eventos em que o *A* e o *C* estão contidos em uma transação. De acordo com *Wu et tal* em muitas aplicações é importante quantificar até que ponto um conjunto de dados é “controverso” para que um analista de dados entenda corretamente os dados.

Sua escala de valores é definida no intervalo [0,1], a proporção é próxima de 0 se as probabilidades condicionais forem semelhantes (ou seja, muito equilibradas) e próxima de 1 se forem muito diferentes, sendo valores acima de 0 interessantes ao objetivo da métrica e 0 totalmente desinteressantes.

Sua fórmula é definida por:

$$IR(A \Rightarrow C) = \frac{|Supp(A) - Supp(C)|}{Supp(A) + Supp(C) - Supp(A \cup C)} = \frac{|P(A|C) - P(C|A)|}{P(A|C) + P(C|A) - P(A|C)P(C|A)}$$

(27)

3.20 IMPORTANCE

A métrica *Importance* [31] é uma métrica definida pela Microsoft em seu algoritmo de regras de associação (onde a métrica *Confidence* é descrita como "probabilidade"), que é uma implementação direta do algoritmo APRIORI. A métrica *Importance* de uma regra calcula a probabilidade do log do lado direito da regra (*C*), dado o lado esquerdo da regra (*A*).

Exemplificando, dada a regra $A \Rightarrow C$, o *Microsoft Analysis Services* calcula a proporção de casos com *A* e *C* sobre casos com *C*, mas sem *A*, e normaliza essa proporção usando uma escala logarítmica. Seu valor é definido no intervalo de (- ∞, +∞) (quanto maior seu valor mais "importante" no contexto do conjunto de regras).

Sua fórmula é definida por:

$$Importance(A \Rightarrow C) = \log_{10}(L(A \Rightarrow C)/L(A \Rightarrow \overline{C})) \qquad \textbf{(28)}$$

onde

L= métrica *Laplace Correctd Confidence* **(38)**

3.21 IMPROVEMENT

Definido por *Bayardo et al.*[30], O *Improvement* de uma regra é a diferença mínima entre sua confiança e a confiança de qualquer sub-regra de mesmo antecedente definido pelo usuário. A ideia é que só se deve estender o *A* da regra se isso melhorar a regra suficientemente. Sua escala de valores é definida no intervalo [0,1] sendo valores próximos de 1 interessantes ao objetivo da métrica.

Sua fórmula é definida por:

$$\text{Improvement}(A \Rightarrow C) = \min_{A' \subset A} (\text{Conf}(A \Rightarrow C) - \text{Conf}(A' \Rightarrow C)) \qquad \textbf{(29)}$$

3.22 JACCARD COEFFICIENT

Desenvolvido por *C.J. van Rijsbergen* [29] e descrito por ele como uma versão normalizada da mais simples medida de associação (Supp(*A* U *C*)). É uma medida nulo-invariante utilizada para medir a sobreposição de *A* e *C* nos registros.

Wu et al. [25] o referencia como uma métrica de nome *Coherence.* Seu valor é definido no intervalo de [0,1] (sendo o 0 caracterizado como independência estatística entre as variáveis). Sua fórmula é definida por:

$$jaccard(A \Rightarrow C) = \frac{Supp(A \cup C)}{Supp(A) + Supp(C) - Supp(A \cup C)} = \frac{P(A \cap C)}{P(A) + P(C) - P(A \cap C)}$$

(30)

3.23 J-MEASURE

Desenvolvido por *Smyth et al.* [28], com a capacidade de medir a entropia cruzada entre variáveis. *Tan et al.* [17] o descreve como uma medida definida através de sua distribuição de probabilidades, assim como as métricas *Mutual Information* e *Gini Index.*

Seu valor é definido no intervalo de [0,1] (sendo o 0 caracterizado como independência estatística entre as variáveis). Sua fórmula é definida por:

$$J - Measure(A \Rightarrow C) = P(A \cap C) \log\left(\frac{P(C \mid A)}{P(C)}\right) + P(A \cap \overline{C}) \log\left(\frac{P(\overline{C} \mid A)}{P(\overline{\overline{C}})}\right) \qquad \textbf{(31)}$$

3.24 KAPPA

Definido por *J.* Cohen [27] como um coeficiente para medir o grau de concordância em escalas nominais. O coeficiente *k* é simplesmente a proporção de discordâncias esperadas ao acaso que não ocorrem, ou alternativamente, é a proporção de concordância depois que o acordo aleatório é removido da consideração.

Tan et al. [17] o descreve como a medida que captura o grau de concordância entre um par de variáveis. Quanto maior a correlação entre as variáveis , maior os valores para $P(A,C)$ e $P(\neg A,\neg C)$, o que resulta em um valor mais alto para k. Seu valor é definido no intervalo de [-1,1] (sendo o valor 0 caracterizando independência) e sua fórmula é definida por:

$$Kappa(A \Rightarrow C) = \frac{P(A \cap C) + P(\overline{A} \cap \overline{C}) - P(A)P(C) - P(\overline{A})P(\overline{C})}{1 - P(A)P(C) - P(\overline{A})P(\overline{C})} \qquad \textbf{(32)}$$

3.25 KLOSGEN

Desenvolvido por *Willi Klösgen* [26] durante o desenvolvimento do *"Statistics Interpreter Explora"*, um sistema assistente protótipo para revelar descobertas interessantes em conjuntos de dados recorrentes.

Seu valor é definido no intervalo de [-1,1] (sendo o 0 caracterizado como independência estatística entre as variáveis). Sua fórmula é definida por:

$$Kn(A \Rightarrow C) = \sqrt{Supp(A \cup C)}\,(Conf(A \Rightarrow C) - Supp(C)) = \sqrt{P(A \cap C)}\,(P(C \mid A) - P(C)) \qquad \textbf{(33)}$$

3.26 KULCZYNSKI

Estudo de *Kulczynski* [24] sobre associações é reexaminado, em conjunto com outras medidas, por *Wu et al.* [25], que mostra que este conjunto de medidas de interesse nulo-invariantes podem ser expressas como a média matemática generalizada levando a uma ordenação total delas.

Descrito por *Wu et tal.* como, “Inúmeras medidas de interesse foram propostas em estatística e mineração de dados para avaliar as relações dos objetos. Isso é especialmente importante em estudos recentes de mineração de padrões de associação ou correlação. No entanto, ainda não está claro se existe algum relacionamento intrínseco entre muitas medidas propostas e qual é realmente eficaz na aferição de relacionamentos de objetos em grandes conjuntos de dados. Estudos recentes identificaram uma propriedade crítica, invariância nula (transação), para medir associações entre eventos em grandes conjuntos de dados, porém, nem toda medida possui essa propriedade. Neste estudo, reexaminamos um conjunto de medidas de interesse invariantes nulos e descobrimos que elas podem ser expressas como a média matemática generalizada, levando a uma ordenação total delas.”

Seu valor é definido no intervalo de [0,1] (0,5 caracteriza neutralidade ou simplesmente desinteressante). Sua fórmula é definida por:

$$Kulc(A \Rightarrow C) = \frac{1}{2}(Conf(A \Rightarrow C) + Conf(C \Rightarrow A)) = \frac{1}{2}(\frac{Supp(A \cup C)}{Supp(A)} + \frac{Supp(A \cup C)}{Supp(C)})$$

(34)

$$Kulc(A \Rightarrow C) = \frac{1}{2}(P(A|C) + P(C|A))$$

(35)

3.27 GOODMAN-KRUSKAL (λ)

Tan et al. [17] descreve o índice λ como índice de associação preditiva, foi proposto inicialmente por *Goodman & Kruskal* [21] e seu valor é definido no intervalo de [0,1]. Eles descrevem que a intuição por trás dessa medida é que, se duas variáveis são altamente dependentes uma da outra, o erro ao prever uma delas seria pequeno sempre que o valor da outra variável for conhecido. É usado para capturar a quantidade de redução no erro de previsão.

Sua fórmula é definida por:

$$\lambda(A \Rightarrow C) = \frac{\sum_{a \in A} max_{c \in C} P(a \cap c) - max_{c \in C} P(C)}{n - max_{c \in C} P(C)} \qquad \textbf{(36)}$$

$$\lambda(A \Rightarrow C) = \frac{\sum_{j} max_{k} P(A_j, C_k) + \sum_{k} max_{j} P(A_j, C_k) - max_{j} P(A_j) - max_{k} P(B_k)}{2 - max_{j} P(A_j) - max_{k} P(C_k)}$$

(37)

3.28 LAPLACE CORRECTED CONFIDENCE

Tan et al.[17] descreve a métrica *LCC* como uma correção para a métrica *Confidence*, foi desenvolvido por *Clark et al* [23]. A métrica *Confidence* é frequentemente usada para medir a precisão de uma determinada regra, no entanto, pode produzir resultados enganosos, especialmente quando o apoio à regra consequente é superior à confiança da regra [22]. A métrica *Laplace Corrected Confidence* diminui com menor suporte para explicar a incerteza de estimativa com contagens baixas. Sua escala de valores pode assumir qualquer valor no intervalo de [-1, 1] e sua fórmula é definida por:

$$Laplace - Corrected - Confidence(A \Rightarrow C) = \frac{c_{AC} + 1}{c_{X} + 2} \qquad \textbf{(38)}$$

3.29 LEAST CONTRADICTION

Definido por *Azé and Kondratoff* [36]. *S*ua escala de valores pode assumir qualquer valor no intervalo de [-1, 1] e sua fórmula é definida por:

$$Least-Contradiction(A\Rightarrow C)=\frac{Supp(A\cup C)-Supp(A\cup\overline{C})}{Supp(C)}=\frac{P(A\cap C)-P(A\cap\overline{C})}{P(C)}$$

(39)

3.30 LERMAN SIMILARITY

Descrita por *Lerman et al.* [37]. *S*ua escala de valores pode assumir qualquer valor no intervalo de [0, 1] e sua fórmula é definida por:

$$Lerman-Similarity(A\Rightarrow C)=\sqrt{n}\,\frac{Supp(X\cup Y)-Supp(A)Supp(C)}{\sqrt{Supp(A)Supp(C)}}=\frac{c_{AC}-\frac{c_A c_C}{n}}{\sqrt{\frac{c_A c_C}{n}}}$$

(40)

3.31 LEVERAGE

Introduzido por *Piatestky-Shapiro* [11], este define a medida *Leverage* como uma das funções mais simples e satisfaz os princípios básicos para funções de interesse em uma regra de associação. Sua escala de valores pode assumir qualquer valor no intervalo de [-1, 1], sendo que:

- 0 se as variáveis são estatisticamente independentes.
- Aumenta monotonicamente se as variáveis ocorrem mais frequentemente juntas.
- Diminui monotonicamente se uma das variáveis sozinha ocorre com mais frequência.

Sua fórmula é definida por:

$$PS(A \Rightarrow C) = Supp(A \Rightarrow C) - Supp(A)\,Supp(C) = P(A \cap C) - P(A)P(C) \qquad \textbf{(41)}$$

Uma característica interessante é que o uso de limites mínimos de *Leverage* incorpora ao mesmo tempo uma restrição de frequência implícita. Exemplificando, para definir um limite min. de *Leverage* para 0,01% (corresponde a 10 ocorrências em um conjunto de dados com 100.000 transações), primeiro é possível usar um algoritmo para localizar todos os conjuntos de itens com min. suporte de 0,01% e, em seguida, filtre os conjuntos de itens encontrados usando esta restrição de *Leverage*. Devido a essa propriedade, a métrica *Leverage* também sofre com o problema associado a itens raros (itens raros são itens que possuem baixa frequência na base transacional, ou seja, que integram *itemsets* com baixo suporte).

3.32 LIFT

A métrica *Lift* foi originalmente definida por *Brin et al*. [22] como *Interest*, e sua escala de valores é definida no intervalo [0, +∞). A medida *Lift* mede quantas vezes mais *A* e *C* ocorrem juntos do que esperado se fossem estatisticamente independentes. Seu objetivo é medir o desvio da independência estatística, definido como:

• quando a elevação é exatamente 1: Sem efeito (independência de *A* e *C*). Nenhuma relação entre eventos.
• para elevação maior que 1: efeito positivo (dado que o *A* é verdadeiro, é mais provável que o *C* seja verdadeiro). Dependência positiva entre eventos.
• se a elevação for menor que 1: Efeito negativo (quando o *A* é verdadeiro, é menos provável que o *C* seja verdadeiro). Dependência negativa entre eventos.

Sua fórmula é definida por:

$$Lift(A \Rightarrow C) = \frac{Conf(A \Rightarrow C)}{Supp(C)} = \frac{P(A \cap C)}{P(A)P(C)} \qquad \textbf{(42)}$$

Uma característica desta métrica é que não está fechada para baixo e não sofre com o problema do item raro, além disso *Lift* é suscetível a ruídos em bancos de dados pequenos. Conjuntos de itens raros com baixa contagem (baixa probabilidade) que por acaso venham a ocorrer algumas vezes (ou apenas uma vez) juntos, podem produzir enormes valores de *Lift*.

3.33 MAXCONFIDENCE

Definida por *Tan et al.* [17] como uma versão simétrica da métrica *Confidence* (confiança), a forma de medida nulo-invariável da métrica para análise de propriedades. Sua escala de valores é encontarda no intervalo [0, 1], sendo valores próximos a 0 menos interessantes (ou associados) e valores próximos a 1 mais interessantes (ou associados).

Sua fórmula é definida por:

$$Max\,Confidence(A \Rightarrow C) = max\{Conf(A \Rightarrow C), Conf(C \Rightarrow A)\} = max\{P(C \mid A), P(A \mid C)\}$$

(43)

3.34 MUTUAL INFORMATION

Descrita por *Tan et tal.* [17], como uma medida fortemente ligada a entropia. Entropia está relacionada à variação de uma distribuição de probabilidades. A entropia de uma distribuição uniforme é grande, enquanto a entropia de uma distribuição inclinada é pequena.

Mutual Information é uma medida baseada em entropia para avaliar as dependências entre variáveis. Representa a quantidade de redução na entropia de uma variável quando o valor de uma segunda variável é conhecida. Se as duas variáveis estiverem fortemente associadas, há quantidade de redução na entropia, ou seja, suas informações mútuas, é alta. Sua escala de valores é encontarda no intervalo [0, 1], sendo o valor 0 representando independência entre as variáveis *A*, *C.*

Sua fórmula é definida por:

$$M(A \Rightarrow C) = \frac{\sum_i \sum_j P(A_i, C_j) \log \frac{P(A_i, C_j)}{P(A_i)P(C_j)}}{min(-\sum_i P(A_i) \log P(A_i), -\sum_j P(B_j) \log P(B_j))} \quad \textbf{(44)}$$

$$M(A \Rightarrow C) = \frac{\sum_{i \in \{A, \overline{A}\}} \sum_{j \in \{Y, \overline{Y}\}} P(i \cap j) \log \frac{P(i \cap j)}{P(i)P(j)}}{min(-\sum_{i \in \{A, \overline{A}\}} P(i) \log P(i), -\sum_{j \in \{C, \overline{C}\}} P(j) \log P(j))} \quad \textbf{(45)}$$

3.35 ODDS RATIO

Odds ratio [17] (razão de chances ou razão de possibilidades) deriva de uma medida de intensidade de associação estatística antiga e representa as chances de obter os diferentes resultados de uma variável. É o valor que define as chances de encontrar o antecedente em transações que contêm o consequente, esua escala de valores é definida no intervalo [0, +∞).

Uma razão de chances de 1 indica que a condição ou regra sob estudo é igualmente provável de ocorrer nos dois grupos (a probabilidade de se encontrar o antecedente em transações que contêm o consequente é a mesma de se encontrar o antecedente em transações que não contêm o consequente), ou seja, não estão associados. Uma razão de chances maior do que 1 indica que a condição ou regra tem maior probabilidade de ocorrer, ou seja, é mais interessante. Uma razão de chances menor do que 1 indica que a probabilidade é menor no primeiro grupo do que no segundo, ou seja, a chance de encontrar o antecedente em transações onde não contém o consequente é maior.

Sua fórmula é definida por:

$$Odds-Ratio(A \Rightarrow C) = \alpha(A \Rightarrow C) = \frac{c_{AC} c_{\overline{A}\,\overline{C}}}{c_{A\,\overline{C}} c_{\overline{A}\,C}} = \frac{P(A \cap C) P(\overline{A} \cap \overline{C})}{P(A \cap \overline{C}) P(\overline{A} \cap C)} \qquad \textbf{(46)}$$

3.36 Φ CORRELATION COEFFICIENT

A medida *Φ Correlation coefficient* [17] é análoga ao coeficiente de correlação produto-momento de Pearson para variáveis contínuas, que mede o grau de correlação entre duas variáveis. Ela está intimamente relacionada à estatística X^2(*Chi-squared*), pois $\Phi^2 = X^2 / N$. embora a estatística X^2 seja frequentemente usada para testes de qualidade do ajuste, ela raramente é usada como uma medida de associação, pois depende do tamanho do banco de dados.

Sua fórmula é definida por:

$$\phi(A \Rightarrow C) = \frac{Nc_{AC} - c_A - c_C}{\sqrt{c_A c_C c_{\overline{A}} c_{\overline{C}}}} = \frac{P(A,C) - P(A)P(C)}{\sqrt{P(A)P(C)(1-P(A))(1-P(C))}} \qquad \textbf{(47)}$$

Seu valor assume uma escala que varia entre:

- -1 → Significa uma correlação positiva perfeita entre as duas variáveis(as duas variam proporcionalmente juntas, ou seja, assumem o comportamento dinâmico igual. Quando uma aumenta a outra aumenta proporcionalmente).
- 0 →Significa que possuem comportamento independente uma da outra.
- 1 →Significa uma correlação negativa perfeita entre as duas variáveis(as duas variam de forma inversamente proporcional).

3.37 RALAMBONDRAINY MEASURE

Definido por *Diatta et tal.* [35] são definidas pelo autor como medidas de qualidade probabilística normalizada (PQM) para regras de associação; isto é, PQMs cujos valores se situam entre menos um e mais um, e que levam em conta situações de referência como incompatibilidade, repulsão, independência, atração e implicação lógica, entre o antecedente e o consequente das regras de associação.

Sua escala de valores é definida no intervalo [0, 1] e o valor mais próximo de 0 indica melhor interesse a regra. Sua fórmula é descrita por:

$$Ralambondrainy(A \Rightarrow C) = \frac{c_{A\overline{C}}}{N} = P(A \cap \overline{C}) \qquad \textbf{(48)}$$

3.38 RELATIVE LINKAGE DISEQUILIBRIUM

De acordo com *Kenett et al* [18], O *Relative Linkage Disequilibrium* pode ser considerado uma adaptação da métrica *Lift* com a vantagem de apresentar de maneira mais eficaz o desvio do suporte de toda a regra do suporte esperado sob independência (D = 0), dados os suportes do A e do C. Utiliza-se uma Variável D (chamada de Desequilíbrio de ligação), que é calculada através da seguinte fórmula:

$$D = Supp(AC)\,Supp(\overline{A}\,\overline{C}) - Supp(A\overline{C})\,Supp(\overline{A}C) \qquad \textbf{(49)}$$

O *Relative Linkage Disequilibrium* avalia o desvio do suporte de toda a regra, do suporte esperado sob independência, considerando os suportes do *A* e do *C*.O *Relative Linkage Disequilibrium* assume um valor que varia de 0 a 1 e o seu cálculo pode ser realizado através do seguinte algoritmo:

Se $D > 0$

Então Se $Supp\ \overline{A}C < Supp\ A\overline{C}$

Então *Relative Linkage Disequilibrium* = $D / (D + Supp\ \overline{A}C)$

Senão *Relative Linkage Disequilibrium* = $D / (D + Supp\ A\overline{C})$

Senão Se $Supp\ AC < Supp\ \overline{A}\overline{C}$

Então *Relative Linkage Disequilibrium* = $D / (D - Supp\ AC)$

Senão *Relative Linkage Disequilibrium* = $D / (D - Supp\ \overline{A}\overline{C})$

3.39 RULE POWER FACTOR

Ochin et al. [20] descreve a métrica *Rule Power Factor* como um "superconjunto da medida de confiança". A medida *Rule Power Factor* foca na importância da associação entre antecedentes e consequentes de regras de associação. A métrica pesa a confiança de uma regra por seu suporte. Sua escala de valores é definida no intervalo [0, 1] e sua fórmula é descrita por:

$$RPF(A \Rightarrow C) = Supp(A \cup C) * Conf(A \cup C) \qquad \textbf{(50)}$$

3.40 SEBAG-SCHOENAUER MEASURE

Sua escala de valores é definida no intervalo [0, 1] e sua fórmula é descrita por:

$$Seabag(A \Rightarrow C) = \frac{Supp(A \cup C)}{Supp(A \cup \overline{C})} = \frac{P(A \cap C)}{P(A \cap \overline{C})} \qquad \textbf{(51)}$$

3.41 VARYING RATES LIAISON

Descrito por *Bernard et al.* [19], com o objetivo de valorar relações de dependência orientada entre variáveis binárias através de sua "Análise Implicativa Bayesiana". Seu valor fica definido no intervalo [-1, $+\infty$) (sendo o valor 0 significando independência orientada de $A \Rightarrow C$).

Sua fórmula é definida por:

$$VRL(A \Rightarrow C) = \frac{Supp(A \cup C)}{Supp(A)\,Supp(C)} - 1 = \frac{P(A \cap C)}{P(A)P(C)} - 1 = Lift(A \Rightarrow C) - 1 \qquad \textbf{(52)}$$

Sendo Lift ($A \Rightarrow C$) o valor da métrica *Lift* **(42)** para está regra

3.42 YULE'S Q AND YULE'S Y

As métricas *Yule's Q & Y* [17] são métricas de identificação de intensidade de associação entre as variáveis(antecedente e consequente). Os coeficientes Q e Y de Yule são variantes normalizadas do odds ratio, definidas de uma maneira que variam de -1 a +1.

Suas fórmulas são definidas por:

$$Q(A \Rightarrow C) = \frac{\alpha - 1}{\alpha + 1} \qquad \textbf{(53)}$$

e

$$Y(A \Rightarrow C) = \frac{\sqrt{\alpha} - 1}{\sqrt{\alpha} + 1} \qquad \textbf{(54)}$$

, onde α é o valor de *Odds-ratio*.

Suas escalas de valor podem ser entendidas da seguinte forma:

- -1 → Significa uma associação positiva perfeita entre as duas variáveis(as duas variam proporcionalmente juntas, ou seja, assumem o comportamento dinâmico igual. Quando uma aumenta a outra aumenta proporcionalmente).
- 0 →Significa que não possuem associação.
- 1 →Significa uma associação negativa perfeita entre as duas variáveis(as duas variam de forma inversamente proporcionais.

4 ANÁLISE EXPERIMENTAL DAS MÉTRICAS CALCULADAS

Descrevemos anteriormente um grupo de métricas de associação, uma versão resumida das métricas apresentadas se encontram no Anexo A, mas existe um número extenso de métricas disponíveis para identificar padrões de associação e um grande número dessas métricas fornece conflito de informações entre si, pois cada métrica é construída com propriedades específicas que revelam informações de acordo com características identificadas como interessantes para um determinado objetivo específico. O grande desafio é identificar qual métrica está alinhada com o interesse de identificação de padrões almejado, pois cada domínio de problema exige um conjunto de características específicas. Desta forma, ao se utilizar a métrica correta para o problema, consegue-se revelar os padrões mais interessante ao objetivo traçado.

Neste capítulo buscamos demonstrar visualmente a relação entre as métricas supracitadas. Abaixo descrevemos algumas propriedades que nos orientaram neste objetivo.

Piatetsky-Shapiro [11] propôs 3 principais propriedades que toda métrica objetiva deve satisfazer:

- (P1) M = 0 se A e C são estatisticamente independentes, ou seja, $P(AC) = P(A)P(C)$. No entanto, alguns pesquisadores relaxaram essa propriedade, portanto, se A e C são independentes, basta que o valor para M seja uma constante [12].
- (P2) M aumenta monotonicamente com $P(AC)$ quando $P(A)$ e $P(C)$ permanecem os mesmos;
- (P3) M diminui monotonicamente com $P(A)$ (ou $P(C)$) quando o restante dos parâmetros ($P(AC)$ e $P(A)$ (ou $P(C)$)) permanecem inalterados.

O princípio (P1) afirma que uma regra de associação que ocorre por acaso não é interessante. Um valor maior que 1 indica uma correlação positiva e um valor menor que 1 indica uma correlação negativa.

O princípio (P2) afirma que, quanto maior o suporte para *AC*, maior o valor de interesse quando o suporte para *A* e *C* é fixo, ou seja, quanto mais correlação positiva *A* e *C* tiver, mais interessante será a regra.

O princípio (P3) afirma que, se os suportes para *AC* e *C* (ou *A*) forem fixos, quanto menor o suporte para *A* (ou *C*), mais interessante será o padrão. De acordo com os princípios (P2) e (P3), quando os suportes de *A* e *C* são idênticos ou o suporte de A contém o suporte de *C* (ou vice-versa), a medida de interesse deve atingir seu valor máximo.

Major e Mangano [13] adicionaram uma quarta propriedade às três acima:

- (P4) M aumenta monotonicamente com $P(A)$ quando $P(C)$, $P(\neg C)$ e a confiança da regra ($P(C|A)$) permanecem os mesmos. Essa propriedade afirma que, se a confiança de uma regra permanecer fixa, então, quanto maior o suporte de A, mais interessante será a regra.

Tan et al. [12] propuseram cinco propriedades baseadas em operações para tabelas de contingência 2×2 (**figura 1**), que seriam:

- (O1) M deve ser simétrico sob permutação variável.
- (O2) M deve ser o mesmo quando dimensionamos qualquer linha ou coluna por um fator positivo.
- (O3) M deve se tornar – M se as linhas ou colunas forem permutadas, ou seja, trocar as linhas ou colunas na tabela de contingência faz com que os valores de interesse alterem seus sinais.
- (O4) M deve permanecer o mesmo se ambas as linhas e colunas forem permutadas.
- (O5) M não deve ter relação com a contagem dos registros que não contêm *A* e *C*.

Diferentemente dos princípios de *Piatetsky-Shapiro*, essas propriedades não devem ser interpretadas como declarações do que é desejável, de acordo com *Tan et* al. [12] o objetivo é que eles sejam usados para classificar as medidas em diferentes grupos.

A propriedade (O1) afirma que as regras $A \Rightarrow C$ e $C \Rightarrow A$ devem ter os mesmos valores de interesse, o que não é verdade para muitas aplicações. Por exemplo, confiança representa a probabilidade de um consequente, dado o antecedente, mas não vice-versa. Portanto, é uma medida assimétrica. Para fornecer medidas simétricas adicionais, *Tan et al.* [12] Transformaram cada medida assimétrica M em simétrica, tomando o valor máximo de M ($A \Rightarrow C$) e M ($C \Rightarrow A$). Por exemplo, eles definiram uma medida de confiança simétrica como $\max(P(C|A), P(A|C))$.

A propriedade (O2) requer invariância com o dimensionamento de linhas ou colunas.

A propriedade (O3) afirma que M ($A \Rightarrow C$) = −M ($A \Rightarrow \neg C$) = −M ($\neg A \Rightarrow C$). Essa propriedade significa que a medida pode identificar correlações positivas e negativas.

A propriedade (O4) afirma que M ($A \Rightarrow C$) = M ($\neg A \Rightarrow \neg C$).

A propriedade (O3) é, de fato, um caso especial da propriedade (O4), porque se a permeação das linhas (colunas) faz com que o sinal seja alterado uma vez e a permuta das colunas (linhas) faz com que seja alterada novamente, o resultado geral da permeação de ambas as linhas e colunas será para deixar o sinal inalterado.

A propriedade (O5) afirma que a medida deve levar em consideração apenas o número de registros que contêm A, C ou ambos. M deve permanecer o mesmo com a contagem de registros que não contêm A, C ou ambos. Apoio, suporte não satisfazem essa propriedade, enquanto a confiança o satisfaz.

Lenca et al. [14, 15] propuseram cinco propriedades para avaliar medidas de associação:

- (L1) M é constante se não houver contraexemplo à regra. Essa propriedade afirma que, se $P(A\neg C) = 0$, o valor da medida deve ser constante ou infinito. Em outras palavras, se a confiança de uma regra é uma, a medida deve ter o mesmo valor de interesse, independentemente do suporte. De alguma forma, essa propriedade está em contradição com a propriedade (P4), que afirma que, se a confiança é fixa, quanto maior o suporte, mais interessante é a regra.
- (L2) M diminui com $P(A\neg C)$ de maneira linear, convexa ou côncava em torno de 0+. Essa propriedade descreve a maneira como uma medida diminui quan-

do alguns contraexemplos são adicionados. a maneira desejada depende do domínio do problema e do usuário. Se alguns registros de contraexemplo puderem ser tolerados, é desejável uma diminuição côncava. Se for necessária uma confiança estrita de 1, é desejada uma diminuição convexa.

- (L3) M aumenta à medida que o número total de registros aumenta. Essa propriedade afirma que o valor da medida aumenta com N (número total de registros), assumindo que $P(A)$, $P(C)$ e $P(AC)$ permaneçam fixos.
- (L4) O limite é fácil de corrigir. Para uma medida que possui essa propriedade, é fácil encontrar um limite que possa separar as regras interessantes das desinteressantes.
- (L5) A semântica da medida é fácil de expressar. Esta propriedade descreve que a semântica da medida é facilmente compreensível pelo usuário.

Geng e Hamilton [16] também propuseram duas propriedades para avaliar a relação entre uma medida, suporte e confiança:

- (G1) M deve ser uma função crescente de suporte se as margens na tabela de contingência forem fixas. Supondo que as margens da tabela de contingência sejam fixas (ou seja, $c_A = a$, $c_{\neg A} = (N - a)$, $c_C = b$; $c_{\neg C} = (N - b)$, se houver suporte igual a x então $P(AC) = x$, $P(\neg AC) = (b/N) - x$, $P(A\neg C) = a/N - x$ e $P(\neg A\neg C) = 1 - (a+b)/N + x$. Substituindo essas fórmulas em medidas, são obtidas funções com a variável x. A função deve aumentar em x. Esta propriedade é igual à propriedade (P2).
- (G2) M deve ser uma função crescente de confiança se as margens na tabela de contingência forem fixas. Como (G1), assumindo que as margens da tabela de contingência sejam fixas e a confiança seja igual a y, então $P(AC) = ay/N$, $P(\neg AC) = (b-ay)/N$, $P(A\neg C) = a(1-y)/N$ e $P(\neg A\neg C) = 1 - (a+b)/N + ay/N$. Novamente, substituindo essas fórmulas em medidas, são obtidas funções com a variável y. A função deve aumentar em y.

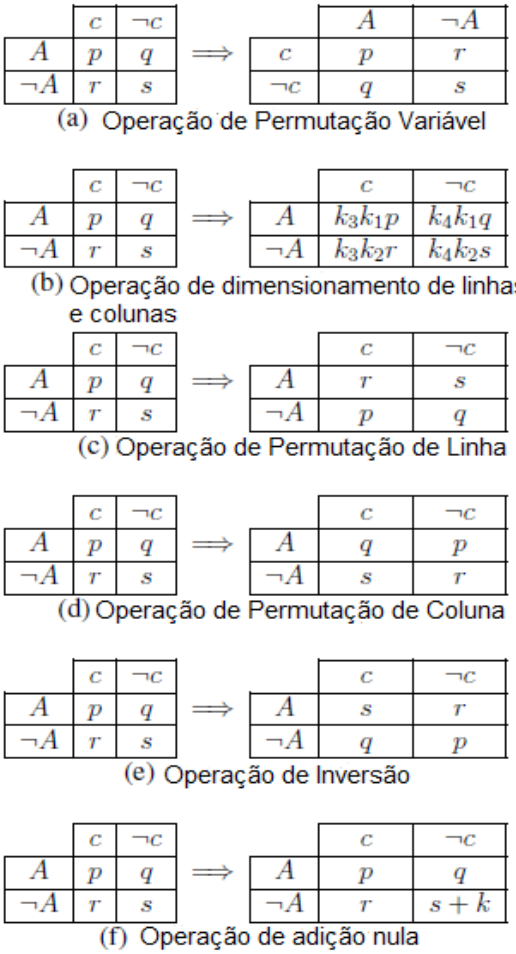

Figura 1: Operações em uma tabela de contingência [17]

A **tabela 16** mostra as 14 propriedades mencionadas para todas as 44 medidas relacionadas neste trabalho. As propriedades L4 e L5 não estão incluídas porque são propriedades subjetivas e dependem do domínio do usuário e do problema.

Para a propriedade P3, “A” significa que a medida diminui monotonicamente com $P(A)$, “C” significa que diminui monotonicamente com $P(C)$, “B” significa que diminui monotonicamente com $P(A)$ e $P(C)$ e “N” significa que não diminui com $P(A)$ ou $P(C)$. Para a propriedade O3, “R” significa que o sinal da medida muda com permutação de linha, “C” significa que o sinal da medida muda com permutação de coluna, “B” significa que o sinal da medida muda com permutações de linha e coluna e "N" significa que o sinal de medida não muda com as permutações de linha ou coluna. Para a propriedade L2, são utilizados números de 0 a 6, que representam respectivamente diminuição convexa, diminuição linear, diminuição côncava, diminuição mas a maneira depende dos parâmetros, invariável, aumento e depende dos parâ-

metros. Para as propriedades G1 e G2, são utilizados números de 0 a 3, que representam respectivamente aumento, invariante, diminuição e depende dos parâmetros. Para outras propriedades, “Y” significa que a medida possui essa propriedade e “N” significa que a medida não possui essa propriedade.

Tabela 16: Propriedades de métricas objetivas para regras de associação

Métricas	**P1**	**P2**	**P3**	**P4**	**O1**	**O2**	**O3**	**O4**	**O5**	**L1**	**L2**	**L3**	**G1**	**G2**
Support	N	Y	N	Y	Y	N	N	N	N	N	1	N	0	0
Confidence	N	Y	C	N	N	N	N	N	Y	Y	1	N	0	0
Added - Value	Y	Y	B	N	N	N	N	N	N	N	1	N	0	0
All-Confidence	-	-	-	-	-	-	-	-	-	-	-	-	-	
Casual-Conf	N	Y	B	N	N	N	N	N	N	Y	0	N	0	0
Casual-Supp	N	Y	B	N	Y	N	N	Y	N	N	1	N	0	0
Loevinger	Y	Y	B	N	N	N	N	N	N	Y	0	N	0	0
Chi-Squared	Y	N	N	Y	Y	N	N	Y	N	Y	6	Y	3	3
X-Supp-Ratio	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Collect-Str	N	Y	A	N	N	N	N	N	N	N	6	N	0	0
Conviction	Y	Y	B	N	N	N	N	N	N	Y	0	N	0	0
Cosine	N	Y	B	Y	Y	N	N	N	Y	N	2	N	0	0
Coverage	-	-	-	-	-	-	-	-	-	-	-	-	-	-
DCC	N	Y	C	N	N	N	C	N	Y	Y	1	N	0	0
Diff-Conf	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ECE-Rate	N	Y	C	N	N	N	N	N	Y	Y	2	N	0	0
Fisher-ET	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Gini-Index	Y	N	N	Y	N	N	N	Y	N	N	6	N	3	3
Hyper-Conf	N	N	N	N	N	N	N	N	N	N	6	Y	3	3
Hyper-Lift	N	N	N	N	N	N	N	N	N	N	6	Y	3	3
Imb-Ratio	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Importance	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Improvement	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Jaccard-Coeff	N	Y	B	Y	Y	N	N	N	Y	N	1	N	0	0
J-Measure	Y	N	A	N	N	N	C	N	N	N	6	N	3	3

Kappa	Y	Y	B	N	Y	N	N	Y	N	N	3	N	0	0
Klosgen	Y	N	B	N	N	N	N	N	N	N	6	N	3	3
Kulezynski	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Predictive-Ass	Y	Y	N	N	Y	N	N	Y	N	N	6	N	3	3
Laplace-CC	N	Y	C	N	N	N	N	N	Y	N	1	Y	0	0
Least-Contrad	N	Y	C	N	N	N	N	N	Y	N	2	N	0	0
Lerman-Simil	-	-	-	-	-	-	-	-	-	-	-	-	-	
Leverage	N	Y	B	N	N	N	N	N	N	N	1	N	0	0
Lift	Y	Y	B	N	Y	N	N	N	N	N	2	N	0	0
MaxConf	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Mutual-Inform	Y	N	N	N	N	N	N	Y	N	N	6	N	3	3
Odds-Ratio	Y	Y	B	N	Y	Y	N	Y	N	Y	0	N	0	0
Corr-Coeff	Y	Y	B	N	Y	N	B	Y	N	N	3	N	0	0
Rala-Measure	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Relative-LD	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Rule-Power- F	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Sea-Scho-Me	N	Y	C	N	N	N	N	N	Y	Y	0	N	0	0
Vary-Rates-Li	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Yule Q and Y	Y	Y	B	N	Y	Y	B	Y	N	Y	3	N	0	0

Todas essas propriedades são introduzidas no contexto da mineração de regras de associação. Elas podem ser usadas para encontrar medidas semelhantes, com comportamento equivalente ou para encontrar a medida apropriada para um domínio de problema se as propriedades de medida necessárias para esse domínio forem conhecidas. Essas definições formais já foram utilizadas como parâmetro de comparação e agrupamento de métricas [12].

Este capítulo almeja comparar de maneira prática o comportamento dessas métricas, fazendo uso de regras de associação mineradas (com um suporte mínimo de 1.33) a partir uma base de teste com 16 atributos e 15 linhas.

Para cada regra de associação, foram calculadas 43 métricas (descritas no capítulo 3), gerando assim uma matriz onde cada linha representa uma regra de associação, e cada coluna representa uma métrica.

A covariância entre as colunas foi utilizada para mensurar a relação entre as métricas. O coeficiente de correlação de Pearson [44] foi calculado para cada par de métricas, gerando assim uma matriz de correlação. Como a matriz é simétrica (o eixo X é idêntico ao eixo Y), ela pode ser representada por uma matriz triangular sem que haja perda de informação, como apresentado **no gráfico 2**.

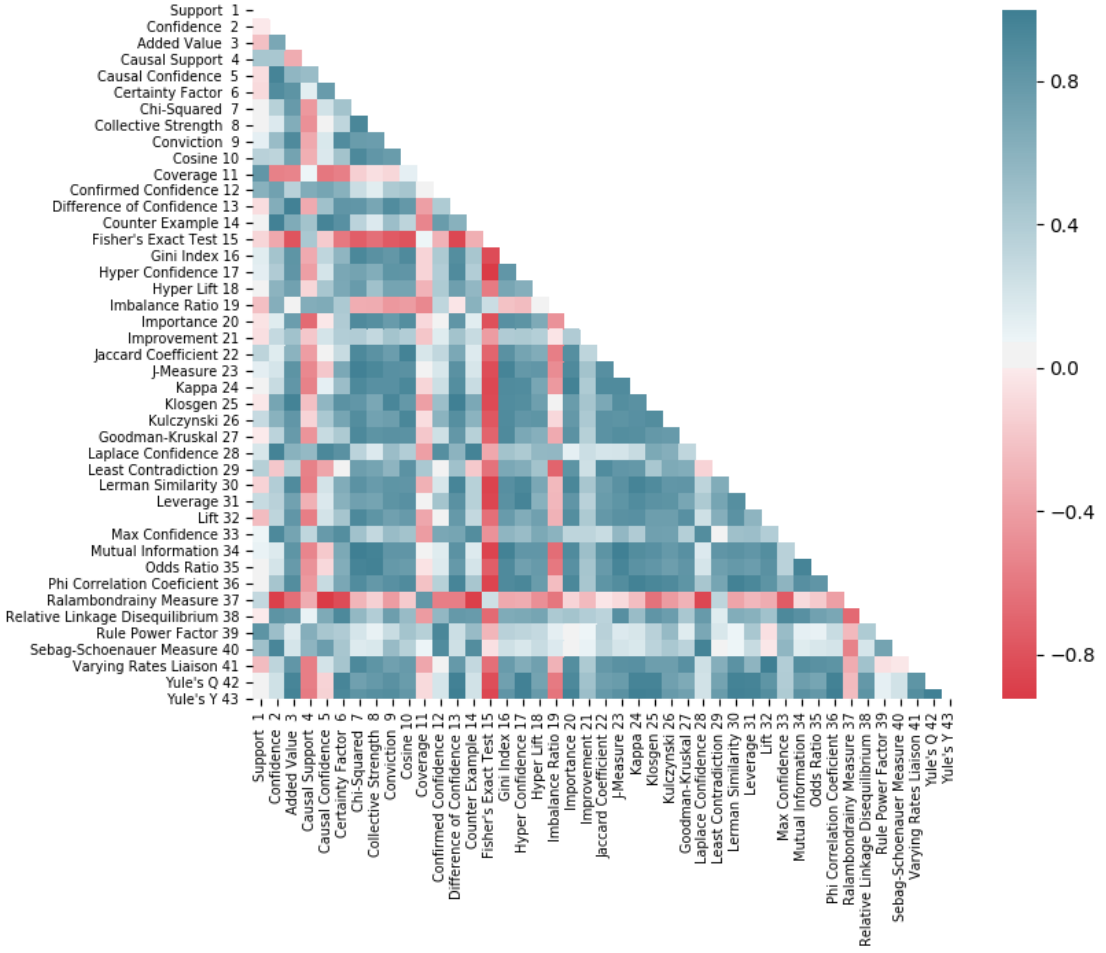

Gráfico 2: Matriz de correlação

É possível ainda agrupar as métricas através de métodos de clusterização hierárquica [45]. Utilizando a própria matriz de correlação como parâmetro de similaridade, podemos agrupar as métricas em *clusters*.

Para o **gráfico 3**, *clusterizamos* a matriz de correlação com o método *complete-linkage* [46]. Após a *clusterização*, cada *cluster* foi separadamente *sub-clusterizado,* para que se formassem grupos ainda menores.

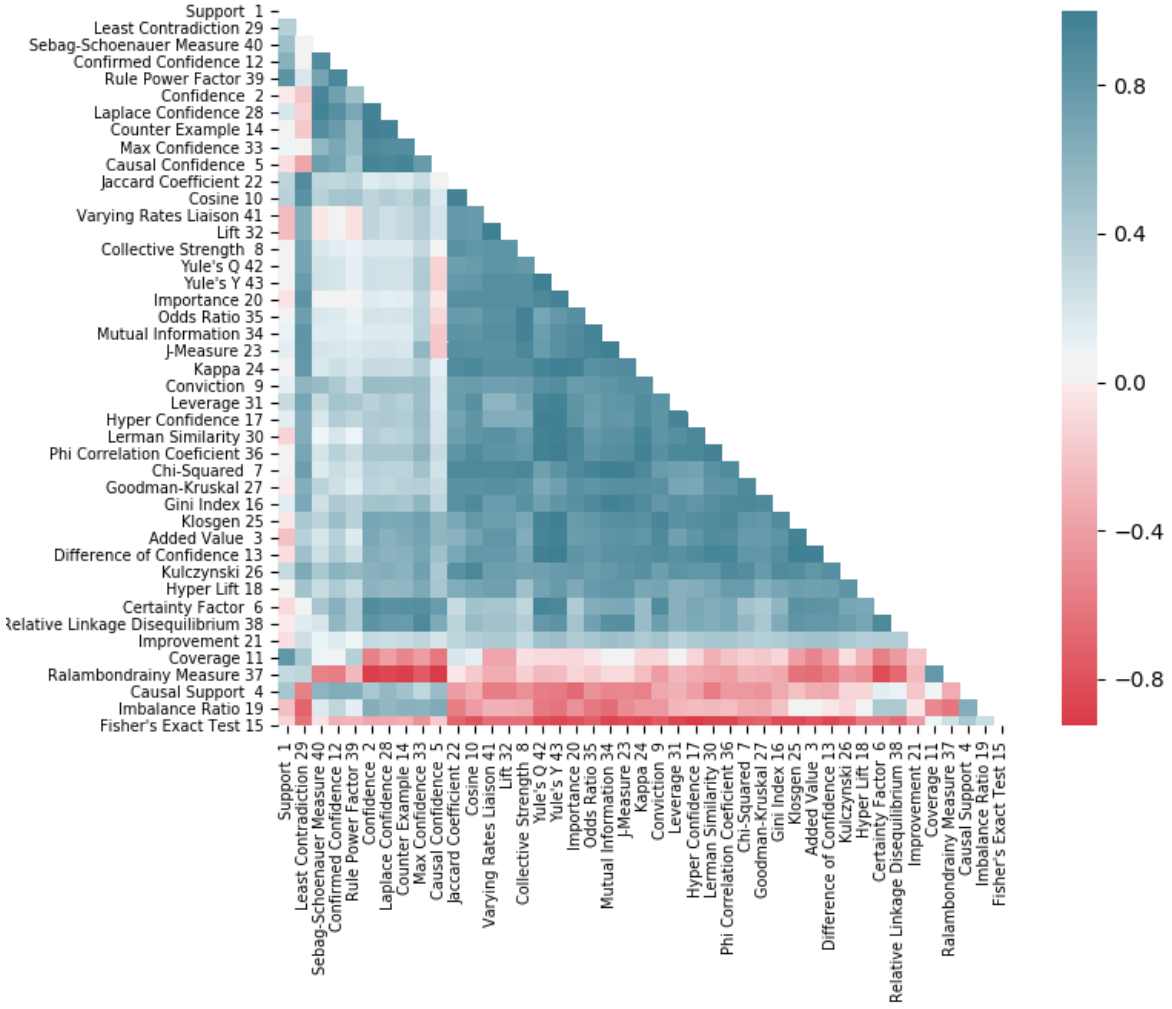

Gráfico 3: Matriz de correlação Clusterizada

Percebe-se que o suporte da regra aparenta estar fortemente correlacionado com apenas 5 outras métricas (Entre elas *Rule Power Factor*, *Coverage* e *Descriptive confirmed confidence*). O **gráfico 4** mostra as 30 métricas com maior módulo da correlação com o suporte da métrica. As correlações negativas foram evidenciadas em vermelho.

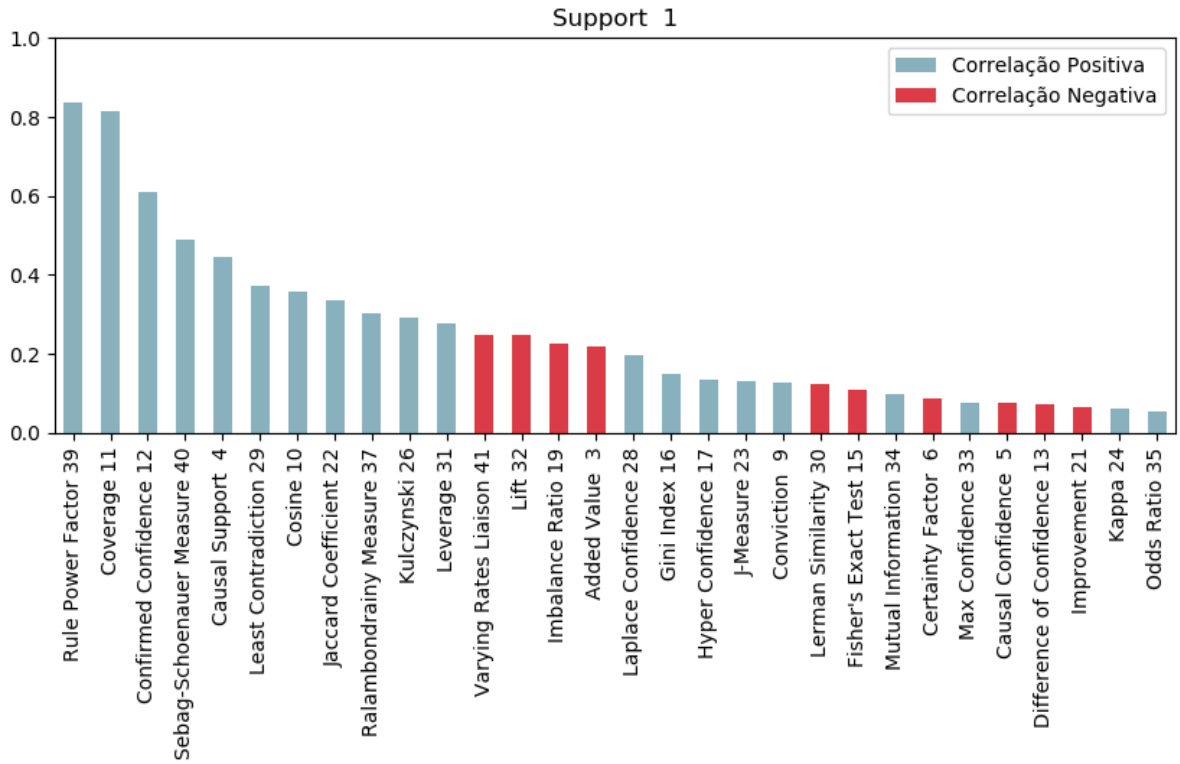

Gráfico 4: Correlação entre Suporte da regra e outras métricas

Observamos que o suporte possui ainda uma correlação negativa com algumas métricas, o que pode parecer contraintuitivo (dada por exemplo a definição da métrica *Added Value* **(equação 8)**).

O suporte é massivamente utilizado em algoritmos de mineração (tal como o *Apriori*) para filtrar e ranquear regras interessantes. Uma baixa covariância com outras métricas pode ser um indício de que podem haver outros parâmetros e critérios que podem ser utilizados na filtragem e ranqueamento de regras de associação, o que beneficiaria significantemente algoritmos de mineração.

5 CONCLUSÕES E TRABALHOS FUTUROS

Durante o projeto foi apresentado um conjunto de métricas que visam filtrar e ranquear regras de associação. Essas métricas são de extrema importância, visto o montante de regras de associação que podem ser mineradas dada uma pequena base de dados.

Através de uma busca na literatura, foram encontradas um total de 45 métricas que foram implementadas em Python. Alguns dos valores retornados pelas funções de cálculo das métricas (por exemplo infinito) não puderam ser representados. Nestes casos, as funções de cálculo retornam "*Not a number*".

Todas as métricas do nosso código que se encontravam disponíveis também no pacote aRules (R) foram validadas através de comparação dos valores calculados por ambos programas, para uma mesma base de dados.

Devido a dificuldades na interpretação semântica de algumas métricas, definimos uma análise de todas as métricas. Calculamos a correlação de Pearson par a par para todas as métricas, na tentativa de evidenciar padrões de comportamento que pudessem existir entre elas.

Para facilitar a interpretação, tivemos o intuito de agrupar as métricas em categorias, utilizando a correlação como parâmetro de similaridade. Assim, *clusterizamos* a matriz para tornar mais aparente o relacionamento entre as métricas definidas.

Para trabalhos futuros, pretendemos fazer o mesmo tipo de análise em diferentes bases de dados, comparar definições formais das métricas com seus resultados práticos, entender o significado semântico das regras, e entender melhor a relação do suporte com outras métricas. Pretendemos também fazer um estudo sobre os testes estatísticos utilizados neste trabalho como métricas de avaliação para determinação de dependência estatística entre o antecedente e o consequente das regras.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] AGRAWAL, R; IMIELINSKI, T; SWAMI, A. Mining Association Rules between Sets of Items in Large Databases. In: ACM SIGMOD CONFERENCE ON MANEGEMENT OF DATA, p. 207 – 216,Washington, DC, USA, 1993. ACM Press - New York, NY, USA.

[2] AGRAWAL, R; SRIKANT, R. Fast Algorithms for Mining Association Rules. In: Bocca, J. B; Jarke, M; Zaniolo, C, editors, PROC. 20TH INT. CONF. VERY LARGE DATA BASES, VLDB, p. 487–499, Washington, DC, USA, 12–15 1994. Morgan Kaufmann.

[3] EDWARD R. OMIECISNKI; Alternative interest measures for mining associations in databases. *IEEE Transactions on Knowledge and Data Engineering*, 15(1):57--69, Jan/Feb 2003.

[4] KODRATOFF, Y. (1999). Comparing Machine Learning and Knowledge Discovery in Databases: An Application to Knowledge Discovery in Texts. Lecture Notes on AI (LNAI) - Tutorial series

[5] BERZAL, FERNANDO, IGNACIO BLANCO, DANIEL SANCHEZ AND MARIA-AMPARO VILA (2002). Measuring the accuracy and interest of association rules: A new framework. Intelligent Data Analysis 6(3), 221-235.

[6] FISHER, RONALD A. (1922). "On the Interpretation of chi-squared from Contingency Tables, and the Calculation of P". Journal of the Royal Statistical Society. 85 (1): 87–94.

[7] AHMED, KHALIL M. AND EL-MAKKY, NAGWA M. AND TAHA, YOUSRY; A Note on "Beyond Market Baskets: Generalizing Association Rules to Correlations"; In: SIGKDD Explor. Newsl. January 2000.

[8] HUI XIONG, PANG-NING TAN, AND VIPIN KUMAR. Mining strong affinity association patterns in data sets with skewed support distribution. In Bart Goethals and Mohammed J. Zaki, editors, *Proceedings of the IEEE International Conference on Data Mining, November 19--22, 2003, Melbourne, Florida*, pages 387-394, November 2003.

[9] C. C. AGGARWAL AND P. S. YU. A new framework for itemset generation. Proceedings of the seventeenth ACM SIGACT-SIGMOT–SIGART, *Symposium on Principles of Database Systems* - in *PODS 1998,* pages 18--24, Seattle, WA, USA.

[10] SERGEY BRIN, RAJEEV MOTWANI, JEFFREY D. ULLMAN, AND SHALOM TSUR. Dynamic itemset counting and implication rules for market basket data. In- *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data*, pages 255--264, Tucson, Arizona, USA, May 1997.

[11] G. PIATETSKY-SHAPIRO. Discovery, analysis and presentation of strong rules, in: G. Piatetsky-Shapiro, W. Frawley (Eds.), Knowledge Discovery in Databases, MIT Press, Cambridge, MA, 1991, pp. 229–248.

[12] TAN, P., KUMAR, V., AND SRIVASTAVA, J. 2002. Selecting the right interestingness measure for association analysis. Department of Computer Science, University of Minnesota, 200 Union Street SE, Minneapolis, MN 55455, USA. Information Systems 29 (2004) 293–313 .

[13] J. A. MAJOR AND J. J. MANGANO. Selecting among rules induced from a hurricane database. Journal of Intelligent Information systems, 4:39–52, 1995.

[14] P. LENCA, P. MEYER, B. VAILLANT, AND S. LALLICH. A multicriteria decision aid for interestingness measure selection. Technical Report LUSSI-TR-2004-01-EN, LUSSI Department, GET/ENST, Bretagne, France, 2004.

[15] P. LENCA, B. VAILLANT, P. MEYER, AND S. LALLICH. Association rule interestingness measures: Experimental and theoretical studies. In Quality Measures in Data Mining, pages 51–76. Springer, 2007.

[16] L. GENG AND H. J. HAMILTON. Interestingness measures for data mining: A survey. ACM Comput. Surv., 38(3):9, 2006.

[17] P. TAN, V. KUMAR, AND J. SRIVASTAVA. Selecting the right objective measure for association analysis. Inf. Syst., 29(4):293–313, 2004.

[18] KENETT, R. & SALINI, S.. Relative Linkage Disequilibrium: A New Measure For Association Rules. Lectures Notes In Computer Science, 189-199, 2008.

[19] BERNARD, JEAN-MARC AND CHARRON, CAMILO. L'analyse implicative bayesienne, une methode pour l'etude des dependances orientees. II : modele logique sur un tableau de contingence Mathematiques et Sciences Humaines, Volume 135 (1996), p. 5-18, 1996.

[20] OCHIN, SURESH AND KUMAR , NISHEETH JOSHI. Rule Power Factor: A New Interest Measure in Associative Classification. 6th International Conference On Advances In Computing and Communications, ICACC 2016, 6-8 September 2016, Cochin, India.

[21] GOODMAN, L.A. AND KRUSKAL, W.H. Measures of association for cross-classifications, J. Am. Stat. Assoc. 49 (1968) 732–764.

[22] S. BRIN, R. MOTWANI AND C.SILVERSTEIN. Beyond market baskets: generalizing association rules to correlations, in: Proceedings of 1997 ACM-SIGMOD International Conference on Management of Data, Tucson, Arizona, June 1997, pp. 255–264.

[23] P. CLARK AND R. BOSWELL. Rule induction with cn2: some recent improvements, in: Proceedings of the European Working Session on Learning EWSL-91, Porto, Portugal, 1991, pp. 151–163.

[24] KULCZYNSKI, S. 1927.Die Pflanzenassoziationen der Pieninen. Bulletin International de l'Academie Polonaise des Sciences et des Lettres, Classe des Sciences Mathematiques et Naturelles B, 57-203.

[25] TIANYI WU, YUGUO CHEN AND JIAWEI HAN. Re-examination of interestingness measures in pattern mining: a unified framework. *Data Mining and Knowledge Discovery*, January 2010.

[26] W. KLÖSGEN, Problems for knowledge discovery in databases and their treatment in the statistics interpreter explora, Int. J. Intell. Systems 7 (7) (1992) 649–673.

[27] J. COHEN, A coefficient of agreement for nominal scales Educ. Psychol. Meas. 20 (1960) 37–46.

[28] P. SMYTH, R.M. GOODMAN. Rule induction using information theory, in: Gregory Piatetsky-Shapiro, William Frawley (Eds.), Knowledge Discovery in Databases, MIT Press, Cambridge, MA, 1991, pp. 159–176.

[29] C.J. VAN RIJSBERGEN, Information Retrieval, 2nd Edition, Butterworths, London, 1979.

[30] R.BAYARDO, R.AGRAWAL AND D.GUNOPULOS. Constraint-based rule mining in large, dense databases.*Data Mining and Knowledge Discovery*, 4(2/3):217--240, 2000.

[31] https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/microsoft-association-algorithm-technical-reference.

[32] L. BREIMAN, J. FRIEDMAN, R. OLSHEN, C. STONE, Classification and Regression Trees, Chapman & Hall, New York, 1984.

[33] HOFMANN, HEIKE AND ADALBERT WILHELM (2001). Visual comparison of association rules. Computational Statistics, 16(3):399-415.

[34] MERCERON, A. AND K. YACEF. Interestingness measures for association rules in educational data. In International Conference on Educational Data Mining, R. Baker and J. Beck, editors, Montreal, Canada, pp. 57–66, 2008

[35] J. DIATTA, H. RALAMBONDRAINY, AND A. TOTOTHASINA (2007). Towards a unifying probabilistic implicative normalized quality measure for association rules. In Quality Measures in Data Mining, 237-250, 2007.

[36] AZÉ, J. AND Y. KODRATOFF. Extraction de pepites de connaissances dans les donnees: Une nouvelle approche et une etude de sensibilite au bruit. In Mesures de Qualite pour la fouille de donnees. Revue des Nouvelles Technologies de l'Information, RNTI (2004).

[37] LERMAN, I. C.(1981). Classification et analyse ordinale des donnees. Paris.

[38] BATANERO, C.; ESTEPA, A.; GODINO, J.D.; GREEN, D. R. (1996). Intuitive Strategies and Preconceptions about Association in Contingency Tables. Journal for reaserch in Mathematics Education, 27(2), 151.

[39] GOLDSCHMIDT, R.; PASSOS, E. Data Mining: Um guia prático. Editora Campus, Rio de Janeiro (2005).

[40] KANG, J. H., YANG, D. H., PARK, Y. B., & KIM, S. B. (2012). A text mining approach to find patterns associated with diseases and herbal materials in oriental medicine. International Journal of Information and Education Technology, *2*(3), 224.

[41] WONG, K. W., ZHOU, S., YANG, Q., & YEUNG, J. M. S. (2005). Mining customer value: From association rules to direct marketing. Data Mining and Knowledge Discovery, 11(1), 57-79.

[42] ABBAS, W. F., AHMAD, N. D., & ZAINI, N. B. (2013, December). Discovering purchasing pattern of sport items using market basket analysis. In 2013 International Conference on Advanced Computer Science Applications and Technologies (pp. 120-125). IEEE.

[43] TAN, N. P.; KUMAR, V.; SRIVASTAVA, J.(2002). Selecting the right interestingness measure for association patterns. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '02). ACM, New York, NY, USA, 32-41.

[44] PEARSON, K.: Contributions to the mathematical theory of evolution, III, Regression, heredity, and panmixia. Philos. Trans. R. Soc. Lond. Ser. A 187, 253–318 (1896)

[45] DANIELI MULLNER, “Modern hierarchical, agglomerative clustering algorithms”, arXiv:1109.2378v1

[46] DANIEL DEFAYS. An efficient algorithm for a complete link method.The Computer Journal, 20(4):364–366, 1977. doi: 10.1093/comjnl/20.4.364.

ANEXOS A: RESUMO DAS MÉTRICAS APRESENTADAS

Métricas	*Range*	Fórmula
Suporte (*Support*)	[0, 1]	$Supp(X)=P(X)=\frac{\lvert\{t\in D; X\in t\}\rvert}{\lvert D\rvert}$ **(6)**
Confiança (*Confidence, Strength*)	[0, 1]	$Conf(A\Rightarrow C)=P(C\mid A)=\frac{Supp(A\Rightarrow C)}{Supp(A)}$ **(7)** ou $Conf(A\Rightarrow C)=\frac{Supp(A\cup C)}{Supp(A)}=\frac{P(A\cap C)}{P(A)}$ **(7)**
Confiança Centrada (*Added Value, Pavillon Index, Centered Confidence*)	[-0.5, 1]	$A-V(A\Rightarrow C)=Conf(A\Rightarrow C)-Supp(C)$ **(8)**
Confiaça Total (*All-Confidence*)	[0, 1]	$AC(X)=\frac{Supp(X)}{max(i)_{i\in X}(Supp(i))}=\frac{P(X)}{max(i)_{i\in X}(P(i))}$ **(9)** ou $All-Conf(X)=min\{P(A\mid C),P(C\mid A)\}$ **(9)**
Confiança Causal (*Causal Confidence*)	[0, 1]	$Casual-Conf=1/2\cdot[Conf(A\Rightarrow C)+Conf(\overline{A}\Rightarrow\overline{C})]$ **(11)** ou $Casual-Conf=1/2\cdot[P(C\mid A),P(\overline{C}\mid\overline{A})]$ **(11)**
Suporte Causal (*Causal*	[0, 2]	$Casual\ Supp=Supp(A\cup C)+Supp(\overline{A}\cup\overline{C})$ **(10)** ou $Casual\ Supp=P(A\cap C)+P(\overline{A}\cap\overline{C})$ **(10)**

Support)		
Fator de Certeza (*Certainty Factor, Loevinger*)	[-1, 1]	$$Certainty-Factor(A\Rightarrow C)=\frac{Conf(A\Rightarrow C)-Supp(C)}{Supp(\overline{C})}$$ **(12)** ou $$Certainty-Factor(A\Rightarrow C)=\frac{P(C\mid A)-P(C)}{1-P(C)}$$ **(12)**
Qui-Quadrado (*Chi-Squared*)	[0, ∞]	$$X^2=\frac{\sum_{i=0}^{1}\sum_{j=0}^{1}(N_{ij}-E_{ij})^2}{E_{ij}}$$ **(13)**
Taxa de Suporte Cruzado (*Cross-Support Ratio*)	[0, 1]	$$Cross-Support(X)=\frac{min_{i\in X}(Supp(i))}{max_{i\in X}(Supp(i))}$$ **(15)**
Força Coletiva (*Collective Strength*)	[0, ∞]	$$Collective-Strength=\frac{1-v(X)}{1-E[v(x)]}\times\frac{E[v(X)]}{v(X)}$$ **(16)** ou $$Collective-Strength=\frac{P(A\cap C)+P(\overline{C}\mid\overline{A})}{P(A)P(C)+P(\overline{A})P(\overline{C})}$$ **(16)**
Convicção (*Conviction*)	[0, ∞]	$$Conviction(A\Rightarrow C)=\frac{1-Supp(C)}{1-Conf(A\Rightarrow C)}$$ **(17)** ou $$Conviction(A\Rightarrow C)=\frac{P(A)P(\overline{C})}{P(A\cap\overline{C})}$$ **(17)**
Coseno (*Cosine*)	[0, 1]	$$Cosine(A\Rightarrow C)=\frac{Supp(A\cup C)}{\sqrt{Supp(A)Supp(C)}}$$ **(18)** ou $$C(A\Rightarrow C)=\frac{P(A\cap C)}{\sqrt{P(A)P(C)}}=\sqrt{P(A\mid C)\times P(C\mid A)}$$ **(18)**
Cobertura (*Coverage*)	[0, 1]	$$Coverage(A\Rightarrow C)=Supp(A)=P(A)$$ **(19)**
Confiança Descritiva	[-1, 1]	$$Descriptive-Conf=Conf(A\Rightarrow C)-Conf(A\Rightarrow\overline{C})$$ **(20)** ou

Confirmada (*Descriptive Confirmed Confidence*)		$Descriptive-Conf=P(C\|A)-P(\overline{C}\|A)$ **(20)**
Diferença de Confiança (*Difference of Confidence*)	[-1, 1]	$Diff-Conf(A\Rightarrow C)=Conf(A\Rightarrow C)-Conf(\overline{A}\Rightarrow C)$ **(21)** ou $Diff-Conf(A\Rightarrow C)=P(C\|A)-P(C\|\overline{A})$ **(21)**
Exemplo e Taxa de Contra-Exemplo (Example and Counter-Example Rate)	[0, 1]	$ECR(A\Rightarrow C)=\frac{P(A\cap C)-P(A\cap\overline{C})}{P(A\cap C)}$ **(22)**
Teste Exato de *Fisher* (*Fisher's Exact Test*)	[0, 1]	$p(A\Rightarrow C)=\frac{(c_{AC})!\times(c_{A\overline{C}})!\times(c_{\overline{A}C})!\times(c_{\overline{A}\overline{C}})!}{(c_A)!\times(c_{\overline{A}})!\times(c_C)!\times(c_{\overline{C}})!\times N!}$ **(23)**
Índice *Gini* (*Gini Index*)	[0, 1]	$Gini(A\Rightarrow C)=P(A)[P(C\|A)^2+P(\overline{C}\|A)^2]+P(\overline{A})[P(C\|\overline{A})^2+P$ **(24)**
Hiperconfiança (*Hyper_Confidence*)	[0, 1]	$Hy-Conf=(A\Rightarrow C)=1-P[C_{AC}\geq c_{AC}\|c_A,c_C]$ **(25)**
Hiper-*Lift* (*Hyper-Lift*)	[0, ∞]	$Hy-Lift(A\Rightarrow C)=\frac{c_{AC}}{Q_\delta[C_{AC}]}$ **(26)**
Taxa de Desequilibrio (*Imbalance Ratio*)	[0, 1]	$IR(A\Rightarrow C)=\frac{\lvert Supp(A)-Supp(C)\rvert}{Supp(A)+Supp(C)-Supp(A\cup C)}$ **(27)** Ou $IR(A\Rightarrow C)=\frac{\lvert P(A\|C)-P(C\|A)\rvert}{P(A\|C)+P(C\|A)-P(A\|C)P(C\|A)}$ **(27)**

Importância (*Importance*)	[-∞,∞]	$Importance(A \Rightarrow C) = \log_{10}(L(A \Rightarrow C)/L(A \Rightarrow \overline{C}))$ **(28)** Onde *L*= Laplace Correctd Confidence
Melhoria (*Improve-ment*)	[0, 1]	$Imp(A \Rightarrow C) = min_{A' \subset A}(Conf(A \Rightarrow C) - Conf(A' \Rightarrow C))$ **(29)**
Coeficiente de *Jacard* (*Jaccard Coefficient*)	[-1, 1]	$jaccard(X \Rightarrow Y) = \frac{Supp(X \cup Y)}{Supp(X) + Supp(Y) - Supp(X \cup Y)}$ **(30)** ou $jaccard(X \Rightarrow Y) = \frac{P(X \cap Y)}{P(X) + P(Y) - P(X \cap Y)}$ **(30)**
Medida *J* (*J-Measure*)	[0, 1]	$J(A \Rightarrow C) = P(A \cap C)\log(\frac{P(C \mid A)}{P(C)}) + P(A \cap \overline{C})\log(\frac{P(\overline{C} \mid A)}{P(\overline{\overline{C}})})$ **(31)**
Kappa	[-1, 1]	$K(A \Rightarrow C) = \frac{P(A \cap C) + P(\overline{A} \cap \overline{C}) - P(A)P(C) - P(\overline{A})P(\overline{C})}{1 - P(A)P(C) - P(\overline{A})P(\overline{C})}$ **(32)**
Klosgen	[-1, 1]	$Kn(A \Rightarrow C) = \sqrt{Supp(A \cup C)}(Conf(A \Rightarrow C) - Supp(C))$ **(33)** ou $Klosgen(A \Rightarrow C) = \sqrt{P(A \cap C)}(P(C \mid A) - P(C))$ **(23)**
Kulezynski	[0, 1]	$Kulc(A \Rightarrow C) = \frac{1}{2}(Conf(A \Rightarrow C) + Conf(C \Rightarrow A))$ **(34)** ou $Kulc(A \Rightarrow C) = \frac{1}{2}(\frac{Supp(A \cup C)}{Supp(A)} + \frac{Supp(A \cup C)}{Supp(C)})$ **(34)** ou $Kulc(A \Rightarrow C) = \frac{1}{2}(P(A \mid C) + P(C \mid A))$ **(35)**
Associação Preditiva (*Goodman-Kruskal, Predictive Association*)	[0, 1]	$\lambda(A \Rightarrow C) = \frac{\sum_{a \in A} max_{c \in C} P(a \cap c) - max_{c \in C} P(C)}{n - max_{c \in C} P(C)}$ **(36)** ou $\lambda(A \Rightarrow C) = \frac{\sum_j max_k P(A_j, C_k) + \sum_k max_j P(A_j, C_k) - max_j$ $2 - max_j P(A_j) - max_k P(C_k$

		(37)
Confiança Corrigida por *LaPlace* (*Laplace Corrected Confidence*)	[0, 1]	$Laplace-Corrected-Conf(A\Rightarrow C)=\frac{c_{AC}+1}{c_X+2}$ **(38)**
Menos Contradição (*Least Contradicti-on*)	[-1, 1]	$Least-Contrad(A\Rightarrow C)=\frac{Supp(A\cup C)-Supp(A\cup\overline{C})}{Supp(C)}$ **(39)** ou $Least-Contradiction(A\Rightarrow C)=\frac{P(A\cap C)-P(A\cap\overline{C})}{P(C)}$ **(39)**
Similaridade de *Lerman* (*Lerman Similarity*)	[0, 1]	$L-S(A\Rightarrow C)=\sqrt{n}\frac{Supp(X\cup Y)-Supp(A)Supp(C)}{\sqrt{Supp(A)Supp(C)}}$ **(49)** ou $Lerman-Similarity(A\Rightarrow C)=\frac{c_{AC}-\frac{c_A c_C}{n}}{\sqrt{\frac{c_A c_C}{n}}}$ **(40)**
Alavanca-gem, Medida *Pia-tetsky-Shapi-ro* (*Leverage, Piatetsky-Shapiro Measure*)	[-1, 1]	$PS(A\Rightarrow C)=Supp(A\Rightarrow C)-Supp(A)Supp(C)$ **(41)** ou $Leverage(A\Rightarrow C)=P(A\cap C)-P(A)P(C)$ **(41)**
Elevador, Interesse, Juros (*Lift, Interest*)	[0, ∞]	$Lift(A\Rightarrow C)=\frac{Conf(A\Rightarrow C)}{Supp(C)}$ **(42)** ou $Lift(A\Rightarrow C)=\frac{P(A\cap C)}{P(A)P(C)}$ **(42)**

Confiança Máxima (*MaxConfidence*)	[0, 1]	$maxConf(A \Rightarrow C) = max\{Conf(A \Rightarrow C), Conf(C \Rightarrow A)\}$ **(43)** ou $maxConf(A \Rightarrow C) = max\{P(C \mid A), P(A \mid C)\}$ **(43)**
Informação Mútua, Incerteza (*Mutual Information, Uncertainty*)	[0, 1]	$M(A \Rightarrow C) = \frac{\sum_i \sum_j P(A_i, C_j) \log \frac{P(A_i, C_j)}{P(A_i)P(C_j)}}{min(-\sum_i P(A_i) \log P(A_i), -\sum_j P(B_j) \log P(B_j)}$ **(44)** ou $M(A \Rightarrow C) = \frac{\sum_{i \in \{A, \overline{A}\}} \sum_{j \in \{Y, \overline{Y}\}} P(i \cap j) \log \frac{P(i \cap j)}{P(i)P(j)}}{min(-\sum_{i \in \{A, \overline{A}\}} P(i) \log P(i), -\sum_{j \in \{C, \overline{C}\}} P(j) \log$ **(45)**
Taxa *Odds* (*Odds Ratio*)	[0, ∞]	$Odss - Ratio(A \Rightarrow C) = \frac{c_{AC} c_{\overline{A}\,\overline{C}}}{c_{A\overline{C}} c_{\overline{A}C}}$ **(46)** ou $\alpha(A \Rightarrow C) = \frac{P(A \cap C)P(\overline{A} \cap \overline{C})}{P(A \cap \overline{C})P(\overline{A} \cap C)}$ **(46)**
Coeficiente de Correlação (*Correlation Coefficient*)	[-1, 1]	$\phi(A \Rightarrow C) = \frac{Nc_{AC} - c_A - c_C}{\sqrt{c_A c_C c_{\overline{A}} c_{\overline{C}}}}$ **(47)** ou $\phi(A \Rightarrow C) = \frac{P(A, C) - P(A)P(C)}{\sqrt{P(A)P(C)(1 - P(A))(1 - P(C))}}$ **(47)**
Medida de *Ralambondrainy* (*Ralambondrainy Measure*)	[0, 1]	$Ralambondrainy(A \Rightarrow C) = \frac{c_{A\overline{C}}}{N}$ **(48)** ou $Ralambondrainy(A \Rightarrow C) = P(A \cap \overline{C})$ **(48)**
Desequilíbrio de Ligação	[0, 1]	$RLD(A \Rightarrow C) = P(A \cap C)P(\overline{A} \cap \overline{C}) - P(A \cap \overline{C})P(\overline{A} \cap C)$ **(49)**

Relativo (*Relative Linkage Disequilibrium*)		
Fator de Potência de Regra (*Rule Power Factor*)	[0, 1]	$RPF(A \Rightarrow C) = Supp(A \cup C) * Conf(A \cup C)$ **(50)**
Medida *Sebag-Schoenauer* (*Seabag-Schoenauer measure*)	[0, 1]	$Seabag(A \Rightarrow C) = \frac{Supp(A \cup C)}{Supp(A \cup \overline{C})} = \frac{P(A \cap C)}{P(A \cap \overline{C})}$ **(51)**
Ligação de Taxas Variáveis (*Varying Rates Liaison*)	[-1, ∞]	$VRL(A \Rightarrow C) = \frac{Supp(A \cup C)}{Supp(A)\,Supp(C)} - 1$ **(52)** ou $VRL(A \Rightarrow C) = \frac{P(A \cap C)}{P(A)P(C)} - 1$ **(52)** ou $VRL(A \Rightarrow C) = Lift(A \Rightarrow C) - 1$ **(52)**
Q de *Yule* e Y de *Yule* (*Yule's Q and Yule's Y*)	[-1 , 1]	$Q(A \Rightarrow C) = \frac{\alpha - 1}{\alpha + 1}$ **(53)** $Y(A \Rightarrow C) = \frac{\sqrt{\alpha} - 1}{\sqrt{\alpha} + 1}$ **(54)**