Estudo de Caso – Regras de Associação

Roteiro

- Definição de Regras de Associação
- Um exemplo de Regras de Associação
- Algoritmos para gerar Regras de Associação
- Estudo de Caso

Regras de Associação

Objetivo:

- Encontrar tendências que possam ser usadas para entender e explorar padrões de comportamento dos dados.

Exemplo:

- Uma Base de Dados de um supermercado teria como regra o fato de que 80% dos clientes que compram um produto *Q,* também adquirem, na mesma ocasião o produto *W.* Em que 80% é o fator de confiabilidade da regra.

Problema:

- Analisar um grande volume de conhecimento extraído no formato de regras.

Alguns conceitos importantes

Itemset

Conjunto de atributos ou itens ordenados lexicograficamente.

Exemplos:

{a,b,c}
{1,2,3}
{André,Marcio,Marcos}

Alguns conceitos importantes

Suporte

Indica a freqüência com que um *itemset* ou com que LHS e RHS ocorrem juntos no conjunto de dados.

Exemplos:

- *Itemset*:
  {de} - Suporte = 1
- Regra:
  ab -> c = {abc} – Suporte = 2

X1	X2	X3
a	b	c
d	e	f
a	b	c

Alguns conceitos importantes

Confiança

Indica a freqüência com que LHS e RHS ocorrem juntos em relação ao número total de registro em que LHS ocorre.

Exemplo: **ab -> c**

$$\text{confiança} = \frac{\text{suporte}(\{\text{LHS}\wedge\text{RHS}\})}{\text{suporte}(\text{LHS})}$$

$$\text{confiança} = \frac{\text{suporte}(\{abc\})}{\text{suporte}(\{ab\})} = \frac{2}{3} = 0.66$$

X1	X2	X3
a	b	c
a	b	e
a	b	c

Alguns conceitos importantes

Itemset freqüente

É um *itemset* com suporte maior ou igual a um suporte mínimo especificado pelo usuário.

Exemplo:

Suporte Mínimo = 2

{ab} – Suporte = 3

{ab} é um *itemset* freqüente

X1	X2	X3
a	b	c
a	b	e
a	b	c

Definição

Seja **D** uma Base de Dados composta por um conjunto de itens $\mathbf{A} = \{a_1, a_2, \ldots, a_m\}$ ordenados lexicograficamente e por um conjunto de transações $\mathbf{T}=\{t_1, t_2, \ldots, t_n\}$, em que, cada transação t_i é composta por um conjunto de itens tal que t_i está contido em **A**.

- Uma Regra de Associação é uma implicação na forma **LHS** ⟶ **RHS**, em que **LHS** e **RHS** estão contidos em **A** e a intersecção de **LHS** e **RHS** é vazia.
- A regra **LHS** ⟶ **RHS** ocorre no conjunto de transações **T** com confiança **c** se **c%** das transações em **T** em que ocorre **LHS** também ocorre **RHS**.
- A regra **LHS** ⟶ **RHS** tem suporte **s** se em **s%** das transações em **D** ocorre **LHS** ∪ **RHS**.

Síntese

Dado uma Base de Dados ***D*** composta por um conjunto de itens $\mathbf{A} = \{a_1, a_2, \ldots, a_m\}$ ordenados lexicograficamente e por um conjunto de transações $\mathbf{T}=\{t_1, t_2, \ldots, t_n\}$, em que, cada transação t_i é composta por um conjunto de itens tal que t_i está contido em **A**.

1. Encontrar todos os *itemsets* que possuem suporte maior que um suporte mínimo especificado pelo usuário (*itemsets* freqüentes).
2. Utilizar todos os *itemsets* freqüentes para gerar todas as regras de association que possuem confiança maior do que a confiança mínima especificada pelo usuário.

Exemplo

- A = {bermuda, calça, camiseta, sandália, tênis} e T = {1, 2, 3, 4}.
- Suporte Mínimo = 50% (2 transações).
- Confiança Mínima = 50%.

Transações	Itens Comprados
1	calça, camiseta, tênis
2	camiseta, tênis
3	bermuda, tênis
4	calça, sandália

Itemsets Freqüentes	Suporte
{tênis}	75%
{calça}	50%
{camiseta}	50%
{camiseta, tênis}	50%

Exemplo

- Suporte Mínimo = 50% (2 transações)
- Confiança Mínima = 50%

Itemsets Freqüentes	Suporte
{tênis}	75%
{calça}	50%
{camiseta}	50%
{camiseta, tênis}	50%

tênis ⟶ ***camiseta***, em q

$$\text{suporte} = \text{suporte}(\{\text{tênis, camiseta}\}) = 50\%$$

$$\text{confiança} = \frac{\text{suporte}(\{\text{tênis, camiseta}\})}{\text{suporte}(\{\text{tênis}\})} = \frac{50}{75} = 66{,}6\%$$

Exemplo

- Suporte Mínimo = 50% (2 transações)
- Confiança Mínima = 50%

Itemsets Freqüentes	Suporte
{tênis}	75%
{calça}	50%
{camiseta}	50%
{camiseta, tênis}	50%

camiseta ⟶ ***tênis***, em q

$$\text{suporte} = \text{suporte}(\{\text{tênis, camiseta}\}) = 50\%$$

$$\text{confiança} = \frac{\text{suporte}(\{\text{tênis, camiseta}\})}{\text{suporte}(\{\text{tênis}\})} = \frac{50}{50} = 100\%$$

Algoritmos para Regras de Associação

- *Apriori e AprioriTid* [Agrawal R. & R. Srikant (1994)];
- *Opus* [Webb G. I. (1995)];
- *Direct Hasing and Pruning* (DHP) [Adamo J.M.(2001)];
- *Dynamic Set Couting* (DIC) [Adamo J.M. (2001)];
- *Charm* [Zaki M. & C. Hsiao (2002)];
- *FP-growth* [J. Han, J. Pei & Y. Yin (1999)];
- *Closet* [Pei J., J. Han & R. Mao (2000)];

Considerações dos Algoritmos

- Os diversos algoritmos devem gerar sempre o mesmo conjunto de conhecimento.
- Mas qual é a diferença entre eles?
  - Forma como os dados são carregados na memória;
  - Tempo de processamento;
  - Tipos de atributos (Numéricos, Categóricos);
  - A maneira como os algoritmos geram os *itemsets.*

Estudo de Caso

- Descrição da Área
- Identificação do Problema
- Descrição do Conjunto de Dados
- Pré-Processamento
- Extração de Padrões
- Pós-Processamento

Descrição da Área

Os experimentos foram relizados tendo como foco um conjunto de variáveis de qualidade de vida consideradas na formulação do IQVU (Indicador da Qualidade de Vida Urbana) de uma cidade de médio porte - São Carlos/SP - utilizando recursos de geoprocessamento.

Cidade de São Carlos

Identificação do Problema

- O problema em questão é tentar encontrar relações entre os atributos que descrevem a qualidade de vida das pessoas enquanto moradoras da cidade.
- Com as relações é possível identificar a existência de um conjunto de determinantes relacionados às condições básicas de vida da população, constituindo um núcleo de aspectos fundamentais (alimentação, saúde, habitação, segurança, educação, emprego) que, uma vez assegurados, determinam o estabelecimento de um conjunto de possibilidades/oportunidades sociais para a população urbana.

Descrição do Conjunto de Dados

Formado por fatores mensuráveis (variáveis de natureza quantitativa) que se referem às características/recursos pessoais dos habitantes da cidade (grau de alfabetização, renda familiar etc.) e às características do lugar urbano (oferta local de infraestruturas e serviços urbanos, grau de acessibilidade espacial às atividades urbanas etc.)

Descrição da Base de Dados QVU

- Características:
  - Formado por 120 exemplos
  - Possui 19 atributos, todos contínuos
  - Atributo-meta: IQVU
  - Não existem valores ausentes

Descrição dos atributos da Base de Dados QVU

	Atributo	Descrição
1	IQVU	Indicador de qualidade de vida urbana
2	acessibi	Nível de acessibilidade às atividades urbanas
3	agua_p	Percentual de domicílios atendidos pelo sistema de abastecimento de água
4	cmd_dom	Número médio de cômodos por domicílio
5	denscons	Densidade construída (m^2/ha.)
6	densidad	Densidade populacional (hab./ha.)
7	dpi_p	Percentual de domicílios particulares improvisados
8	esgoto_p	Percentual de domicílios atendidos pelo sistema de coleta de esgoto sanitário
9	kilomete	Distância do n-ésimo setor ao setor central
10	lixo_p	Percentual de domicílios atendidos pelo sistema de coleta de resíduos sólidos
11	perocppr	Percentual de domicílios de propriedade do morador
12	pess_dom	Número médio de habitantes por domicílio
13	porpdimp	Percentual de habitantes em domicílios improvisados

Descrição dos atributos da Base de Dados QVU

	Atributo	Descrição
14	porpdp	Percentual de habitantes em domicílios permanentes
15	porpopal	Percentual de indivíduos alfabetizados
16	rendchef (SM)	Renda média dos chefes de família (Salários Mínimos)
17	rotas	Número de rotas de ônibus
18	sinstsan	Número de domicílios sem instalações sanitárias
19	stop_pto	Número de pontos de parada das rotas de ônibus

Pré-Processamento

- Etapa do processo de *Data Mining* que consiste em preparar os dados para serem submetidos a algum algoritmo.
- Realiza operações como: seleção dos dados, transformações, adequação de formato, redução da dimensão, dentre outras.

Pré-Processamento

Primeira Atividade Realizada

Com a ajuda de algumas informações estatísticas e de um especialista, no domínio em questão, foram removidos alguns atributos, que apresentam valores irrelevantes, sendo em seguida realizado uma discretização dos dados.

Atributos Removidos

Por sugestão do especialista, no domínio em questão, foram removidos 3 atributos da Base de Dados QVU :

dpi_p - Percentual de domicílios particulares improvisados;

porpdimp - Percentual de habitantes em domicílios improvisados;

porpdp - Percentual de habitantes em domicílios permanentes.

Intervalos de discretização

	Atributo	Intervalos
1	IQVU	(... -0.544035], (-0.544035 ... -0.02481], (-0.02481 ... 0.494415], (0.494415 ...]
2	acessibi	(... 0.6615], (0.6615 ... 0.738], (0.738 ... 0.8145], (0.8145 ...]
3	agua_p	(... 95], (95 ... 98], (98 ...]
4	cmd_dom	(... 4], (4 ... 5], (5 ... 6], (6 ... 7], (7 ...]
5	denscons	(... 117308], (117308 ... 213916], (213916 ... 310523], (310523 ...]
6	densidad	(... 25], (25 ... 50], (50 ... 80], (80 ...]
7	esgoto_p	(... 95], (95 ... 98], (98 ...]
8	kilomete	(... 1], (1 ... 3], (3 ...]
9	lixo_p	(... 95], (95 ... 98], (98 ...]
10	perocppr	(... 60.6167], (60.6167 ... 73.5633], (73.5633 ...]
11	pess_dom	(... 2], (2 ... 3], (3 ...]

Intervalos de discretização

	Atributo	Intervalos
12	porpopal	(... 78.7067], (78.7067 ... 84.4033], (84.4033 ...]
13	rendchef (SM)	(... 1.60408], (1.60408 ... 2.40398], (2.40398 ... 3.20388], (3.20388 ... 4.00378], (4.00378 ...]
14	rotas	(... 17], (17 ... 24], (24 ... 31], (31 ... 59], (59 ...]
15	sinstsan	(... 1], (1 ... 2], (2 ... 3], (3 ... 4], (4 ...]
16	stop_pto	(... 60], (60 ... 150], (150 ...]

Legenda:

(e) ou] e [- intervalo aberto

[e] - intervalo fechado

Pré-Processamento

Segunda Atividade Realizada

Foram realizadas algumas operações para adequar o conjunto de dados ao formato de entrada do algoritmo *Apriori* (utilizado na realização dos experimentos).

O algoritmo *Apriori*

- A implementação do algoritmo Apriori (disponível em http://fuzzy.cs.uni-magdeburg.de/~borgelt/software.html) recebe como entrada, um arquivo de dados contendo transações.
- Cada linha do arquivo representa uma transação e os itens (valores discretos) dentro de uma transação são separados por um espaço em branco.

Extração de Padrões

- Essa etapa consiste na aplicação de algum algoritmo ao conjunto de exemplos para a extração de padrões.
- Os padrões extraídos a partir de um determinado conjunto de exemplos armazenam o conhecimento adquirido.

O algoritmo *Apriori*

- O algoritmo *Apriori* é executado com o comando:

  apriori.exe –o arquivo_entrada.tab arquivo_saida.rul

- O resultado do algoritmo são regras no formato:

  RHS <- LHS (suporte%,confiança%)

O algoritmo *Apriori*

- Com suporte = 1%, confiança = 50% e considerando no máximo 3 itens por regra, o algoritmo *apriori* aplicado à Base de Dados QVU gerou 14299 regras.
- Já com suporte = 1%, confiança = 50% e considerando no máximo 5 itens por regra, o algoritmo *apriori* aplicado à Base de Dados QVU gerou 502048 regras.

Exemplos de Regras

stop_pto=(60...150] <- SINSTSAN=(2...3] (1.7%, 100.0%)

CMD_DOM=(7...] <- RENDCHEF(SM)=(4.00378...] (4.2%, 62.5%,)

RENDCHEF(SM)=(4.00378...] <- CMD_DOM=(7...] (4.2%, 50.0%)

PEROCPPR=(60.6167...73.5633] <- rotas=(17...24] agua_p=(98...] (9.2%, 68.8%)

lixo_p=(98...] <- rotas=(17...24] PESS_DOM=(2...3] (8.3%, 83.3%)

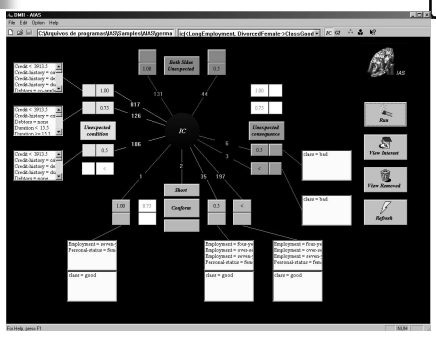
PESS_DOM=(2...3] <- rotas=(17...24] lixo_p=(98...] (8.3%, 58.8%)

QV=(0.494415...] <- SINSTSAN=(3...4] (1.7%, 100.0%)

Pós-Processamento

- A técnica de Regras de Associação possui o inconveniente de gerar um grande volume de regras.
- Nessa etapa, os padrões extraídos devem ser filtrados, com o uso de medidas, para que se possa realizar uma análise buscando identificar se os padrões obtidos geram um conhecimento novo, interessante, válido e útil para o usuário.
- Avaliação das regras com o aplicativo AIAS – *Association Interestingness Analysis System*

Aplicativo AIAS



Aplicativo AIAS

- O procedimento para análise das regras inicia-se com a definição de um arquivo de análise de interessabilidade, contendo:
  - arquivo de regras (formato .nar ou .rul);
  - especificação da taxonomia do conhecimento (opcional);
  - definição do conhecimento impreciso;
  - definição de impressões gerais;
- Na elaboração dos três últimos itens é necessário o conhecimento do especialista.

Arquivo de Regras

- Formato .nar

$ant_1, ant_2, ..., ant_n$ -> $cons_1$, $cons_2$, ..., $cons_n$ (suporte,confiança)

- Formato .rul

Rule X:
ant_1
ant_2
...
ant_n
-> $cons_1$
$cosn_2$
...
$cons_n$
(suporte, confiança, NLHS, NRHS)

NLHS é número de itens com LHS.
NRHS é o número de itens com ambos os lados da regra.

Conhecimento Impreciso

Conhecimento Impreciso (ic): possibilita informar um conhecimento que o especialista supõe ser verdadeiro. É expresso na forma:

ic(<ant_A, ant_B **->** Cons>) (Suporte,Confiança)

Impressões Gerais

Impressões Gerais (gi): possibilita informar uma relação que o especialista acredita existir entre os atributos especificados. É expressa na forma:

gi(<Atr_1, Atr_2, Atr_N>) (Suporte,Confiança)

Tipos de Análise

- Identificação de conformidade: Identifica e classifica as regras que estão em conformidade com uma impressão geral ou um conhecimento impreciso fornecido pelo usuário do domínio.
- Conseqüente inesperado: Permite avaliar se o conseqüente da regra é inesperado.
- Antecedente inesperado: Identifica outros antecedentes que estão associados aos conseqüentes especificados na impressão geral ou no conhecimento impreciso.
- Antecedente e conseqüente inesperados: Permite avaliar se o antecedente e o conseqüente da regra são inesperados.

Análise das Regras

- Utilizou-se 14299 regras geradas com o algoritmo *apriori* (suporte = 1%, confiança = 50% e máximo de 3 itens por regra).
- No aplicativo AIAS, os símbolos de "( )" são utilizados somente como delimitadores das medidas suporte e confiança.
- Os símbolos de aberto "( )" das regras foram substituidos por "] [". O atributo RENDCHEF(SM) foi substituido por RENDCHEF_SM.

Análise das Regras

ic(< SINSTSAN-]3...4] -> RENDCHEF_SM-]...1.60408] >) (0.017, 1)

- **Identificação de conformidade:** 1 regra com 100% de conformidade e 10 regras com conformidade entre [50% ... 75%).
- **Conseqüente inesperado:** 10 regras com 100% e 100 regras entre [50% ... 75%) do conseqüente inesperado.
- **Antecedente inesperado:** 76 regras com 100% e 10 regras entre [50% ... 75%) do antecedente inesperado.
- **Antecedente e conseqüente inesperados:** 2411 regras com 100% e 110 regra entre [50% ... 75%) do antecedente e conseqüente inesperados.

Análise das Regras

gi(< lixo_p−]98...], agua_p−]98...], esgoto_p−]...95] >) (0.1,1)

- **Identificação de conformidade:** 23 regras com conformidade entre [50% ... 75%) e 62 regras com conformidade entre [0% ... 50%).
- **Conseqüente inesperado:** 9 regras com conseqüente inesperado entre [50% ... 75%).
- **Antecedente inesperado:** 53 regras com 100% e 23 regras entre [50% ... 75%) do antecedente inesperado.
- **Antecedente e conseqüente inesperados:** 99 regras com 100% e 9 regra com antecedente e conseqüente inesperados entre [50% ... 75%).

Bibliografia

- Adamo J.M. (2001). Data Mining for Association Rules and Sequential Patterns. Springer-Verlag.
- Agrawal R. & R. Srikant (1994). Fast Algorithms for Mining Association Rules in Large Databases. In Proceedings of the 20th International Conference on Very Large Databases.
- Webb G. I. (1995). Opus: An Efficient Admissible Algorithm for Unordered Search. Journal of Artificial Intelligence Research 3, 431–465.
- Zaki M. & C. Hsiao (2002). Charm: An Efficient Algorithm for Closed Association Rule Mining. In R. Grossman, J. Han, V. Kumar, H. Mannila, and R. Motwani (Eds.), 2nd SIAM International Conference on Data Mining.
- Pei J., J. Han & R. Mao (2000). Closet: An Efficient Algorithm for Mining Frequent Closed Itemsets. In ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, pp. 21–30.
- J. Han, J. Pei & Y. Yin (1999). Mining Frequent Patterns Without Candidate. Technical Report TR-99-12, Computing Science Technical Report, Simon Fraser University, October 1999.