Mineração de padrões frequentes

Fabrício Jailson Barth

Dezembro de 2012

Sumário e Objetivos

- Mineração de itens frequentes.
- Definição de suporte e confiança.
- Caso: Identificação de áreas correlatas.

Mineração de itens frequentes

- Dado:
  - ⋆ um conjunto $A = \{a_1, \cdots, a_m\}$ de itens,
  - ⋆ uma tabela $T = (t_1, \cdots, t_n)$ de transações sobre A,
  - ⋆ um número β_{min} que $0 < \beta_{min} \leq 1$, o **suporte mínimo**.
- Objetivo 1:
  - ⋆ encontrar o conjunto de **itens frequentes**, tais que o **suporte** de cada conjunto de itens é maior ou igual ao β_{min} definido pelo usuário.

Exemplo de transações

	Itens
1	{a,d,e}
2	{b,c,d}
3	{a,c,e}
4	{a,c,d,e}
5	{a,e}
6	{a,c,d}
7	{b,c}
8	{a,c,d,e}
9	{b,c,e}
10	{a,d,e}

0 itens	1 item	2 itens	3 itens
{}: 10	{a}: 7	{a,c}: 4	{a,c,d}: 3
	{b}: 3	{a,d}: 5	{a,c,e}: 3
	{c}: 7	{a,e}: 6	{a,d,e}: 4
	{d}: 6	{b,c}: 3	
	{e}: 7	{c,d}: 4	
		{c,e}: 4	
		{d,e}: 4	

Figure 1: Um banco de dados de transações, com 10 transações, e a enumeração de todos os conjuntos de itens frequentes usando o suporte mínimo = 0,3

Mineração de itens frequentes

- Objetivo 2:
  - ⋆ encontrar o conjunto de regras de associação com confiança maior que um mínimo definido pelo utilizador.

Suporte e Confiança

O suporte de um conjunto de itens Z, $suporte(Z)$, representa a porcentagem de transações na base de dados que contêm os itens de Z.

O suporte de uma regra de associação $A \rightarrow B$, $suporte(A \rightarrow B)$, é dado por $suporte(A \cup B)$.

$$confianca(A \rightarrow B) = \frac{P(A \cup B)}{P(A)} = \frac{suporte(A \cup B)}{suporte(A)} \quad (1)$$

Exemplo de regras geradas

Premises	Conclusion	Support	Confidence
b	c	0.300	1
e, d	a	0.400	1
e	a	0.600	0.857
a	e	0.600	0.857
d	a	0.500	0.833
a, d	e	0.400	0.800

Figure 2: Regras extraídas com confiança maior que 0,8 (processo executado usando o RapidMiner.)

Estudo de caso

- Candidatos podem definir várias áreas no objetivo profissional.
- Empresas podem definir várias áreas em cada vaga.
- **É possível identificar áreas correlatas através de regras de associação?**

Características do experimento

- Dados utilizados: vagas criadas no 2^o semestre de 2012.
- 51.311 vagas (*transações*)
- 124 áreas (*atributos*)

Fase de pré-processamento

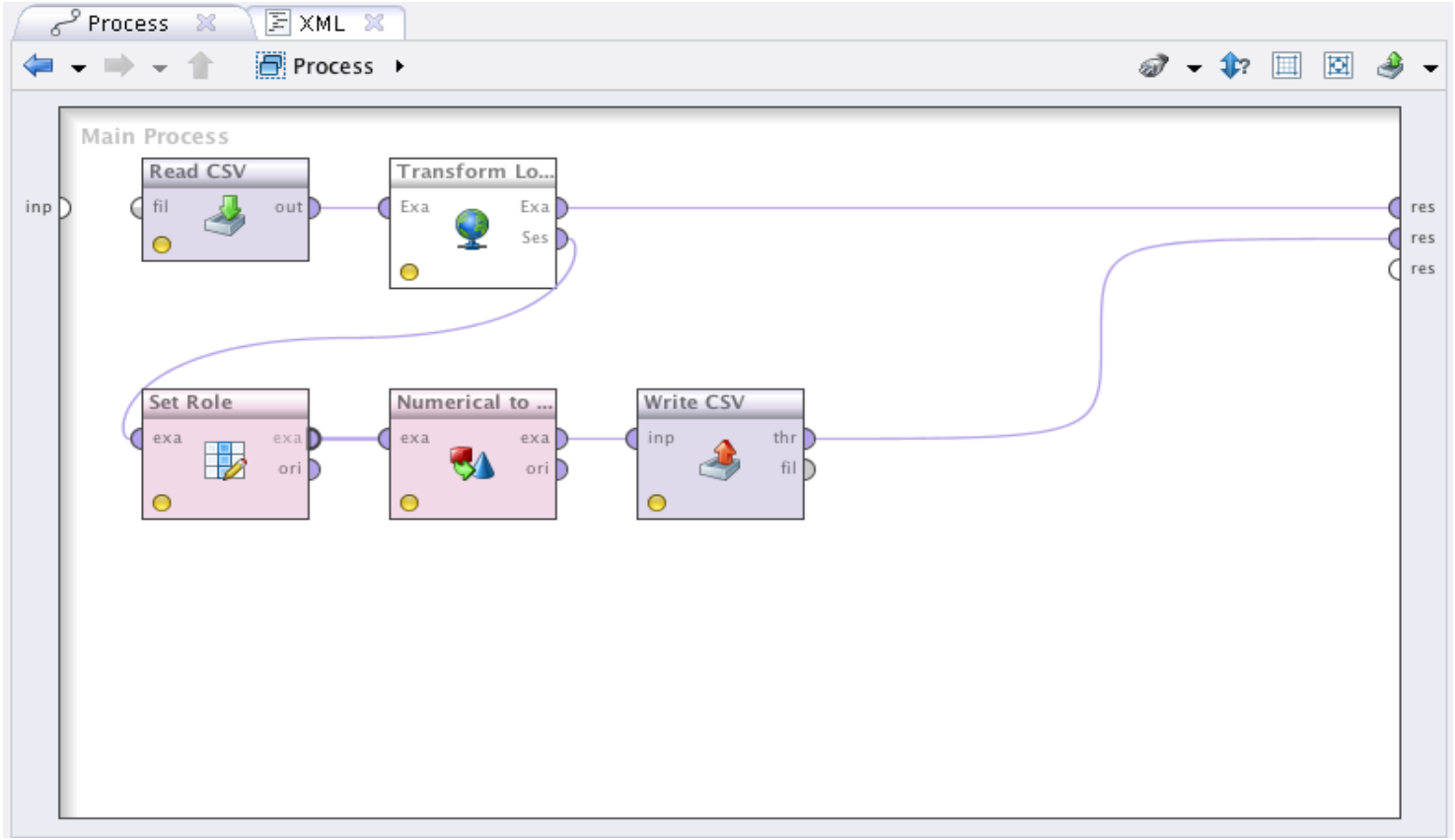

Figure 3: Processo especificado na ferramenta RapidMiner

Fase de modelagem

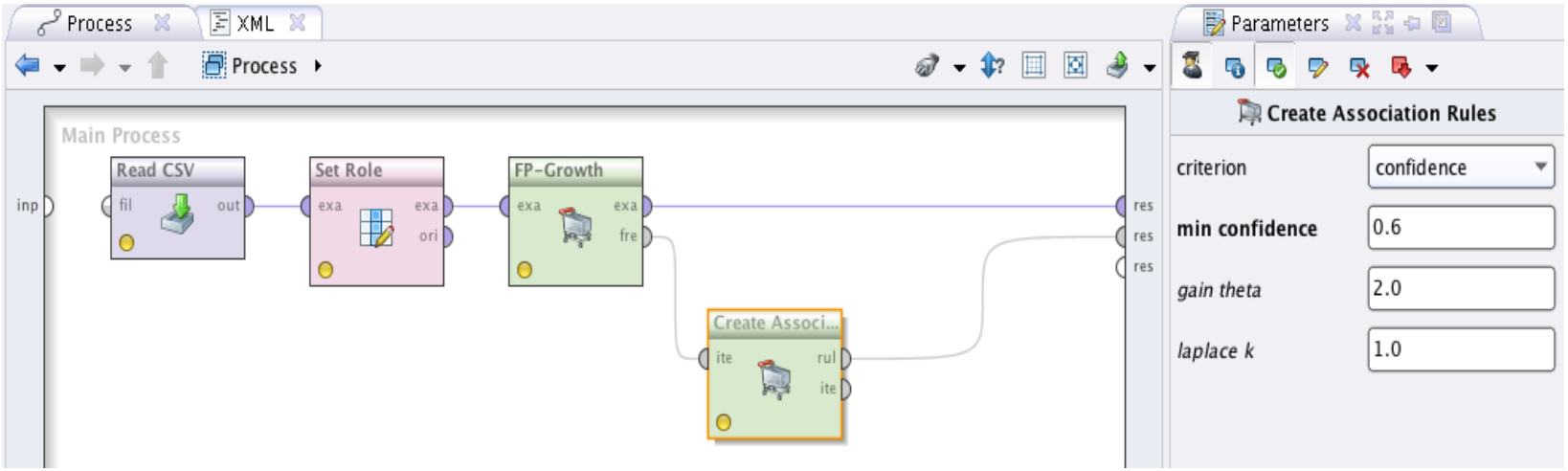

Figure 4: Processo especificado na ferramenta RapidMiner

Regras geradas

Premises	Conclusion	Support	Confidence	LaPlace	Gain	p-s	Lift	Conviction
Economia	Administração de Empresas	0.088	0.920	0.993	-0.104	0.056	2.705	8.245
Finanças	Administração de Empresas	0.066	0.812	0.986	-0.097	0.039	2.388	3.514
Contabilidade	Administração de Empresas	0.074	0.719	0.974	-0.132	0.039	2.113	2.344
Vendas	Administração Comercial/Vendas	0.073	0.644	0.964	-0.153	0.054	3.794	2.331

Figure 5: Regras geradas para Suporte = 0,05 e Confiança = 0,6

Representação visual das regras geradas

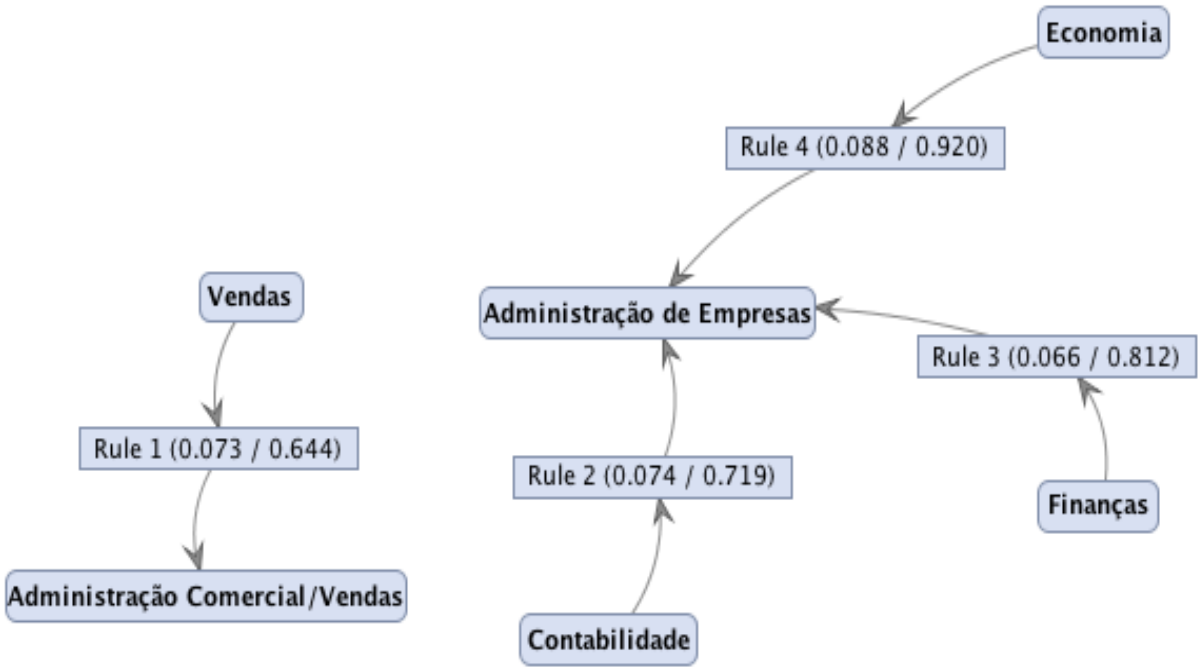

Figure 6: Regras geradas para Suporte = 0,05 e Confiança = 0,6

Heurísticas para Seleção de Regras de Associação

- *Lift*: um valor de *lift* para uma regra $(A \rightarrow B)$ superior a 1 indica que A e B aparecem mais frequentemente juntos do que o esperado, isso significa que a ocorrência de A tem um efeito positivo sobre a ocorrência de B.
- *Convicção*: este valor indica que a probabilidade do item A ocorrer sem o item B é X vezes menor.

Material de **consulta**

- Fabrício Barth. Mineração de regras de associação em servidores Web com RapidMiner[a].
- Iah H. Witteh and Eibe Frank. Data Mining: Practical Machine Learning Tools and Techniques (Third Edition), 2011.
- Gonçalves. Regras de Associação e suas Medidas de Interesse Objetivas e Subjetivas. INFOCOMP Journal of Computer Science, 2005, 4, 26-35.
- *Faceli, Lorena, Gama, Carvalho. Inteligência Artificial: uma abordagem de aprendizado de máquina, 2011.*

[a] http://fbarth.net.br/materiais/webMining/webUsageMining.pdf

Arquivos RapidMiner utilizados

- 20121228_exemplo_regras_associacao.rmp
- 20121228_geracao_regras_por_vaga_fase_0.rmp
- 20121228_geracao_regras_por_vaga.rmp