Representação de Conhecimento Usando Teoria de Conjuntos Aproximados

MARCOS AURÉLIO DOMINGUES
JOAQUIM QUINTERO UCHÔA

UFLA - Universidade Federal de Lavras
DCC - Departamento de Ciência da Computação
Cx Postal 37 - CEP 37200-000 Lavras (MG)
domingues@comp.ufla.br
joukim@ufla.br

Resumo. A Teoria de Conjuntos Aproximados (TCA), tem sido utilizada em várias áreas de pesquisa, principalmente naquelas relacionadas com representação de conhecimento e aprendizado de máquina. Este trabalho tem por objetivo apresentar os principais conceitos da TCA e demonstrar sua viabilidade para utilização como um formalismo matemático para a representação de conhecimento na presença de incerteza.

Palavras-Chave: Teoria de Conjuntos Aproximados (TCA), espaços e conjuntos aproximados, sistemas de representação de conhecimento (SRCs).

1 Introdução

A Teoria de Conjuntos Aproximados (TCA) foi proposta em [Pawlak (1982)], como um novo modelo matemático para representação do conhecimento, tratamento de incerteza e classificação aproximada. Devido a estas características, tem-se utilizado esta teoria em Inteligência Artificial, especialmente nas áreas de:

- aprendizado de máquina;
- aquisição de conhecimento;
- raciocíonio indutivo;
- descoberta de conhecimento em base de dados.

Como pode ser verificado em [Pawlak (1991)], conjuntos aproximados podem ser considerados conjuntos com fronteiras nebulosas, ou seja, conjuntos que não podem ser caracterizados precisamente, utilizando-se dos atributos disponíveis.

A TCA se difere de outras teorias (como a Teoria de Conjuntos Fuzzy [Zadeh (1965)] e a Teoria de Evidência de Dempster-Shafer [Shafer (1976)]) por não necessitar de nenhuma informação adicional a respeito de dados, tais como: distribuição de probabilidade, atribuição de crenças, grau de pertinência ou possibilidade. Isto tem levado a TCA a ser uma grande fonte de inspiração para pesquisadores que buscam o desenvolvimento de sistemas lógicos e métodos dedutivos para a representação, manipulação e raciocínio na presença de informações incompletas.

2 Principais Conceitos da TCA

2.1 Espaços Aproximados

Um *espaço aproximado* é um par ordenado $A = (U, R)$, onde:

- U é um conjunto não vazio, denominado *conjunto universo*;
- R é uma relação de equivalência sobre U, denominada *relação de indiscernibilidade*. Dados $x, y \in U$, se xRy então x e y são *indiscerníveis* em A, ou seja, a classe de equivalência definida por x é a mesma que a definida por y, i.e., $[x]_R = [y]_R$.

As classes de equivalência induzidas por R em U são denominadas *conjuntos elementares*. Se E é um conjunto elementar, $des(E)$ denota a descrição dessa classe de equivalência. Essa descrição é função do conjunto de atributos que define R. Note que, dados $x, y \in E$, onde E é um conjunto elementar em A, x e y são indiscerníveis, i.e., no espaço $A = (U, R)$ não se consegue distinguir x de y, pois $des(x) = des(y) = des(E)$.

2.2 Aproximação Inferior e Superior

Dado um espaço aproximado $A = (U, R)$ e um conjunto $X \subseteq U$, com o objetivo de verificar o quão bem X é representado pelos conjuntos elementares de A, são definidas:

- a *aproximação inferior* de X em A, $A_{A-inf}(X)$, como a união de todos os conjuntos elementares que estão contidos em X, em símbolos,

$$A_{A-inf}(X) = \{x \in U \mid [x]_R \subseteq X\}$$

- a *aproximação superior* de X em A, $A_{A-sup}(X)$, como a união de todos os conjuntos que possuem intersecção não vazia com X, em símbolos,

$$A_{A-sup}(X) = \{x \in U \mid [x]_R \cap X \neq \varnothing\}$$

Nas notações utilizadas, quando o espaço aproximado for conhecido e não houver risco de confusão, a referência ao espaço será abolida. Assim, por exemplo, $A_{sup}(X)$ será usado em substituição a $A_{A-sup}(X)$.

2.3 Regiões do Espaço Aproximado

Dado um espaço aproximado $A = (U, R)$ e $X \subseteq U$, as aproximação inferior e aproximação superior permitem a classificação do espaço aproximado em regiões:

- *região positiva* de X em A, formada por todas as classes de equivalência de U contidas inteiramente no conjunto X e dada por $pos_A(X) = A_{A-inf}(X)$. Pode-se afirmar que todos os elementos desta região pertencem a X.

- *região negativa* de X em A, formada pelos conjuntos elementares de A que não estão contidos na aproximação superior de X e dada por $neg_A(X) = U - A_{A-sup}(X)$. Pode-se afirmar que nenhum dos elementos de A pertencem a X.

- *região duvidosa* de X em A, formada pelos elementos que pertencem a aproximação superior mas não pertencem à aproximação inferior e dada por $duv_A(X) = A_{A-sup}(X) - A_{A-inf}(X)$. A pertinência de um elemento desta região a X é incerta.

Exemplo 1 *Seja U um conjunto universo e R uma relação de equivalência em U, definindo o espaço aproximado $A = (U, R)$. Seja também X, como ilustra a Figura 1. A aproximação inferior e a aproximação superior de X em $A = (U, R)$ são mostradas na Figura 2(a) e 2(b). A Figura 3, por sua vez, apresenta as regiões de X.*

Seja $A = (U, R)$ um espaço aproximado e seja $X \subseteq U$. O conjunto X pode ou não ter suas fronteiras claramente definidas em função das descrições dos conjuntos elementares de A. Isso leva ao conceito de conjuntos aproximados: um *conjunto aproximado* em A é a família de todos os subconjuntos de U que possuem a mesma aproximação inferior e a mesma aproximação superior em A. Ou seja, possuem a mesma região positiva, negativa e duvidosa.

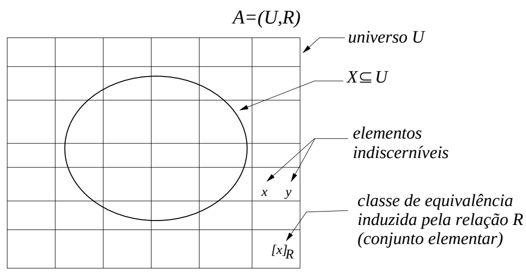
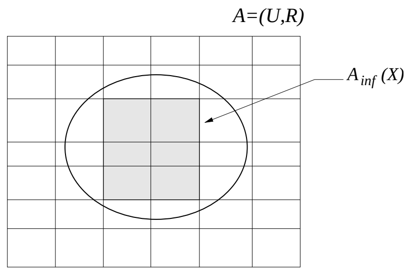
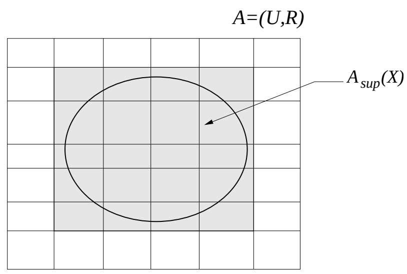


Figura 1: Conjunto X no espaço aproximado $A = (U, R)$



(a) Aproximação Inferior de X



(b) Aproximação Superior de X

Figura 2: Aproximações de X em A

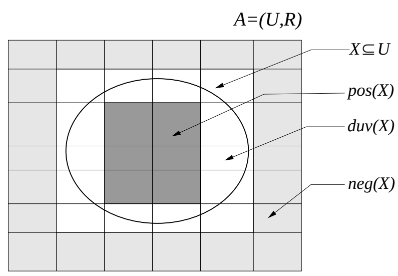


Figura 3: Regiões de X em A

2.4 Outras Medidas da TCA

Seja um espaço aproximado $A = (U, R)$ e um conjunto $X \subseteq U$. Com o objetivo de "medir" quão bem o espaço A pode refletir as funções de pertinência de objetos em X, as seguintes medidas são definidas:

- *medida interna* de X em A, $\omega_{A-inf}(X) = |A_{A-inf}(X)|$
- *medida externa* de X em A, $\omega_{A-sup}(X) = |A_{A-sup}(X)|$
- *qualidade da aproximação inferior* de X em A, $\gamma_{A-inf}(X) = \omega_{A-inf}(X)/|U|$
- *qualidade da aproximação superior* de X em A, $\gamma_{A-sup}(X) = \omega_{A-sup}(X)/|U|$
- *acurácia* de X em A, $\omega_A(X) = \omega_{A-inf}(X)/\omega_{A-sup}(X)$
- *pertinência aproximada* de x a X, $\mu_X(x) = \frac{|[x]_R \bigcap X|}{|[x]_R|}$

Também com estas medidas, quando não houver dúvida, a referência ao espaço aproximado será omitida da notação. Estes conceitos serão exemplificados mais à frente, no Exemplo 4.

3 Sistemas de Representação de Conhecimento (SRCs)

Os conceitos da TCA são utilizados principalmente no contexto de Sistemas de Representação de Conhecimento. Um *Sistema de Representação de Conhecimento* (SRC) é uma quádrupla $S = (U, Q, V, \rho)$, onde U é o universo finito de S.

Os elementos de U são chamados objetos, que são caracterizados por um conjunto de atributos Q e seus respectivos valores. O conjunto de valores de atributos é dado por $V = \bigcup_{q \in Q} V_q$, onde V_q é o conjunto de valores do atributo q. Por sua vez, $\rho\colon U \times Q \to V$ é uma *função de descrição* tal que $\rho(x, q) \in V_q$, para $x \in U$ e $q \in Q$.

Exemplo 2 *O SRC representado pela Tabela 1 descreve sete tipos de dietas, em termos de cinco atributos, que descrevem as quantidades de vitamina A, vitamina C, vitamina D, proteínas e lipídios de cada dieta, e que são apresentados sobre a seguinte notação: vitA, vitC, vitD, prot e lip. Assim,* $U = \{d_1, d_2, d_3, d_4, d_5, d_6, d_7\}$, $Q = \{vitA, vitC, vitD, prot, lip\}$ *e* $V = \{1, 2, 3, 4, 5\}$. *Neste SRC tem-se:*

- $\rho(d_1, vitA) = 1$;
- $\rho(d_2, vitC) = 3$;
- $\rho(d_5, vitD) = 2$;
- $\rho(d_7, lip) = 2$.

Tabela 1: SRC onde $U = \{d_1, d_2, d_3, d_4, d_5, d_6, d_7\}$ e $Q = \{vitA, vitC, vitD, prot, lip\}$

U	$vitA$	$vitC$	$vitD$	$prot$	lip
d_1	1	3	4	2	3
d_2	1	3	3	3	2
d_3	1	3	4	3	1
d_4	3	5	2	5	2
d_5	4	5	2	5	1
d_6	3	5	2	3	4
d_7	4	4	1	3	2

Dado um SRC $S = (U, Q, V, \rho)$, é importante observar que cada subconjunto de atributos $P \subseteq Q$ define um único espaço aproximado $A = (U, \tilde{P})$ onde $\tilde{P}$ é a relação de indiscernibilidade (equivalência) induzida por P.

Exemplo 3 *Seja S o SRC do Exemplo 2 e $P = \{lip\}$. Neste caso tem-se que os elementos d_2, d_4 e d_7 são indiscerníveis com relação a P pois possuem o mesmo valor. Além disto, $A = (U, \tilde{P})$ é um espaço aproximado e seus conjuntos elementares são $E_1 = \{d_1\}$, $E_2 = \{d_2, d_4, d_7\}$, $E_3 = \{d_3, d_5\}$ e $E_4 = \{d_6\}$.*

3.1 Índice Discriminante de Atributos

Dado um espaço aproximado $A = (U, \tilde{P})$, definido por um subconjunto de atributos $P \subseteq Q$ em um sistema de representação de conhecimento $S = (U, Q, V, \rho)$ e $X \subseteq U$, define-se o *índice discriminante* de X em relação ao subconjunto de atributos $P \subseteq Q$, notado por $\alpha_P(X)$, como uma medida do grau de certeza na determinação da pertinência de um elemento de U ao conjunto X, de acordo com os atributos de P, que é dado por:

$$\alpha_P(X) = \frac{|U| - |A_{sup}(X) - A_{inf}(X)|}{|U|}.$$

Exemplo 4 *Seja S o SRC definido no Exemplo 2 e seja $P = \{vitC, vitD\}$. Nesse caso, os conjuntos elementares do espaco aproximado $A = (U, \tilde{P})$ são: $E_1 = \{d_1, d_3\}$, $E_2 = \{d_4, d_5, d_6\}$, $E_3 = \{d_7\}$ e $E_4 = \{d_2\}$. Seja $X = \{d_1, d_2\}$. Tem-se que:*

$$A_{inf}(X) = \{d_2\}$$
$$A_{sup}(X) = \{d_1, d_2, d_3\}$$

$$\alpha_P(X) = \frac{7-|3-1|}{7} = \frac{5}{7} = 0.71$$

$$\omega_{inf}(X) = 1$$
$$\omega_{sup}(X) = 3$$
$$\omega(X) = 0.333$$

$$\gamma_{inf}(X) = 0.142$$
$$\gamma_{sup}(X) = 0.428$$

$$\mu_X(d_1) = 0.5$$

3.2 Tabelas de Decisão

No contexto da TCA o interesse recai, principalmente, sobre tabelas de decisão, um tipo particular de SRC. Uma *tabela de decisão* é um SRC onde os atributos de Q são divididos em condições e decisões. Tem-se então $Q = C \cup D$, onde C é o conjunto das condições e D o conjunto das decisões. Como geralmente o conjunto D é unitário, uma tabela de decisão é descrita por $S = (U, C \cup \{\delta\}, V, \rho\}$, onde U, V e ρ são tais como num SRC, C é o conjunto de condições e δ é o atributo de decisão. Por $Class_S(\delta)$ entende-se a classificação de S, i.e., a família de conjuntos elementares do conjunto aproximado induzido por $\{\delta\}$.

Exemplo 5 *Seja S o SRC do Exemplo 2. Seja lip o atributo de decisão em S, i.e., $\delta = lip$. S é uma tabela de decisão, onde $C = \{vitA, vitC, vitD, prot\}$. Os conjuntos elementares do espaço aproximado induzido por C são $\{d_1\}$, $\{d_2\}$, $\{d_3\}$, $\{d_4\}$, $\{d_5\}$, $\{d_6\}$ e $\{d_7\}$. Por sua vez, a classificação de S é dada por $Class_S(\delta) = \{\{d_1\}, \{d_2, d_4, d_7\}, \{d_3, d_5\}, \{d_6\}\}$.*

Dada uma tabela de decisão $S = (U, C \cup \{\delta\}, V, \rho\}$, é importante verificar o quão bem a família de conjuntos elementares induzidos pelas condições $P \subseteq C$ espelha a família de conjuntos elementares induzidos por $\{\delta\}$. Para isso, considerando o espaço aproximado induzido por P, são definidas:

- *região positiva* de δ induzida por P,

$$pos(P, \delta) = \bigcup_{X \in Class_S(\delta)} A_{P-inf}(X)$$

- *grau de dependência* de δ com relação a P,

$$\kappa(P, \delta) = \frac{|pos(P, \delta)|}{|U|}$$

- *fator de significância* de um atributo $a \in P$, com relação à dependência existente entre δ e P,

$$FS(a, P, \delta) = \frac{(\kappa(P, \delta) - (\kappa(P - \{a\}, \delta))}{\kappa(P, \delta)}$$

se $\kappa(P, \delta) > 0$

Exemplo 6 *Seja S o SRC definido no Exemplo 2 e seja $P = \{vitA, vitC\}$. Os conjuntos elementares do espaço aproximado induzido por P são: $E_1 = \{d_1, d_2, d_3\}$, $E_2 = \{d_4, d_6\}$, $E_3 = \{d_5\}$ e $E_4 = \{d_7\}$. Por sua vez, $Class_S(vitD) = \{\{d_1, d_3\}, \{d_2\}, \{d_4, d_5, d_6\}, \{d_7\}\}$. Com isso, é possível verificar os seguintes resultados:*

$$pos(P, vitD) = \bigcup_{X \in Class_S(vitD)} A_{P-inf}(X) =$$
$$= A_{P-inf}(\{d_1, d_3\}) \cup A_{P-inf}(\{d_2\}) =$$
$$= \cup A_{P-inf}(\{d_4, d_5, d_6\}) \cup A_{P-inf}(\{d_7\}) =$$
$$= \{d_4, d_6\} \cup \{d_5\} \cup \{d_7\} =$$
$$= \{d_4, d_5, d_6, d_7\}$$

$$\kappa(P, vitD) = \frac{|pos(P, vitD)|}{|U|} = \frac{4}{7} = 0.571$$

Para o cálculo de:

$$FS(vitA, P, vitD) =$$
$$= \frac{(\kappa(P, vitD) - (\kappa(P - \{vitA\}, vitD))}{\kappa(P, vitD)},$$

bem como o de

$$FS(vitC, P, vitD) =$$
$$= \frac{(\kappa(P, vitD) - (\kappa(P - \{vitC\}, vitD))}{\kappa(P, vitD)}$$

é necessário calcular antes o valor de $\kappa(vitC, vitD)$ e $\kappa(vitA, vitD)$, respectivamente, o que é feito a seguir.

Considerando-se que $P_1 = \{vitC\}$ e $P_2 = \{vitA\}$, tem-se que as famílias dos conjuntos elementares dos espaços aproximados induzidos por P_1 e P_2 são dadas, respectivamente, por $\{\{d_1, d_2, d_3\}, \{d_4, d_5, d_6\}, \{d_7\}\}$ e $\{\{d_1, d_2, d_3\}, \{d_4, d_6\}, \{d_5, d_7\}\}$. Donde:

$$pos(P_1, vitD) = \{d_4, d_5, d_6, d_7\} \text{ e}$$
$$\kappa(P_1, vitD) = \frac{4}{7} = 0.571$$

$$pos(P_2, vitD) = \{d_4, d_6\} \text{ e}$$
$$\kappa(P_2, vitD) = \frac{2}{7} = 0.285$$

Portanto:

$$FS(vitA, P, vitD) =$$
$$\frac{(\kappa(P, vitD) - (\kappa(\{vitC\}, vitD))}{\kappa(P, vitD)} =$$
$$= \frac{\frac{4}{7} - \frac{4}{7}}{\frac{4}{7}} = 0$$

$$FS(vitC, P, vitD) =$$
$$\frac{(\kappa(P, vitD) - (\kappa(\{vitA\}, vitD))}{\kappa(P, vitD)} =$$
$$= \frac{\frac{4}{7} - \frac{2}{7}}{\frac{4}{7}} = \frac{2}{3} = 0.6$$

Diz-se, ainda com respeito a uma tabela de decisão $S = (U, C \cup \{\delta\}, V, \rho\}$ e $P \subseteq C$ que P é *independente* com relação à dependência existente entre δ e P se, para todo subconjunto próprio $R \subset P$, for verdade que $pos(P, \delta) \neq pos(R, \delta)$, i.e., $\kappa(P, \delta) \neq \kappa(R, \delta)$. Caso haja algum $R \subset P$ tal que $pos(P, \delta) = pos(R, \delta)$, i.e., $\kappa(P, \delta) = \kappa(R, \delta)$, então P é dito ser *dependente* com relação à dependência existente entre δ e P. Um conjunto $R \subset P$ é dito ser um *reduto* de P com relação à dependência existente entre δ e P se for independente com relação à dependência existente entre δ e R, e $pos(P, \delta) = pos(R, \delta)$, i.e., $\kappa(P, \delta) = \kappa(R, \delta)$.

Exemplo 7 *Seja S o SRC do Exemplo 2, onde lip é o atributo de decisão em S. Como já visto, S e uma tabela de decisão, onde a classe de condições é dada por $C = \{vitA, vitC, vitD, prot\}$ e a decisão é dada por $\{\delta\} = \{lip\}$. A família $Class_S(lip)$ foi determinada no Exemplo 5. Tem-se os seguintes resultados em S:*

$$\kappa(C, lip) = \frac{|pos(C, lip)|}{|U|} = \frac{|7|}{|7|} = 1$$

$$\kappa(\{vitA, vitC, vitD\}, lip) = \frac{|2|}{|7|} = 0.285$$

$$\kappa(\{vitA, vitD, prot\}, lip) = \frac{|7|}{|7|} = 1$$

$$\kappa(\{vitC, vitD, prot\}, lip) = \frac{|5|}{|7|} = 0.714$$

$$\kappa(\{vitA, vitC, prot\}, lip) = \frac{|5|}{|7|} = 0.714$$

Neste caso C é dependente e $\{vitA, vitD, prot\}$ é o único subconjunto de C a possuir um reduto de C, pois $\{vitA, vitD, prot\}$ é o único subconjunto com o mesmo grau de dependência de C. Tem-se ainda:

$$\kappa(\{vitA, prot\}, lip) = \frac{|5|}{|7|} = 0.714$$

$$\kappa(\{vitA, vitD\}, lip) = \frac{|3|}{|7|} = 0.428$$

$$\kappa(\{vitD, prot\}, lip) = \frac{|5|}{|7|} = 0.714$$

Portanto, o conjunto $R = \{vitA, vitD, prot\}$ é o único reduto do conjunto de condições com relação a dependência entre δ e C.

4 Considerações Finais

Este trabalho apresentou os principais conceitos da Teoria de Conjuntos Aproximados (TCA), evidenciando o seu formalismo matemático para a representação de conhecimento. E também exemplificou a aplicação de tais conceitos em Sistemas de Representação de Conhecimentos (SRCs).

Como perspectivas futuras, pretende-se implementar uma versão multiplataforma de um sistema que implemente os principais conceitos da TCA, bem como algoritmos de aprendizado baseados nestes conceitos e já apresentados em [Uchôa (1998)] e [Uchôa & Nicoletti (1999)].

Referências

[Pawlak (1982)] Pawlak, Z. Rough sets. *International Journal of Computer and Information Sciences*, 11(5):341-356, 1982.

[Pawlak (1991)] Pawlak, Z. *Rough sets: theoretical aspects of reasoning about data*. London, Kluwer, 1991.

[Shafer (1976)] Shafer, G. *A mathemathical theory of evidence*. Princeton, Princeton University Press, 1976.

[Uchôa (1998)] Uchôa, J. Q. *Representação e indução de conhecimento usando teoria de conjuntos aproximados*. São Carlos, UFSCar, 1998. 237p. (Dissertação de Mestrado).

[Uchôa & Nicoletti (1999)] Uchôa, J. Q. & Nicoletti, M. C. ILROS: um sistema de aprendizado de máquina para domínios incompletos. In: *Anais do 4. SBAI - Simpósio Brasileiro de Automação Inteligente*. Escola Politécnica da USP, São Paulo (SP), 8 a 19 de Setembro de 1999. p. 314-319.

[Zadeh (1965)] Zadeh, L. A. Fuzzy sets. *Information and Control*, (8):338-353, 1965.