ALGORÍTMO A* NA APRENDIZAGEM DE CLASSIFICADORES PROBABILÍSTICOS EM DADOS INCOMPLETOS

Gustavo Burgath (PIBIC/CNPq/FA/UEPG), José Carlos Ferreira da Rocha (Orientador), e-mail: jrocha@uepg.br;

Universidade Estadual de Ponta Grossa/Departamento de Informática

Ciência da Computação, Metodologia e Técnicas da Computação

Palavras-Chave: *Aprendizagem de classificadores, Algoritmo A*, variáveis latentes, redes bayesianas hierárquicas.*

Resumo

Este trabalho apresenta o procedimento A^*_H que estende o procedimento de aprendizagem de redes bayesinas baseado no algoritmo A* descrito por Yuan, Malone e Wu (2011). O A^*_H explora restrições topológicas das redes bayesianas hierárquicas para limitar tamanho do espaço de busca na aprendizagem de classificadores com variáveis latentes. Além disso, emprega o algoritmo EM para tratar os dados faltantes. Os testes mostraram que o custo computacional do A^*_H foi menor que aquele do algoritmo original.

Introdução

A aprendizagem de redes bayesianas é um problema NP-Completo que pode ser abstraído em: aprendizagem da topologia da rede e o treinamento dos parâmetros. Yuan, Malone e Wu (2011) empregam o algoritmo A* e a função MDL nesta tarefa. Aqueles autores aplicam um procedimento combinatório que não explora restrições topológicas quando da geração do espaço de busca. Considerando isto, este trabalho avaliou o procedimento A^*_H, que explora restrições topológicas das redes bayesianas hierárquicas (RBH) na aprendizagem de classificadores. Além disso, o A^*_H emprega o algoritmo EM para tratar as variáveis latentes das RBH.

Materiais e Métodos

O procedimento A^*_H emprega uma técnica de poda na aprendizagem baseada em escores (Correia, Cussens e Campos, 2020) a fim de limitar o número de conjuntos de nós pais de cada variável nas hipóteses de solução geradas pelo A*. O procedimento de poda explora restrições topológicas impostas por RBHs restritas e organizadas em camadas que especificam classificadores bayesianas com variáveis latentes (Njah, Jamoussi e Mahdi, 2018). As restrições foram aplicadas na seleção de variáveis que compõe uma solução e no cômputo das funções de custo e heurística. A função heurística analisa hipóteses que relaxam o conjunto de soluções válidas e, portanto, é admissível. Usou busca em grafo.

Resultados e Discussão

As Tabelas 1 e 2 listam o número de nós expandidos por A^*_H e uma implementação que não usava as restrições topológicas na seleção de variáveis para compor uma solução, A^*_0. Os testes foram realizados em bases da UCI e da PUCHUY. Também é indicado o número de nós das camadas latentes, o número máximo de pais e o número de ciclos do EM.

Tabela 1. Desempenho dos procedimento/ bases da UCI

Base	Atributos (# EM)	#Exp;Ger A^*_H	#Exp;Ger A^*_0	#Var. Lat. #pais
63	19 (8)	420;640	4553;17439	[2,3,4],3
69	60(10)	571;613	722;971	[2,3],3
73	22 (10)	412;583	829;1754	[2,3],3
90	35 (8)	537;741	3938;14539	[2,3],3

Tabela 2. Desempenho dos procedimento/ bases PUCHUY

Base (# EM)	# Atr.	#Exp;Ger A^*_H	#Exp;Ger A^*_0	#Var. Lat. #pais
Bordetella	33	368;408	491;739	[2,3],3
Escherichia	33	397;547	867;1717	[2,3],3
Klebsiella	31	344;398	548;835	[2,3],3
Salmonella	29	324;366	525;810	[2,3],3

Conclusão

Os procedimentos A^*_H e A^*_0 permitem o tratamento de dados faltantes em classificadores com variáveis latentes. Os resultados mostram que o procedimento A^*_H gerou/expandiu um número de estados inferior ao A^*_0. O algoritmo EM insere aleatoriedade durante a aprendizagem. Apesar disso, a maior diferença observada com relação aos valores ótimos do MDL em todas nas bases UCI(PUCHUY) foi inferior a 0.9% (5%).

Agradecimentos

CNPq, Fundação Araucária, UEPG

Correia, A.H.C.; On Pruning for Score-Based Bayesian Network Structure Learning. **Anais do 33º ICAI, PMLR 108**, 2020, 2709-2718.

Njah, H.; Jamoussi, S.; Mahdi, W. Deep Bayesian network architecture for Big Data mining. **Concurr. and Comp.** v.31, n.1, cpe4418 p. 2018.

Yuan, C.; Malone, B.; Wu, X. . Learning Optimal Bayesian Networks Using A* Search. **Anais da 22ª IJCAI**, 2011, p.2186-2191.