Aquisição de Conhecimento Automatizada para Sistemas Especialistas Probabilísticos

Priscyla Waleska Targino de Azevedo Simões[1,2], Daiane de Nez Manarin[1], Merisandra Côrtes de Mattos[1], Jane Bettiol[2]

[1]Curso de Ciência da Computação - Universidade do Extremo Sul Catarinense (UNESC)
Av. Universitária, 1105 - Bairro Universitário – Caixa Postal 3.167 - 88806-000
Criciúma – SC - Brasil

[2]Curso de Medicina - Universidade do Extremo Sul Catarinense (UNESC)
Av. Universitária, 1105 - Bairro Universitário – Caixa Postal 3.167 - 88806-000
Criciúma – SC - Brasil

{pri,mem,ddn}@unesc.net, janebettiol@matrix.com.br

Abstract: *This paper consists of the accomplishment of the process of Knowledge Discoverer in Expert Probabilistic Systems, by knowledge discovery (KDD) in Databases, for the generation of Bayesian Networks. The work presents the analysis of the versions freeware of the following shell of Data Mining in Bayesian Networks: Belief Network Power Constructor (BNPC), Bayesian Knowledge Discoverer (BKD) and Hugin Expert. In the end of the comparative study of the shell, it was opted to the use in this study, of the BNPC, for offering a good interface, and presenting some options of types of archive for importation of databases, advantages among others observed. For the accomplishment of tests in the BNPC, it was used a database on Diabetes Mellitus type 2, and generated the Bayesian Network was evaluated by the expert of the application domain, and considered adjusted.*

Resumo: *O presente trabalho consiste da realização do processo de Aquisição do Conhecimento em Sistemas Especialistas Probabilísticos, por meio da descoberta do conhecimento (KDD) em Bases de Dados, para a geração de Redes Bayesianas. O trabalho apresenta a análise das versões freeware das seguintes ferramentas de Mineração de Dados em Redes Bayesianas: Belief Network Power Constructor (BNPC), Bayesian Knowledge Discoverer (BKD) e Hugin Expert. Ao final do estudo comparativo das ferramentas, optou-se pela utilização nesse estudo, do BNPC, por oferecer uma interface intuitiva, e por apresentar várias opções de tipos de arquivo para importação de bases de dados, entre outras vantagens observadas. Para a realização de testes no BNPC, utilizou-se uma base de dados sobre Diabetes Mellitus tipo 2, e a Rede Bayesiana gerada foi avaliada pela especialista do domínio de aplicação, e considerada adequada.*

1. Introdução

Em aplicações baseadas em conhecimento, os sistemas especialistas probabilísticos se propõem a resolver problemas geralmente de natureza incerta. Nesses sistemas, umas das etapas mais trabalhosas é a de aquisição do conhecimento, que envolve a interação entre o engenheiro do conhecimento e o especialista do domínio de aplicação [Nassar 2003] e [Barreto 2001].

Nos sistemas especialistas probabilísticos voltados a áreas específicas, como por exemplo, a médica, considera-se essa etapa trabalhosa no sentido que ambos os profissionais envolvidos, tanto da área computacional, quanto médica, dispedem muito tempo em reuniões, onde o engenheiro do conhecimento busca compreender o raciocínio do especialista, o qual deve ser explicitado na forma de probabilidades, que costumam ser numerosas, sendo feitas todas as combinações possíveis dos diagnósticos apresentados, para os sinais e sintomas que estão presentes na base de conhecimento [Stein 2000]. Uma das técnicas utilizadas para facilitar este processo é a de Descoberta do Conhecimento em Base de Dados (KDD) para construção de redes *bayesianas (*RB*)* [Vellasco e Lopes 1999], podendo-se assim diminuir o tempo despendido nesta interação, no que se refere ao desenvolvimento de redes *bayesianas,* as quais se propõem a servir como base de conhecimento em sistemas especialistas probabilísticos.

Nesse sentido, essa pesquisa consistiu na utilização de um algoritmo (por meio de uma ferramenta) para a realização da extração de conhecimento de uma base de dados médica, para a geração automática de uma rede *bayesiana*, sendo esse processo intermediado por um especialista, podendo-se dessa forma, contribuir no processo de extração de conhecimento de base de dados.

2. Metodologia

Para a geração da RB, utilizou-se a base de dados sobre prevalência de diabetes mellitus tipo 2 nos bairros São Sebastião, Mineira Velha, Renascer, Laranjinha, Rio Maina, São Luiz e Boa Vista do município de Criciúma no estado de Santa Catarina.

A partir dessa base de dados, iniciou-se o processo de descoberta do conhecimento na ferramenta BNPC, utilizando-se como metodologia as etapas que compõem o processo de KDD aplicado em RBs [Vellasco e Lopes 1999].

Essa ferramenta não permite a visualização da parte quantitativa gerada na RB, bem como a realização de inferências, e por esse motivo, especificou-se a utilização da ferramenta Netica para avaliação dos resultados obtidos com a RB.

Considerando-se o KDD, as etapas de pré-processamento e data mining foram realizadas por meio do aprendizado supervisionado.

O pré-processamento engloba a limpeza, integração, seleção e transformação dos dados. No que se refere a limpeza dos dados, com relação aos valores ausentes, a base de dados apresentou muitos campos não preenchidos que foram tratados, adicionando-se os valores não informados nesses campos, de forma a refletir um estado a ser considerado na geração da RB.

Referindo-se as alterações dos nomes das variáveis, como o Netica não permite títulos de nós com nomes compostos, acentuados ou com caracteres especiais, e a base de dados os apresentou, foi necessário retirar os caracteres especiais, os espaços de nomes compostos e acentuações. Na fase de seleção, definiu-se no BNPC os seguintes campos (relacionados aos fatores de risco e complicações) para o aprendizado da RB: antecedentes familiares, hipertensão arterial, tabagismo, sedentarismo, obesidade, infarto agudo miocárdio, coronariopatias, acidente vascular cerebral, pé diabético, amputação e doença renal.

Na etapa de Data Mining (DM), que refere-se a aplicação de algoritmos específicos para a extração de conhecimento, foram informadas ao BNPC as configurações sobre o domínio de conhecimento, sendo estipuladas nesse item, as configurações de causa e efeito, nós raízes e nós folhas (Figura 1).

Nós Raízes	Nós Folhas
Antecedentes Familiares	Infarto Agudo Miocárdio
Hipertensão Arterial	Coronariopatia
Tabagismo	Acidente Vascular Cerebral
Sedentarismo	Pé Diabético
Obesidade	Amputação
	Doença Renal

Figura 1. Especificações dos nós raízes e nós folhas

Para a conclusão do Data Mining, e ao finalizar a utilização do BNPC, gerou-se a parte quantitativa e qualitativa da RB, baseando-se nas configurações feitas sobre o domínio de conhecimento e no algoritmo árvores de Chow Liu [Acid e Campos 1996].

3. Resultados

Após a geração da RB no BNPC, a mesma foi exportada para o formato da ferramenta Netica para realização do pós-processamento, e análise dos resultados obtidos.

Como no restante do trabalho, e pelo fato do aprendizado ser supervisionado, a análise dos resultados também teve o acompanhamento da especialista do domínio de

aplicação, que deu maior destaque à parte qualitativa, onde observou-se que os resultados obtidos com a RB gerada pela aquisição do conhecimento estão corretos, que os nós e seus *links* refletem corretamente a relação a ser observada entre fatores de risco e complicações, no que se refere a base de dados sobre a prevalência de diabetes mellitus tipo 2 em alguns bairros de Criciúma.

Na parte quantitativa, a especialista não achou necessário ajustes nas Probabilidades Condicionais, as quais apresentaram resultados adequados, refletindo assim o comportamento da população estudada, permitindo-se, dessa forma, relacionar fatores de risco e complicações.

4. Discussão e Conclusões

Em consultas realizadas a RB gerada, observou-se que cada fator de risco pode ser relacionado a determinadas complicações, e desse modo, é possível estimar quais são os fatores de risco mais freqüentes e a quais complicações estão relacionados.

Ressalta-se o potencial das RBs em domínios de aplicações na área Médica que não envolvam diagnóstico, como por exemplo, em aplicações voltadas a área de Bioestatística, voltando-se à questões que envolvem fatores de risco e complicações.

5. Referências

Nassar, S. M. (2003), *Tratamento de Incerteza: Sistemas Especialistas Probabilísticos*. Disponível em: <http://www.inf.ufsc.br/~silvia/disciplinas/sep/MaterialDidatico.pdf>.

Barreto, J.M. (2001), *Inteligência Artificial: No limiar do Século XXI,* Florianópolis: J. M. Barreto PPP edições.

Stein, C. E. (2000), *Sistema Especialista Probabilístico: Base de Conhecimento Dinâmica*. 91 f. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Santa Catarina, Florianópolis, 2000.

Aurélio, M., Vellasco, M., Lopes, H. (1999), *Descoberta de Conhecimento e Mineração de Dados*. Disponível em: <http://www.ica.ele.pucrio.br//cursos/download/DM-apostila.pdf>.

Acid, S.; Campos, L. M. (1996), *Benedict: an algorithm for learning probabilistic belief Networks*. Granada, 1996. Disponível em: <http://decsai.ugr.es/gte/tr.html>.