Mineração de Dados Aplicada à Modelagem da Incidência da Anemia Infecciosa Equina (AIE), no Pantanal Sul-matogrossense

Helano P. de Lima[1], Urbano G. P. de Abreu[2], Stanley R. de M. Oliveira[1], Silvia M.F.S. Massruhá[1]

[1] Embrapa Informática Agropecuária, [2]Embrapa Pantanal
Caixa Postal 6041, CEP 13083-970 – Campinas – SP – Brazil

{helano.lima, urbano.abreu, stanley.oliveira, silvia.massruha}@embrapa.br

Abstract. Embrapa Pantanal conducted a control program of the EIA, which involved 28 farms forming a data set with 3,285 equidaes.The aim of this work was to model the problem of IEA using the induction technique of decision trees. We performed a step of pre-processing of experimental data and three decision tree models were induced, considering different levels of pruning. The overall accuracy of the models was around 85%, while the accuracy for the target class was 80%. The number of tests was the attribute that stood out, revealing a direct relationship with the correct monitoring of the program.

Resumo.A Embrapa Pantanal conduziu um programa de controle da AIE, que envolveu 28 fazendas formando um conjunto de dados de 3.285 equídeos. Objetivou-se neste trabalho, modelar o problema da AIE utilizando a técnica de indução de árvores de decisão. Foi realizada uma etapa de pré-processamento dos dados do experimento e três modelos de árvore de decisão foram induzidos, considerando níveis de poda diferentes. A acurácia geral dos modelos girou em torno de 85%, enquanto a precisão para a classe alvo foi de 80%. A quantidade de exames foi o atributo que mais se destacou, revelando uma relação direta com o correto acompanhamento do programa.

1. Introdução

A Anemia Infecciosa Equina (AIE) é causada por um retrovírus pertencente à subfamília dos lentivírus, o qual infecta membros da família *Equidae*. A AIE compromete irreversivelmente o desempenho dos equídeos, afetando indiretamente a pecuária extensiva e, sendo até o momento, uma doença incurável [Abreu et al., 2004a]. Segundo Hammer (1999), a AIE é uma doença bastante disseminada na América Central e do Sul. Em algumas regiões ou países o percentual de equídeos positivos para a doença varia entre 30% e 40%. No Pantanal, Silva et al. (2001) verificaram uma prevalência média de 24,8% em fazendas monitoradas, no período entre 1990 e 1995. Mais recentemente, Borges et al. (2013) observaram 31,5% de prevalência no município de Poconé no Pantanal de Mato Grosso.

A transmissão de AIE é relacionada, principalmente, à transferência de sangue de um animal infectado a outro equino sadio, sua disseminação está intimamente ligada

às práticas de manejo e sanidade do rebanho. A Embrapa Pantanal conduziu um programa de controle da AIE na região, que envolveu 28 fazendas no período entre 1990 e 1995. Silva, et al. (2001) descreveram que, trimestralmente, foram coletadas amostras de sangue da grande maioria dos equídeos das propriedades que participaram do programa de controle; após a coleta inicial, somente amostras dos animais negativos eram coletadas. O objetivo deste trabalho foi modelar, utilizando a técnica de indução de árvores de decisão, os resultados do acompanhamento da AIE realizado pela Embrapa Pantanal, traçando o perfil do sucesso ou fracasso do controle da doença.

2. Material e Métodos

Do acompanhamento, formou-se um conjunto de dados sobre exames realizados em 3.285 equídeos com atributos referentes ao nome, sexo, ano de nascimento e função na fazenda (classe), bem como informações sobre o exame, sendo mês e ano de execução, qual fazenda foi realizado, a quantidade de exames já realizados para o cavalo (ordem) e o resultado sorológico, totalizando 8.701 instâncias.

Tais dados já haviam sido analisados usando métodos estatísticos, porém, observando cada variável separadamente. Desejava-se obter um modelo que fosse capaz de levar em conta o conjunto das variáveis do problema. Dentre as técnicas de mineração de dados, as árvores de decisão mostram-se particularmente úteis neste contexto, pois permitem explicitar no modelo o conhecimento induzido, sendo a escolha para este trabalho.

Os dados passaram por processo de limpeza de dados descritas por Abreu et al. (2004b). Algumas instâncias cujos atributos de datas continham valores inconsistentes tiveram estes valores substituídos por valores vazios.

Segundo comparações efetuadas por Batista e Monard (2003) quanto ao tratamento de valores ausentes, o desempenho do algoritmo dos K-vizinhos mais próximos *(KNN*) é mais eficiente em relação a outros métodos de imputação na grande maioria dos cenários testados, provando ser um método robusto e eficaz, até mesmo em relação ao do tratamento de valores ausentes embutido no algoritmo de indução de árvores de decisão *C4.5* [Quinlan, 1988], sendo o algoritmo utilizado nesta modelagem.

O método foi efetuado com objetivo de entender a magnitude do problema dos valores ausentes e se havia aleatoriedade no processo. Optou-se por utilizar os agrupamentos das fazendas conforme o observado por Abreu et al. (2004a), que aplicaram o procedimento do vizinho mais próximo da análise de tipologia a partir de estimativas de escores fatoriais das fazendas para delimitá-las. Isso facilitaria a análise e validação do resultado obtido, pois obteria uma árvore mais concisa e permitiria comparar os resultados com os obtidos em Abreu et al. (2004a) e (2004b). Os Grupos formados podem ser vistos na Figura 1, onde os Grupos 1, 2 e 3 foram formados por uma única fazenda cada, o Grupo 4 foi formado por três fazendas e o Grupo 5 foi formado por 14 fazendas.

Posteriormente, foram efetuadas transformações nos atributos do conjunto de dados, visando adequar o mesmo para a tarefa de mineração de dados e ao que se desejava obter como resultado. Alguns atributos foram agregados, discretizados ou calculados partindo dos dados. Alguns atributos foram descartados manualmente, pois não contribuiriam para a classificação do atributo meta após as transformações. Para a modelagem, foi utilizada a implementação do algoritmo *C4.5* dentro da ferramenta *Weka* versão 3.6.5 [Hall, 2009]. Foi induzido um modelo de árvore de decisão a partir dos dados transformados utilizando-se o algoritmo *C4.5*. Este, por sua vez, nos permite generalizar ou especializar o modelo através de pós-poda. Foram geradas então três árvores, uma com pouca poda, uma com poda intermediária e outra mais podada. Objetivou-se com isso, fornecer vários níveis de detalhamento para facilitar a interpretação dos resultados.

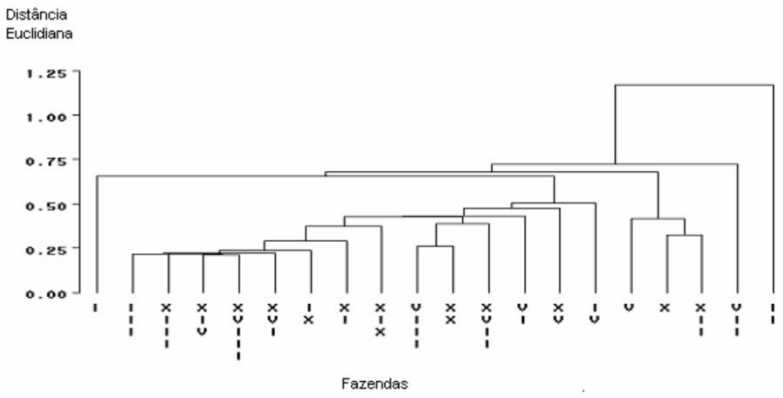

Figura 1. Agrupamentos das fazendas (Fonte: Abreu, et al., 2004a).

Foi utilizada a técnica de validação cruzada, pois esta permite que se use a totalidade do conjunto de dados tanto para o treinamento quanto para o teste. Para a fase de pós-podamento, foi variado o parâmetro fator deconfiança do algoritmo, segundo o sugerido por Witten (2011), usando os valores de 25% para uma poda razoável, mas foram produzidas árvores com os valores 50% (menos poda) e 10% (mais poda).

3. Resultados e Discussão

Observou-se que a acurácia geral dos três modelos foi bastante semelhante, girando em torno de 85%, que pode ser considerada muito boa dada as características do problema e do objetivo, que era extrair conhecimento, e não gerar um classificador preciso. Mais importante que isso, a precisão para a classe de interesse (Positivo) também se mostrou elevada, apresentando uma precisão aproximada de 78% para a árvore mais podada e 80% para as outras duas. Ambos os efeitos eram esperados, visto que a árvore mais sintética tende a errar mais, e com o fator de confiança de 25% (Figura 2) obteve-se o melhor resultado na maioria dos casos [Witten et al., 2011]. Isso se deve ao fato de que quando se trata de classificadores, deve-se buscar um limiar que evite o superajuste do

modelo aos dados. Uma comparação das medidas de precisão dos modelos induzidos pode ser vista na Tabela 1.

Tabela 1. Medidas de precisão para os três modelos gerados.

Medida	Fator conf. 10%		Fator conf. 25%		Fator conf. 50%	
Folhas	26		30		39	
Tam. Árvore	35		40		52	
Instâncias	2518					
Class. Corretamente	2135	0,848	2149	0,853	2154	0,845
Class. Incorretamente	383	0,152	369	0,147	364	0,145
Estatística Kappa	0,604		0,615		0,622	
Erro Médio Absoluto	0,229		0,225		0,219	
Erro Médio Quadrático	0,342		0,339		0,336	
Acurácia por classe	Positivo	Negativo	Positivo	Negativo	Positivo	Negativo
Taxa de VP	0,641	0,930	0,637	0,939	0,649	0,937
Taxa de FP	0,070	0,359	0,061	0,363	0,063	0,351
Precisão	0,783	0,844	0,805	0,867	0,802	0,871

Aparecendo no segundo nível das árvores encontra-se a classe, onde se observa que para as classes Reprodução e Redomão (equinos em processo de doma) o resultado é Negativo imediatamente. Isso corrobora os trabalhos de Abreu et al. (2004b), Hammer (1999) e Silva (2001), que verificaram que as classes de animais menos manejadas pelo homem foram as que apresentaram menos contaminação pela doença. Entretanto, quanto à classe Redomão, segundo o observação de Abreu et al. (2004b), este fato pode estar ligado à classe de Redomão serprovisória, onde os animais permanecem por pouco tempo, mas nas fazendas com manejo inadequado a incidência da doença pode passar de 10% nesta classe. Isto se comprova no modelo, pois na árvore menos podada esta folha é expandida para o atributo grupo_faz, sendo que para as fazendas do Grupo 1, o resultado é Positivo. Analogamente, na árvore menos podada, nas fazendas do Grupo 1, até para os animais da classe Reprodução e do sexo masculino (M) o resultado é Positivo. Conforme Abreu et al. (2004b), as probabilidades de infecção e os percentuais de exames positivos da classe Chucro são relativamente altos, entretanto, pôde ser visto nas três árvores geradas, a ramificação, que parte deste nó é a que mais se expande, apresentando características completamente distintas entre os grupos de fazendas acompanhadas.

Em relação às fazendas, nota-se no modelo que as fazendas do Grupo 4 são as únicas sempre apontadas com resultado Negativo, denotando terem alcançado o controle da AIE por meio do programa de controle da Embrapa Pantanal de maneira satisfatória (Abreu et al., 2004a). As fazendas do Grupo 1, possuem a maioria dos ramos com resultado Positivo, sendo este resultado encontrado para todas as classes de animais. Entretanto, na árvore com menos poda, aparece um ramo para a classe Reprodução indicando um resultado positivo para animais machos, sugerindo que talvez uma parte dos animais de reprodução machos deste grupo de fazendas seja usada como animais de serviço também, contribuindo para a não obtenção do controle da AIE.

As fazendas dos Grupos 2 e 3 apresentaram comportamentos semelhantes no modelo, tendo resultados Negativo para as classes de animais Redomão e Reprodução e resultados Positivo para as classes Chucro e Serviço. Essa igualdade permanece se observada somente a relação percentual entre animais positivos e negativos, mas um olhar mais atento nos dados revela que nas fazendas do Grupo 2, a grande parte dos animais negativos teve uma quantidade de exames razoável, enquanto nas fazendas do Grupo 3, a maior parte dos animais teve uma pequena quantidade de exames realizados, sugerindo que as fazendas do Grupo 2 esforçaram-se mais para seguir o controle.

Figura 2. Árvore com poda moderada (fator de confiança 25%).

No Grupo 5, o mais numeroso, formado por 14 fazendas, a maioria dos resultados foi Positivo, sendo notado que apenas os equídeos com idade Novo mostraram-se livre de contaminação nestas. Este resultado vai ao encontro com o relatado por Abreu et al. (2004a), que observaram que neste grupo encontravam-se as fazendas que não obtiveram sucesso no cumprimento do Programa de controle da AIE.

A utilização de técnicas de estatística multivariada, e de algoritmos de aprendizado de máquina para análise de dados epidemiológicos veterinários é recente no Brasil. O uso combinado das técnicas na avaliação de vigilância permite o monitoramento de tendências, o progresso do controle da doença, a avaliação de intervenções e programas preventivos e a detecção de surtos da doença. A coleta, análise e interpretação de forma continua e sistemática de dados sobre saúde animal são essenciais para o planejamento, e a implementação das prática de prevenção. Em paralelo, a observação oportuna de determinados padrões que indiquem a possibilidade de surtos em estágios precoces, permite aos tomadores de decisões implementar medidas sanitárias em tempo real. O que caracteriza a vigilância epidemiológica

sindrômica com significativos ganhos na eficiência de prevenção das doenças [Dórea, et al., 2013].

4. Conclusões

A técnica de indução de árvores de decisão através do algoritmo *C4.5* mostrou-se adequada ao estudo do acompanhamento do Programa de Prevenção e Controle da AIE no Pantanal, permitindo identificar entre os grupos de fazendas, os indicadores de sucesso ou fracasso. O processo de preparação de dados também se mostrou adequado aos requisitos da modelagem.

A classe de animais Chucros, conforme o resultado diversificado obtido no modelo mostra-se carente de melhor investigação, pois é a classe menos manejada pelo homem e hipoteticamente, deveria ter uma prevalência maior de animais negativos.

5. Referências

Abreu, U. G. P. de; Silva, R. A. M. S.; Barros, A. T. M. de. (2004a) "Avaliação do Controle da Anemia Infecciosa Eqüina em Fazendas na Sub-Região da Nhecolândia, Pantanal Sul-Mato-Grossense". In: *Simpósio sobre Recursos Naturais e Socioeconômicos do Pantanal*, 4. Corumbá. (cd rom)

Abreu, U. G. P. de; Silva, R. A. M. S.; Barros, A. T. M. de. (2004b) "Modelagem da probabilidade de incidência da anemia infecciosa equina em fazendas da Sub-região da Nhecolândia, Pantanal Sul-Mato-Grossense". In: *Simpósio sobre Recursos Naturais e Socioeconômicos do Pantanal*, 4. Corumbá. (cd rom)

Batista, G. E. A. P. A.; Monard, M. C. (2003) "An analysis of four missing data treatment methods for supervised learning", Applied Artificial Intelligence, v. 17, p. 519-533.

Borges, A. M. C. M. ; Silva, L. G. ; Nogueira, M. F. ; Oliveira, A. C. S. ; Segri, N. J. ; Ferreira, F. ; Witter, R. ; Aguiar, D. M. (2013) "Prevalence and risk factors for Equine Infectious Anemia in Poconé municipality, northern Brazilian Pantanal". Research in Veterinary Science, v. 95, p. 76-81.

Dórea, F. C.; McEwen, B. J.; McCnab, W. B.; Revie, C. W.; Sanchez, J.(2013) "Syndromic surveillance using veterinary laboratory data: data pre-processing and algorithm performance evaluation". Journal of the Royal Society Interface, v. 10, p. 20130114-20130114.

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. (2009) "The WEKA Data Mining Software: An Update". SIGKDD Explorations, Volume 11, Issue 1. New York, USA.

Hammer, M. (1999) "Controlling equine infectious anaemia". Animal Research Development, v. 50, p. 44-57.

Quinlan, J. R. (1988) "C4.5 Programs for Machine Learning". Morgan Kaufmann, CA, USA.

Silva, R. A. M. S.; Abreu, U. G. P. de; Barros, A. T. M. de. (2001) "Anemia infecciosa equina: epizootologia, prevenção e controle no Pantanal". Corumbá: EmbrapaPantanal, 30 p. (EmbrapaPantanal. Circular Técnica, 29).

Witten, I. H.; Frank, E.; Hall, M. (2011) "Data Mining: Practical Machine Learning Tools and Techniques", San Francisco, Morgan Kaufmann, 3rd edition.