Predição de Indicadores Zootécnicos de Carcaças Bovinas a Partir de Variáveis de Cria

Denizar S. Souza
URCAMP (Bagé/RS)
Av. Tupy Silveira, 2099
+55 53 32428244
denizarsouza@urcamp.edu.br

Thales V. Maciel
IFSul (Bagé/RS)
Av. Leonel Brizola, 2501
+55 53 32473237
thalesmaciel@ifsul.edu.br

Vinícius do N. Lampert
EMBRAPA (Bagé/RS)
BR 153, Km 603
+55 53 32404650
vinicius.lampert@embrapa.br

Rodrigo R. da Silva
IFSul (Bagé/RS)
Av. Leonal Brizola, 2501
+55 53 32473237
orki2008@gmail.com

RESUMO

Este artigo descreve uma metodologia para obtenção de árvores de decisão para previsão de indicadores zootécnicos de qualidade de carcaças bovinas com base em variáveis de cria dos animais. Para tal, procedeu-se a tarefas de mineração de dados com classificação após pré-processamento com discretização dos atributos numéricos por particionamento igualitário do intervalo ou por descoberta de agrupamentos em experimentos distintos de classificação. Os resultados obtidos mostraram que a descoberta de agrupamentos como forma de discretização pode gerar classes com balanceamento de melhor qualidade em comparação ao método tradicional, permitindo a indução de modelos utilizáveis em problemas reais.

Palavras-chave

pecuária; árvore de decisão; qualidade; zootécnica

ABSTRACT

This paper describes a method for obtaining decision trees for predicting carcasse zootechnical quality indicators for bovine based on their breeding data. For such, data mining classification tasks were performed after data preprocessing, where all numeric attributes were discretized by non-equal frequency binning or by cluster discovery in distinct classification experiments. Obtained results showed that clustering techniques as means for discretization may generate classes in better balancing conjecture when in comparison to the non-equal frequency binning method, allowing the discovery of models that may be applied to real world problems.

Keywords

livestock; decision tree; quality; zootechnical

1. INTRODUÇÃO

O sistema de produção de gado de corte é o conjunto de tecnologias e práticas de manejo, tipo de animal, propósito de criação, raça e ecorregião onde a atividade é desenvolvida [1]. Compreende uma das principais atividades de exploração econômica no Brasil, onde, há décadas, tem-se afastado o cenário de resistência ao emprego tecnológico, de modo a permitir estudos para o melhoramento dos índices de qualidade na produção de carne, por exemplo, através de computação aplicada [2].

Em [3], foram analisados dados zootécnicos de 401 animais bovinos da raça Hereford com vistas em prever o peso de fazenda e bonificação dos indivíduos. No estudo, foram empregadas redes neurais artificiais como ferramenta para o processo de descoberta de conhecimento em experimentos distintos para as duas variáveis. Todos os dados envolvidos foram do tipo numérico. Segundo o autor, o trabalho obteve resultados satisfatórios na previsão do peso de fazenda, mas insatisfatórios na previsão da bonificação, atribuindo o não cumprimento do objetivo específico à má qualidade de dados. O estudo não considerou a praticidade da utilização de redes neurais artificiais pelos produtores pecuários em meio às tarefas cotidianas, tampouco apresentou comparações com outros métodos para descoberta de conhecimento em bancos de dados.

Em [4], foram empregadas tecnologias de armazém de dados, consultas analíticas e mineração de dados para 1142230 registros de abates bovinos. O objetivo foi o de prever o grau de acabamento e o rendimento das carcaças em experimentos individuais, que foram conduzidos com algoritmos de classificação e redes neurais artificiais. Os resultados, segundo os autores, foram promissores, devido às acurácias alcançadas nos experimentos, cuja média em acertos de classificação foi de 62%. Embora tenham composto médias de acurácias da aplicação de diferentes algoritmos nas tarefas preditivas, os autores não apresentaram a comparação das acurácias dos algoritmos utilizados individualmente, tampouco avaliações aprofundadas dos resultados, que fossem além das acurácias observadas nos experimentos, ou apresentar os modelos gerados, que sriam passíveis desta avaliação.

Nota-se que trabalhos correlatos publicados recentemente, mesmo que parcialmente eficazes segundo os respectivos autores, não explicam as predições realizadas pelos experimentos que documentam, ou pela impossibilidade disto ser característica do algoritmo empregado (caixa-preta) ou por não apresentar a totalidade dos resultados da classificação nos resultados obtidos nos testes (matrizes de confusão, por exemplo).

O problema de pesquisa abordado no presente estudo é fundamentado em “quais variáveis podem ser coletadas, pelos criadores, sobre os indivíduos de rebanhos bovinos em etapa de desenvolvimento de cria e que explicam a obtenção de indicadores de qualidade zootécnicos das carcaças ótimos após o abate?”

A hipótese trabalhada é que existe uma relação estatística entre o mês de nascimento, o mês de desmame, a idade de desmame e o peso de desmame com o peso e a idade de abate.

SBSI 2018, June 5th–8th, 2018, Lavras, Minas Gerais, Brazil.

O objetivo é obter um modelo gráfico, de fácil interpretação, capaz de orientar os criadores bovinos sobre o desempenho de seus rebanhos, ainda em etapa anterior ao desmame, com previsões dos futuros índices de qualidade que são obtidos apenas após o abate dos animais.

2. METODOLOGIA

Procedeu-se à descoberta de conhecimento em bancos de dados (DCBD), especificamente com as tarefas de mineração de dados descritas nesta seção.

O processo de DCBD pode ser dividido em três etapas [5]: o pré-processamento, onde o conjunto de dados original é preparado para as próximas etapas do processo através de tarefas de filtragem conforme necessário; o processamento, onde algoritmos de mineração de dados são aplicados sobre o conjunto de dados pré-processado e; o pós-processamento, onde os padrões descobertos no processamento são analisados e transformados em conhecimento útil sobre o domínio estudado.

Para fins de realização das tarefas e experimentos descritos neste estudo, foi empregado o Waikato Environment for Knowledge Analysis (WEKA) [6]. Trata-se de uma coleção de implementações de algoritmos que podem ser utilizados em atividades de mineração de dados diversas, como classificação, regressão, associação e clustering, pré-processamento de dados e visualização de resultados através de interface gráfica, linha de comando ou interface de programação [7].

O conjunto de dados analisado teve sua apresentação original em 167 instâncias de animais bovinos e 6 atributos, conforme descrição na tabela 1.

Tabela 1. Descrição do conjunto de dados analisado

Nome do Atributo	Significado	Tipo de Dado	Intervalo
nascimento_mes	mês de nascimento (01-12)	nominal	-
desmame_mes	mês de desmame (01-12)	nominal	-
desmame_idade	idade de desmame em meses	numérico	2 a 8
desmame_peso	peso de desmame em quilogramas	numérico	78 a 242
abate_idade	idade de abate em meses	numérico	19 a 42
abate_peso	peso de abate em quilogramas	numérico	352 a 574

No pré-processamento, os atributos numéricos foram discretizados, conforme a escala de Likert [8], de forma a criar segmentos nominais dentre o intervalo numérico com as denominações: muito baixo, baixo, intermediário, alto e muito alto.

Discretização é o particionamento de um intervalo numérico e sucessiva atribuição de um valor categórico como rótulo de cada partição criada [7]. No âmbito deste estudo, dois métodos de discretização distintos foram experimentados.

No primeiro deles, os atributos numéricos tiveram seus intervalos divididos em 5 frações iguais, para a criação de 5 classes, sem balanceamento na distribuição de frequência. No segundo, as classes foram descobertas de forma automatizada por aplicação do algoritmo Simple k-means [9] sobre cada atributo numérico individualmente, com base na distância de Manhattan.

As figuras 2 e 3 respectivamente apresentam os histogramas referentes às distribuições de frequência das instâncias de bovinos nas categorias propostas pela escala de Likert nos atributos discretizados pelos métodos do fracionamento igualitário do intervalo numérico e com a descoberta automatizada dos agrupamentos baseados na distância de Manhattan.

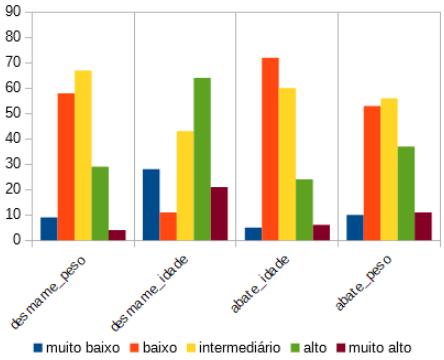

Figura 1. Histogramas referentes aos atributos discretizados por segmentação

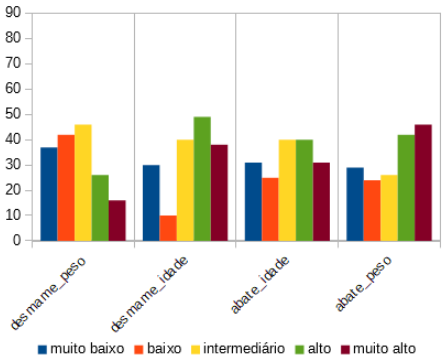

Figura 2. Histogramas referentes aos atributos discretizados por descoberta de agrupamentos.

No total, foram realizados 4 experimentos de predição. Neles, as variáveis de cria (mês de nascimento, mês de desmame, peso de desmame e idade de desmame) foram utilizadas para prever os indicadores zootécnicos de qualidade das carcaças (idade de abate e peso de abate) em experimentos distintos.

Cada indicador zootécnico de qualidade das carcaças foi alvo de predição após discretização dos atributos numéricos através dos dois métodos apresentados. A tabela 2 elenca as atividades de pré-processamento pelas quais cada atributo foi submetido em cada experimento. Os atributos cujo tipo de dado original é o nominal não passaram pois quaisquer tarefas de pré-processamento (N/A). Em cada experimento, houve, ainda, a remoção dos atributos não referentes à variáveis de cria e o próprio atributo alvo de predição, que é destacado em sublinhado em cada linha da tabela 2.

A mineração de dados é a tarefa de identificação de padrões a partir de dados, de forma automatizada em ambiente computacional, que compreende a etapa de processamento no processo de DCBD [5].

A classificação é um tipo de tarefa da mineração de dados que visa categorizar instâncias supostamente novas com base na análise de dados de instâncias pregressas [7]. Há a etapa de treinamento, onde um algoritmo aprende as características inerentes à cada classe e a etapa de teste, onde é verificada a acurácia do modelo criado.

Árvores de decisão são um tipo de modelo de dados utilizado como resultado de tarefas de classificação [10] e apreciado no contexto deste estudo em virtude de sua simplicidade e interpretabilidade.

Tabela 2. Descrição das tarefas de pré-processamento aplicadas a cada atributo do conjunto de dados nos 4 experimentos realizados.

#	nascimento_mes	desmame_mes	desmame_peso	desmame_idade	abate_idade	abate_peso
1	N/A	N/A	segm.	segm.	segm.	removido
2	N/A	N/A	segm.	segm.	removido	segm.
3	N/A	N/A	agrup.	agrup.	agrup.	removido
4	N/A	N/A	agrup.	agrup.	removido	agrup.

O J48 [6, 10, 11] é um algoritmo de mineração de dados, especificamente para tarefas de classificação, capaz de induzir árvores de decisão, sendo um dos algoritmos mais utilizados em aplicações do tipo no mundo real.

A etapa de processamento em todos experimentos foi realizada com o algoritmo J48 configurado para permitir apenas divisões binárias em galhos formados por atributos nominais e desconsiderar limites inferiores de ocorrências de instâncias em folhas para critérios de poda em seu treinamento. Os demais parâmetros do algoritmo foram mantidos em conformação padrão. A etapa de testes do modelo descoberto foi realizada sobre o mesmo conjunto de dados de entrada para treinamento, devido à baixa representatividade de algumas das classes apresentadas em diversos atributos do conjunto de dados.

3. RESULTADOS OBTIDOS

Os resultados obtidos nas tarefas de classificação descritas na seção 2 foram apresentados na forma de árvores de decisão, matrizes de confusão e acurácias dos respectivos modelos. Também é discutida a praticidade dos modelos descobertos. As acurácias alcançadas em todos experimentos realizados neste estudo estão dispostas na tabela 3.

Tabela 3. Acurácias resultantes dos experimentos realizados

Experimento	#1 (%)	#2 (%)	#3 (%)	#4 (%)
Acurácia	63,47	49,70	53,29	51,50

Foi feita comparação dos resultados obtidos na classificação com os conjuntos de dados cujos atributos numéricos foram discretizados por segmentação (segm.) igualitária dos intervalos numéricos os conjuntos de dados cujos atributos numéricos foram discretizados pela descoberta automatizada de agrupamentos (agrup.) por aplicação do algoritmo Simple k-means. Foram analisadas as acurácias e matrizes de confusão resultantes dos experimentos, onde foram evidenciadas falhas cruciais em alguns dos modelos gerados. A tabela 4 apresenta as matrizes de confusão encontradas nos testes.

Para previsão da idade de abate, foram realizados dois experimentos, #1 e #3, cujos modelos gerados apresentaram acurácias de 63,47% e 53,29% respectivamente, com diferença de 10,18% em favor do primeiro. Contudo, a matriz de confusão referente ao experimento #1 evidencia a incapacidade do modelo em classificar as instâncias nas idades de abate muito baixa e muito alta, o que sobremaneira impede que o mesmo tenha proveito prático em alinhamento com os objetivos do presente trabalho. Esta problemática não foi presente nos resultados obtidos no experimento #3, cuja árvore de decisão resultante é apresentada na figura 3.

Tabela 4. Matrizes de confusão resultantes dos experimentos realizados

	abate_idade					**abate_peso**				
classe verdadeira\prevista	mb	b	i	a	ma	mb	b	i	a	ma
muito baixo (mb)	0	5	0	0	0	0	3	6	1	0
baixo (b)	0	57	11	4	0	0	25	22	6	0
intermediário (i)	0	20	32	8	0	0	9	43	4	0
alto (a)	0	5	2	17	0	0	6	16	15	0
muito alto (ma)	0	5	0	1	0	0	1	6	4	0
	Experimento #1					**Experimento #2**				
classe verdadeira\prevista	mb	b	i	a	ma	mb	b	i	a	ma
muito baixo (mb)	24	1	3	0	3	16	1	5	2	5
baixo (b)	11	16	4	2	7	2	6	3	1	14
intermediário (i)	6	1	15	3	6	3	3	20	2	14
alto (a)	5	1	1	11	7	1	0	6	7	10
muito alto (ma)	7	3	2	5	23	1	3	4	1	37
	Experimento #3					**Experimento #4**				

A raíz da árvore de decisão apresentada na figura 3 propõe que o atributo mais relevante para predição da idade de abate é a idade de desmame e, no caso desta ser muito baixa, deve ser considerado o mês de nascimento. Nesta hipótese, e de que o mês de nascimento seja outubro, o modelo prediz com 47% de precisão que a idade de abate é alta, mas caso o mês de nascimento seja diferente de outubro, o modelo prediz com 62% de precisão que a idade de abate é intermediária. Maiores interpretações, sobre os demais galhos da árvore, podem ser feitas de forma semelhante.

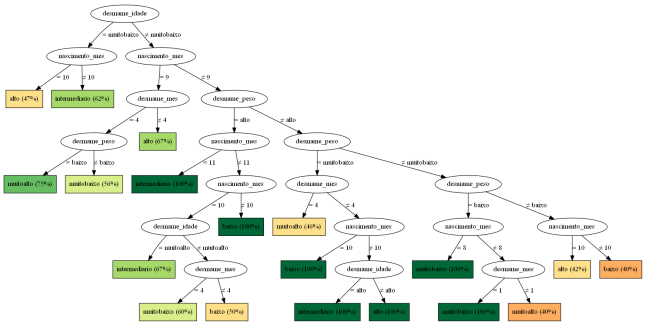

Figura 3. Árvore de decisão para predição da idade de abate

Os experimentos #2 e #4, referentes à previsão do peso de abate, apresentaram acurácias de 49,70% e 51,50% respectivamente, com diferença de 1,80% em favor do segundo. Observou-se que o experimento #2, além de não ter logrado melhor acurácia em comparação com o experimento #4, expõe a mesma problemática apresentada pelos resultados do experimento #1. À exemplo deste, o experimento #2 foi incapaz de classificar as instâncias de animais bovinos nas categorias muito baixo e muito alto, neste caso acerca do peso de abate. A árvore de decisão resultante do experimento #4 é apresentada na figura 4.

Novamente, na árvore de decisão apresentada na figura 4, o atributo de maior relevância foi a idade de desmame. O modelo descreve que, caso este atributo contenha um valor muito baixo, logo, o peso de abate é muito alto, tendo precisão de 50% nesta predição. Caso a idade de desmame não seja muito baixa, outras possibilidades são apresentadas.

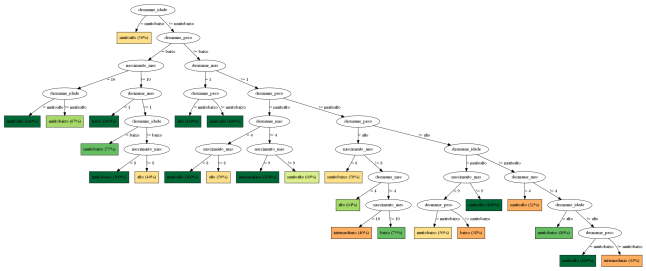

Figura 4. Árvore de decisão para predição do peso de abate

Nos modelos descobertos (figuras 3 e 4), as árvores de decisão tiveram os nós folhas destacados em uma escala cromática de vermelho a verde, denotando, respectivamente, baixa e alta precisão para classificação em cada folha. A escala conta, ainda, com a cor amarela, representando valores intermediários na escala. Esta formatação dos modelos, em cores, foi concebida com vistas em melhorar a legibilidade dos mesmos perante profissionais de domínios de negócios específicos que buscam fácil e rápida avaliação das predições realizadas pelos modelos.

Em sumário, nos testes realizados, os conjuntos de dados cujos atributos numéricos foram discretizados pela descoberta automatizada dos 5 agrupamentos com aplicações do algoritmo Simple k-means foram as entradas mais adequadas para processamento com tarefas de classificação com o algoritmo J48 sobre as variáveis idade de abate e peso de abate.

Isto ocorreu em virtude do desbalanceamento entre as frequências das categorias após as tarefas de discretização, que podem ser observadas nos histogramas apresentados nas figuras 1 e 2. Estas figuras evidenciam as diferenças de balanceamento das classes e permitem a comparação dos resultados da discretização para os dois métodos de discretização utilizados sobre os atributos originalmente numéricos. Entende-se que o desbalanceamento observado em alguns histogramas são a causa da impossibilidade do algoritmo em gerar modelos que não negligenciam quaisquer classes a partir dos respectivos conjuntos de dados.

4. CONCLUSÃO

O presente trabalho buscou um método para descoberta de árvores de decisão capazes de auxiliar os produtores de gado de corte na previsão de indicadores zootécnicos da qualidade das carcaças com base nas respectivas variáveis de cria, ou seja, dados que podem ser coletados entre o nascimento e o desmame dos animais.

Foram realizados experimentos de classificação com o algoritmo J48 após pré-processamento do conjunto de dados para discretização dos atributos numéricos. A discretização pelo método tradicional se mostrou problemática ao produzir categorias com frequências desbalanceadas. Diante desta situação, foi proposto que as tarefas de discretização fossem realizadas através da descoberta automatizada das categorias por medida da distância de Manhattan em aplicação do algoritmo Simple k-means, o que melhorou a qualidade do balanceamento entre as categorias criadas.

Finalmente, foi possível descobrir árvores de decisão capazes de explicar a influência das variáveis de cria mês de nascimento, mês de desmame, peso de desmame e idade de desmame nos indicadores zootécnicos de qualidade de carcaças, como idade de abate e peso de abate, além de fazê-lo de forma didática, através do emprego de escalas de cores para denotar a precisão de classificação em cada possibilidade proposta no modelo. Outrossim, considera-se que o objetivo do trabalho foi cumprido de forma satisfatória.

Trabalhos futuros envolvem novos esforços pela coleta de dados, com vistas no aumento de representatividade dos dados localizados nos extremos das distribuições de frequência analisadas, de forma a viabilizar a realização de testes dos modelos com conjuntos de dados diferentes daqueles de treinamento. Posteriormente, será investigado o melhoramento da acurácia dos modelos descobertos, o que pode ser abordado por diferentes métodos de discretização dos atributos numéricos, diferentes métodos de descoberta de agrupamentos para aplicações em discretização, experimentações com outros algoritmos de indução de árvores de decisão, empilhamento de classificadores e aprendizado sensível à custo.

5. REFERÊNCIAS

[1] Euclides Filho, K. (2000) Produção de bovinos de corte e o trinômio genótipo-ambiente-mercado. Embrapa Gade de Corte - Documentos (Infoteca-E).

[2] Barbosa, P. (1999) Raças e estratégias de cruzamento para produção de novilhos precoces. Embrapa Pecuária Sudeste. In: Simpósio de Produção de Gado de Corte, 1. Viçosa, Brasil.

[3] Costa, C. L. (2016). Utilização de caracteríticas zootécnicas e de manejo na pecuária para previsão do peso final e bonificação de bovinos empregando redes neurais artificiais. Tabalho de conclusão de curso, Universidade Federal do Pampa.

[4] Mota, F., Souza, K., Ishii, R. and Gomes, R. (2017) BovReveals: uma plataforma OLAP e data mining para tomada de decisão na pecuária de corte. In:: Congresso Brasileiro de Agroinformática, 11. Campinas, Brasil.

[5] Maciel, T., Seus, V., Machado, K. and Borges, E. (2015). Mineração de dados em triagem de risco de saúde. Revista Brasileira de Computação Aplicada, 7(2), 26-40.

[6] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. (2009) The WEKA Data Mining Software: An Update. SIGKDD Explorations, Volume 11, Issue 1.

[7] Witten, I., Frank, E., Hall, M. and Pal, C. (2017) Data mining: practical machine learning tools and techniques. Morgan Kaufmann.

[8] Likert, R. (1932) A Technique for the Measurement of Attitudes. Archives of Psychology. 140: 1–55.

[9] Arthur, D. and Vassilvitskii S. (2007) k-means++: the advantages of carefull seeding. In: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, 1027-1035.

[10] Quinlan, R. (1993) C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA.

[11] Quinlan, J. R. (1996). Improved use of continuous attributes in C4. 5. Journal of artificial intelligence research, 4, 77-90. Chicago, IL.