Aplicações de Mineração de Dados na Pecuária de Corte: Previsão de Indicadores de Qualidade de Carcaças

Rodrigo R. da Silva
Instituto Federal de Educação, Ciência e Tecnologia Sul-rio-grandense (IFSUL)
Av.Leonel de M. Brizola, 2501
(Bagé/RS)
orki2008@gmail.com

Thales V. Maciel
Instituto Federal de Educação, Ciência e Tecnologia Sul-rio-grandense (IFSUL)
Av.Leonel de M. Brizola, 2501
(Bagé/RS)
thalesmaciel@ifsul.edu.br

Vinícius do N. Lampert
Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA)
BR 153, Km 603
(Bagé/RS)
vinicius.lampert@embrapa.br

Denizar S. de Souza
Universidade da Região da Campanha (URCAMP)
Av. Tupy Silveira, 2099
(Bagé/RS)
denizarsouza@urcamp.edu.br

RESUMO

Considerando que o produtor rural pode obter algumas variáveis de influência ao longo do processo produtivo do gado de corte, objetiva-se prever se as variáveis de influência obtidas até o desmame dos bovinos podem explicar a bonificação, ganho médio diário, idade de abate e peso de fazenda. Para tanto procede-se a mineração de dados através da regressão linear, em um conjunto de dados de 167 bovinos. Deste modo, observa-se que para a bonificação e ganho médio diário os modelos gerados apresentaram erros baixos, enquanto que para idade de abate e peso de fazenda os erros foram maiores, o que permite concluir que os atributos não foram o suficientes para predizer a idade de abate e peso de fazenda, mas bons para a bonificação e ganho médio diário.

Palavras-Chave

Pecuária; mineração de dados; indicadores.

ABSTRACT

Considering that cattle breeders are able to acquire influence variables along the breeding process, this paper aims to provide a method for predicting carcasse bonus, daily average weight gain, age at slaughter and weight at slaughter based on influce variables to be collected until the bovine ablactation. For such, data mining applications were performed through linear regression applied to a 167 bovine instances dataset. Obtained results showed that carcasse bonus and daily average weight gain may be predicted with zero or insignificant error, maywhile age and weight at slaughter produced higher error rates upon prediction.

SBSI 2018 June 4^{th} – 8^{th}, 2018, Caxias do Sul, Rio Grande do Sul, Brazil

Keywords

Livestock; data mining; indicators

1. INTRODUÇÃO

No contexto da pecuária de corte, o sistema produtivo pode ser conceituado como um conjunto de tecnologias e práticas de manejo, bem como o perfil do animal, a intenção da criação, a raça ou grupamento genético e a região onde a atividade é desenvolvida [4].

Para analisar um sistema produtivo da pecuária de corte, é indispensável mensurar seus indicadores de qualidade, pois somente assim o produtor rural terá embasamento para tomada de decisão. Para [9] a mensuração e análise de indicadores que retratam o funcionamento rural são fundamentais para a tomada de decisão. Estes indicadores, de acordo com [8], são conhecidos como variáveis de influência, ou seja, informações gerenciais de ordem técnica ou econômica que contribuem com avaliações precisas dos processos internos da propriedade rural.

Ainda segundo [9], deve ficar claro que para a empresa rural, interessa, sobretudo, a rentabilidade, que é o elemento mais importante na avaliação da atividade econômica praticada em moldes capitalistas. Este indicador de desempenho deve situar-se em nível adequado para que o investimento se justifique. No âmbito do criador e das informações que estão acessíveis a ele, os indicadores devem possuir relevância para serem aplicados em situações de estudos de caso.

O problema de pesquisa abordado neste trabalho é "existem variáveis de cria, ou seja, dados coletados sobre indivíduos de rebanhos bovinos a partir do nascimento, que explicam bons indicadores de qualidade zootécnicas?". A hipótese é que os dados: ano de abate, ano de desmame, ano de nascimento, mês de nascimento, mês de desmame, peso de desmame e idade de desmame têm correlação suficiente com o peso de abate na fazenda, idade de abate, ganho médio

diário de peso e bonificação, para explicar bons valores em tais indicadores.

O objetivo deste trabalho é descobrir a relação estatística entre as variáveis de cria e os indicadores zootécnicos de qualidade de carcaças após abate. Quantificar o peso dos atributos e hipóteses dos respectivos domínios de valores nos indicadores de qualidade inferidos. Para tal, foram realizadas tarefas de mineração de dados no âmbito de descoberta de conhecimento em banco de dados. O foco da atividade ocorreu com experimentos de regressão, conforme descrito na metodologia.

2. REFERENCIAL TEÓRICO

Nesta seção é apresentado um referencial teórico sobre descoberta de conhecimento com mineração de dados, seguido de um levantamento de suas aplicações na pecuária de corte.

2.1 Descoberta de Conhecimento em Banco de Dados

Segundo [5], o processo de descoberta de conhecimento em bases de dados (DCBD) é definido como um processo não trivial que busca identificar padrões novos, potencialmente úteis, válidos e compreensíveis, com o objetivo de melhorar o entendimento de um problema ou um procedimento de tomada de decisão.

O processo de DCBD compreende três principais etapas: pré-processamento, mineração de dados e pós-processamento [10]. No pré-processamento os dados são coletados e tratados para serem utilizados nas próximas etapas. A limpeza e a remoção de dados ruidosos também ocorre no pré-processamento, visando assegurar a qualidade dos dados selecionados. Subsequentemente, ocorre a mineração de dados, que são processos aplicados para explorar e analisar os dados em busca de padrões, previsões, erros, associações entre outros [1]. A etapa final consiste no pós-processamento, que engloba a interpretação dos padrões descobertos e a possibilidade de retorno a qualquer um dos passos anteriores. Assim, a informação extraída é analisada (ou interpretada) em relação ao objetivo proposto, sendo identificadas e apresentadas as melhores informações [2]. As tarefas de mineração de dados podem ser dividas em quatro grupos: classificação, regressão, agrupamentos e regras de associação.

A regressão é um tipo específico de classificação. Enquanto a classificação trata de previsão de valores nominais ou categóricos, chamados de classes, a regressão mantém o objetivo de realizar previsões, mas tem como alvo valores numéricos. No agrupamento não existe classe, o objetivo é criar grupos e atribuir instâncias a estes grupos a partir de características, ou atributos destas instâncias. Regras de associação buscam relações entre os ítens, gerando regras que determinam a associação entre esses ítens [1]. Este estudo tem foco em tarefas de regressão.

2.2 Revisão dos Trabalhos Correlatos

No âmbito da pecuária de corte, foram identificados trabalhos relacionados ao problema investigado nesta pesquisa.

No trabalho [7], foram utilizadas duas ferramentas computacionais para fins de auxiliar tomadas de decisões na produção de bovinos de corte, criados de maneira extensivas, em condições de manejo encontrados no Brasil. A primeira parte do trabalho visou à construção de um software utilizando a técnica de Simulação Monte Carlo para analisar características de produção (ganho de peso) e manejo (fertilidade, anestro pós-parto, taxa de natalidade e puberdade). Na segunda parte do trabalho foi aplicada a técnica de Redes Neurais Artificiais para classificar animais, segundo ganho de peso nas fases de crescimento (nascimento ao desmame, do desmame ao sobreano) relacionado com o valor genético do ganho de peso do desmame ao sobreano (GP345) obtidos pelo BLUP. Ambos modelos mostraram potencial para auxiliar a produção de gado de corte.

Na pesquisa de [11] foram utilizados dados de 19240 animais Tabapuã, provenientes de 152 fazendas localizadas em diversos estados brasileiros, nascidos entre 1976 e 1995, foram utilizados para predição do valor genético do peso aos 205 dias de idade (VG_P205) por meio de redes neurais artificiais (RNA's) e usando o algoritmo LM - Levenberg Marquardt - para treinamento dos dados de entrada. Por se tratar de rede com aprendizado supervisionado, foram utilizados, como saída desejada, os valores genéticos preditos pelo BLUP para a característica P205. Os valores genéticos do P205 obtidos pela RNA e os preditos pelo BLUP foram altamente correlacionados. A ordenação dos valores genéticos do P205 oriundos das RNA's e os valores preditos pelo BLUP (VG_P205_RNA) sugeriram que houve variação na classificação dos animais, indicando riscos no uso de RNA's para avaliação genética dessa característica. Inserções de novos animais necessitam de novo treinamento dos dados, sempre dependentes do BLUP.

Já no estudo de [3] foi analisado um conjunto de características zootécnicas para gerar um modelo a fim de prever o rendimento dos bovinos, através das variáveis peso de fazenda (PF) e bonificação (BN). Para tanto o autor utilizou a técnica de Redes Neurais Artificiais (RNA's). Segundo aponta o autor, o resultado para o modelo de previsão de bonificação apresentou erro bem elevado, baixa correlação e generalização insatisfatória devido a uma limitação da ferramenta e da escolha dos dados utilizados na matriz de entrada da rede. Cabe ressaltar o trabalho não proveu comparações de desempenho com outros métodos de inferência de dados, tampouco indicações de peso de cada variável de entrada no produto de saída.

O trabalho difere-se dos demais por usar a tarefa de regressão como técnica de processamento na descoberta de conhecimento, além disto, os atributos utilizados são diferentes, pois neste trabalho optou-se por analisar a influência das variáveis de cria em relação as variáveis de qualidade zootécnicas, contribuindo desta maneira para novas abordagens, relatos e discussões sobre a temática da pecuária de corte.

3. METODOLOGIA

O conjunto de dados analisado foi constituído pro 167 instâncias de animais bovinos da raça Hereford. As nomenclaturas e respectivas descrições dos atributos do conjunto analisado são apresentadas na Tabela 1.

Como ferramenta para a realização das tarefas de pré-processamento e aplicações dos algoritmos de mineração de dados, foi utilizados o *software Waikato Environment for Knowledge Analysis* (WEKA), um ambiente para análise de conhecimento desenvolvido pela Universidade de Waikato, Nova Zelândia [6].

O experimento realizado dividiu-se em três etapas. Na primeira etapa os dados foram recuperados em formato .CSV a fim de serem utilizados no *software* WEKA. O conjunto de

Tabela 1: Atributos selecionados

Nomenclatura	Tipo	Descrição
nascimento_ano	Nominal	Ano de nascimento(2010,2011,2012)
desmame_ano	Nominal	Ano de desmame(2011,2012,2013)
abate_ano	Nominal	Ano de abate(2013,2014)
nascimento_mes	Nominal	Mês de nascimento(1,8,9,10,11,12)
desmame_peso	Numerico	Peso de desmame
desmame_idade	Numerico	Idade de desmame
desmame_mes	Nominal	Mês de desmame(1,4,5)
abate_idade	Numerico	Idade de abate
gmd	Numerico	Ganho médio diário de peso
abate_peso	Numerico	Peso de abate na fazenda
bonificação	Numerico	Bonificação

dados original constava com 53 atributos e 1015 instâncias de bovinos de diversas raças. A etapa de pré-processamento deste conjunto de dados contou com tarefas de transformação, remoção de atributos irrelevantes, remoção atributos com dados faltantes e ruidozos, bem como os que não faziam parte do escopo dos experimentos, resultando no conjunto de dados descritos pela Tabela 1.

Após o WEKA ser alimentado com os dados, foi aplicado o filtro *weka.filters.unsupervised.attribute.NumericToNominal* sobre os ano de nascimento, ano de desmame, ano de abate, mês de desmame e mês de nascimento, de modo que os dados foram convertidos do formato numérico para nominal, a fim de evitar que na forma numérica os dados constituíssem pesos quem afetassem os modelos descobertos.

A segunda etapa consistiu no processamento do conjunto de dados, que ocorreu com a tarefa de regressão linear, através do algoritmo *weka.classifiers.functions.LinearRegression* [12]. A regressão linear é utilizada basicamente com duas finalidades, prever o valor de y a partir do valor de x e estimar quanto x influencia ou modifica y. Adotou-se este algoritmo pois ele gera um modelo de comportamento. Também calcula o valor da correlação entre os atributos preditores e o atributo alvo. Além disso, só usa as colunas que contribuem estatisticamente para a precisão, descartando e ignorando as colunas que não ajudam a criar um bom modelo. Foram executados experimentos para cada uma das 4 variáveis alvo. Tabela 2 apresenta os atributos selecionados para cada um dos experimentos. A terceira etapa consistiu na análise dos resultados obtidos.

Tabela 2: Variáveis utilizadas nos experimentos

Variáveis	abate_idade	gmd	abate_peso	bonificacao
nascimento_ano				
desmame_ano				
abate_ano				
desmame_peso				
desmame_idade				
desmame_mes				
nascimento_mes				
abate_idade	Alvo	Removido	Removido	Removido
gmd	Removido	Alvo	Removido	Removido
abate_peso	Removido	Removido	Alvo	Removido
bonificacao	Removido	Removido	Removido	Alvo

4. ANÁLISE DOS RESULTADOS

Na Figura 1 observa-se os modelos descobertos para os quatro experimentos, peso de abate na fazenda, ganho médio diário, bonificação e idade de abate. O modelo é o resultado gerado pela tarefa de regressão linear. Nele, os atributos relevantes têm pesos atribuídos, de forma a comporem uma fórmula matemática para o cálculo do atributo alvo.

Para o peso de abate na fazenda o experimento descobriu um modelo onde os atributos utilizados foram ano de abate,

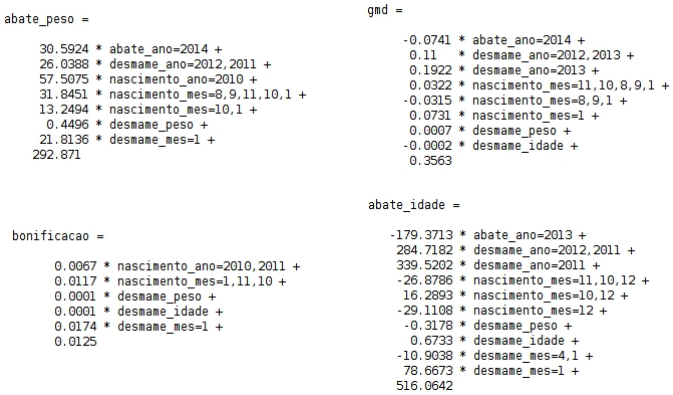

Figura 1: Modelos Descobertos

ano de desmame, ano de nascimento, mês de nascimento, peso de desmame e mês de desmane, sendo o atributo ano de nascimento = 2010 e ano de mês de nascimento= 10 e 1 os mais relevante pois apresentam dois coeficientes no modelo, dando um maior peso a estes atributos, outro fato a se ressaltar está na circunstância de os mês de nascimento = 12, ano de desmame = 2013, ano de abate = 2013 não serem utilizados no modelo, o mesmo ocorre com o atributo mês de desmame = 4 e 5, sendo utilizado apenas o mês de desmame = 1. Idade de desmame sequer foi utilizado.

Para o ganho médio diário o modelo descoberto apresenta mês de nascimento = 1 como atributo de maior relevância, pois é gerado três coeficientes para este atributo, sendo que o mês de nascimento = 12 não foi utilizado no modelo. Nota-se também que o ano de desmame = 2013 apresenta dois coeficientes enquanto mês de desmame não foi utilizado no modelo.

O modelo descoberto para a bonificação utilizou o ano de nascimento = 2010 e 2011, mês de nascimento = 1, 11 ou 10, peso de desmame, idade de desmame e mês de desmame = 1, desconsiderando as outras informações. Nota-se que o modelo descoberto para a bonificação foi o que utilizou menos informações fornecidas pelas instâncias do experimento. No contraponto, o modelo descoberto que utilizou mais informações foi da idade de abate, sendo que para o ano de desmame = 2011 e mês de desmame = 1 foram gerados dois coeficientes, para o mês de nascimento = 12, foram gerados três coeficientes, o que acaba configurando este atributo como o de maior relevância para o modelo.

Com referência aos modelos descobertos, os atributos não utilizados podem ter influenciado o resultado dos experimentos, podendo ,os modelos, terem seus desempenhos afetados por essas exclusões de informações.

Além dos modelos, cada experimento apresentou o relatório de valores reais para cada instância, o valor previsto e a diferença entre eles(erro na previsão). Analisando os erros de cada instância com o valor real, observa-se que os erros para o ganho médio diário e bonificação foram baixos, as maiores diferenças entre o valor real e o previsto ocorreram para o peso de abate. A idade de fazenda apresentou um desenho razoável por não apresentar uma variação elevada do erro.

A Tabela 3 apresenta os valores para comparação dos coeficientes de correlação e erros médios absolutos, calculados

pelo algoritmo de regressão linear.

Tabela 3: Correlação e erro médio absoluto

Classes	Correlation coefficient	Mean absolute error
Bonificação	0.3715	0.015
GMD	0.8761	0.0313
Idade de Abate	0.9746	23.5903
Peso de Abate	0.5036	29.1061

A correlação é uma medida estatística que indica a força e a direção da relação entre variáveis numéricas [1]. Ou seja, a correlação é um índice que indica o quanto duas variáveis estão relacionadas, sendo os valores retornados sempre dentro do intervalo de -1 e 1. Quanto mais próximas de -1 e 1, maior será a correlação entre as variáveis, e da mesma forma, quanto mais próxima de 0, mais fraca ela é.

O indicador de direção é dado pelo sinal da correlação, uma correlação positiva indica que enquanto uma variável cresce, a outra, correlacionada, também cresce, já na correlação negativa, enquanto uma variável cresce a outra diminui [1].

Analisando a Tabela 3 nota-se que, para o ganho médio diário e a idade de abate o índice de correlação foi alto entre as variáveis preditoras e as variáveis alvo, indicando que os modelos descobertos obtiveram uma boa métrica de qualidade, pois todas as variáveis utilizadas estão fortemente correlacionadas. Observa-se também que a direção do coeficiente de correlação é positivo para todos os experimentos.

Ainda foi possível analisar o valor de R^2, que é o coeficiente de determinação. Ele fornece uma informação auxiliar ao resultado da análise de variância da regressão, como maneira de se verificar se o modelo proposto é adequado ou não para descrever o fenômeno estudado. O valor de R^2 varia no intervalo de 0 a 1. Valores próximos de 1 indicam que o modelo proposto é adequado para descrever o fenômeno. Tabela 4 apresenta os valores do R^2 para os experimentos realizados.

Tabela 4: Valores dos coeficientes de determinação

Classes	Coeficiente de determinação - R^2
Bonificação	0.138
GMD	0.7675
Idade de Abate	0.9497
Peso de Abate	0.2536

Analisando os valores de R^2 encontrados, observa-se que o ganho médio diário e a idade de abate, apresentaram bons coeficientes de determinação. Os outros R^2 indicam que os modelos descobertos não são adequados para descrever as variáveis zootécnicas de qualidade estudadas.

5. CONCLUSÃO

Pode-se concluir que os resultados foram parcialmente alcançados, pois com as tarefas de regressão configuradas conforme descritas na metodologia, mostraram que as variáveis de cria usadas possuem boa correlação e R^2 para a idade de abate e ganho médio diário de peso, estes foram os modelos que utilizaram um maior número de atributos preditores(variáveis de cria). Os modelos descobertos para a bonificação e ganho médio diário apresentaram erros baixos. Para o peso de fazenda e idade de abate, os modelos apresentaram diferenças maiores entre o valor real e o valor previsto. A baixa correlação e R^2, para peso de abate e bonificação pode significar que apenas as variáveis de cria usadas não sejam o suficientes para explicar os fenômenos.

Trabalhos futuros envolvem à adoção de novos indicadores, como por exemplo o peso de nascimento, tipo de alimentação da mãe do bovino enquanto este ainda mama, entre outros, pode-se também empregar outras técnicas de mineração de dadoscomo o algoritmo M5P e redes neurais, com treinamento e configuração das camadas ocultas. Também se sugere a expansão do banco de dados, através de parcerias com outros produtores rurais, e do estudo para consideração de outras raças bovinas. Apresentando ao produtores os resultados obtidos e demonstrando que é possível aumentar seu rendimento com técnicas adequadas.

6. REFERÊNCIAS

[1] F. Amaral. *Aprenda Mineração de Dados - Teoria e Prática*. Rio de Janeiro: Alta Books, 1th edition, 2016.

[2] Â. M. J. Corrêa and H. Sferra. Conceitos e aplicações de data mining. *Revista de ciência & tecnologia*, 11:19–34, 2003.

[3] C. L. Costa. *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, 2016.

[4] K. Euclides Filho. Produção de bovinos de corte e o trinômio genótipo-ambiente-mercado. *Embrapa Gado de Corte-Documentos (INFOTECA-E)*, 2000.

[5] U. M. Fayaad, G. P. Shapiro, and P. Smyth. From data mining to knowledge discovery: An overview. 1996.

[6] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[7] F. Meirelles. *Modelo computacional de um rebanho bovino de corte virtual utilizando simulação Monte Carlo e redes neurais artificiais*. PhD thesis, Universidade de São Paulo, 2005.

[8] R. Oiagen and J. Barcellos. Gerenciamento e custo de produção. *MOURA, JA et al. Programa de atualização em medicina veterinária. Porto Alegre: ARTMED*, pages 51–88, 2008.

[9] R. P. Oiagen. *Avaliação da competitividade em sistemas de produção de bovinocultura de corte nas regiões sul e norte do Brasil*. Tese de doutorado em zootecnia, Universidade Federal do Rio Grande do Sul, 2010.

[10] P.-N. Tan, M. Steinbach, and V. Kumar. Association analysis: basic concepts and algorithms. *Introduction to Data mining*, pages 327–414, 2005.

[11] R. Ventura, M. Silva, T. Medeiros, N. Dionello, F. Madalena, A. Fridrich, B. Valente, G. Santos, L. Freitas, R. Wenceslau, et al. Use of artificial neural networks in breeding values prediction for weight at 205 days in tabapuã beef cattle. *Arquivo Brasileiro de Medicina Veterinária e Zootecnia*, 64(2):411–418, 2012.

[12] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.