

Evaluierung des Entscheidungsbaumalgorithmus für die Remontierung von Sauen in Ferkelerzeugerbetrieben

KATRIN KIRCHNER, KIEL
KARL-HEINZ TÖLLE, KIEL
JOACHIM KRIETER, KIEL

Abstract

The decision tree technique is an effective instrument for making large datasets accessible and different sow herd data comparable. This technique could be used to improve the farm management using detected differences and weak points. In datasets from two Northern German sow herds the decision tree technique was used to generate such trees and classify exemplarily the binary farmer decision regarding replacing a sow with a gilt or not. The C4.5-decision tree algorithm generated decision trees which showed different sizes between 15 and 55 nodes. In relation to the sow herd performance the threshold values of the attributes at the branches varied between the trees. The sensitivity, the kappa value and the error rate were the reasonable evaluation parameters for estimating the algorithm performance in classifying present skewed datasets. Through deleting indefinite replacement decisions in the dataset the evaluation parameters were improved after classifying the reduced datasets. Henceforth, the sensitivity reached values of between 69.4 and 87.5% and the error rate (10.1–15.0%) decreased.

1 Einleitung

Die Ferkelerzeugung ist in den letzten Jahren durch ein deutliches Wachstum der Sauenanzahl je Betrieb gekennzeichnet. Diese Entwicklung führt zu einer steigenden Nachfrage nach Management-Informationssystemen, die eine Unterstützung des Herdenmanagement gewährleisten und die Entscheidungsfindung des Landwirtes unterstützen. Ziel dieses Projektes ist es, Daten der Ferkelerzeugung mit Methoden des Maschinellen Lernens auszuwerten, um Schwachstellen im Produktionsablauf zu erkennen. Das angewendete Entscheidungsbaum-Verfahren soll Entscheidungsregeln und Zusammenhänge am Beispiel der Sauenersatzentscheidung abbilden.

2 Material und Methoden

Die verfügbaren Daten stammen von zwei Ferkelerzeugerbetrieben. Nach Durchführung einer Plausibilitätskontrolle des Datenmaterials, liegen für den Betrieb A 14.897 Instanzen (Beobachtungen) in dem Zeitraum von 1994 bis 2001 vor. Betrieb B verfügt über 21.818 Datensätze von 1984 bis 1999. Tabelle 1 zeigt die Mittelwerte und Standardabweichungen einiger Reproduktionsparameter der Sauen beider Betriebe.

Das Entscheidungsbaum-Verfahren des Maschinellen Lernens klassifiziert unbekannte Daten aufgrund vorhandener Muster und stellt diese Entscheidungsregeln in Form von grafischen Bäumen dar. In Neuseeland liegen im Agrarbereich erste Ansätze vor, diese Verfahren zur Informationsunterstützung des Managements einzusetzen (MCQUEEN et al., 1995). Zur Analyse der Daten wurde der C4.5-Algorithmus des Open Source-Programmpaketes WEKA der Universität Waikato (Neuseeland) angewendet (WITTEN und FRANK, 2000).

Zum Aufbau des Baumes und der Klassifikation wurden folgende Reproduktionsmerkmale genutzt: Wurfnummer, Absetz-Konzeptions-Intervall in Tagen, Umrauschhäufigkeit, gesamt geborene, tot geborene, lebend geborene und abgesetzte Ferkel je Wurf.

Tabelle 1: Mittelwerte (\bar{x}) und Standardabweichung (s) wichtiger Reproduktionsmerkmale von Betrieb A und Betrieb B.

Merkmale	Betrieb A <i>n</i> = 14.897		Betrieb B <i>n</i> = 21.818	
	\bar{x}	s	\bar{x}	s
Wurfnummer	3,7	2,4	3,6	2,2
Anzahl gesamt geborener Ferkel/Wurf	11,6	3,3	10,5	2,4
Anzahl tot geborener Ferkel/Wurf	1,1	1,6	0,5	0,7
Anzahl lebend geborener Ferkel/Wurf	10,5	3,1	10,0	2,3
Anzahl abgesetzter Ferkel/Wurf	9,5	1,5	9,1	2,2
Absetz-Konzeptionsintervall (Tage)	10,2	15,2	13,5	19,7
Umrauschquote (%)	10,9		13,9	

Neben dem Originaldatensatz wurden im zweiten Teil der Analyse die Datensätze von Betrieb A und B um die Abgangsgründe reduziert, die sich nicht in den Leistungen der Sauen widerspiegeln bzw. nicht für jede Abferkelung vorliegen und dadurch auch nicht durch den angewendeten Algorithmus aufgefunden werden können. Hierzu zählen beispielsweise aggressive oder streßanfällige Sauen sowie Tiere, die aufgrund eines schlechten Fundaments gemerzt wurden. Für die Analyse dieser reduzierten Daten (A' und B') wurde ein Genauigkeitszuwachs erwartet.

Die Reihenfolge der Attribute innerhalb des Baumes sowie die Schwellenwerte der Äste wurden mit Hilfe des maximalen Informationsgewinnverhältnisses berechnet. Die Validierung der Klassifizierung der binären Entscheidung, ob eine Sau gemerzt wird oder im Bestand bleibt, erfolgte durch die stratifizierte 10-fache Kreuzvalidierung, auf deren Basis die folgenden Evaluierungsparameter berechnet wurden:

Die allgemeine Klassifikationsgenauigkeit gibt den Anteil der korrekt klassifizierten Tiere im Verhältnis zu allen Instanzen an. Sie macht jedoch noch keine Aussage darüber, wie gut die einzelnen Ausprägungen des Zielattributes klassifiziert werden. Die Sensitivität beschreibt den Anteil der korrekt klassifizierten, gemerzten Sauen im Verhältnis zu allen gemerzten Sauen. Die Spezifität gibt einen Überblick über die Qualität der Klassifikation der im Bestand gebliebenen Tiere. Die Kappa Statistik beschreibt den Anteil der Übereinstimmung der Klassifikation der gemerzten und im Bestand gehaltenen Sauen. Die Fehlerrate trifft eine Aussage über den Anteil der Sauen, die im Bestand geblieben sind, jedoch fälschlicherweise in die Klasse der gemerzten Tiere eingeordnet wurden, im Verhältnis zu allen Sauen, die in die Klasse der abgegangenen Sauen klassifiziert wurden.

3 Ergebnisse

In Tabelle 2 sind die unterschiedlichen Evaluierungsparameter der beiden Original-Sauendatensätze A und B dargestellt. Die Generierung der Entscheidungsbäume erfolgte mit einer unterschiedlichen Mindestanzahl Instanzen je Klasse, die in WEKA auf 20, 100 und 200 spezifiziert wurde (Beispiel: A₂₀ = generiert mit mindestens 20 Instanzen je Klasse). Sämtliche Evaluierungsparameter wurden durch eine stratifizierte zehnfache Kreuzvalidierung berechnet.

Unabhängig von der Mindestanzahl Instanzen je Klasse zeigt der Betrieb B bessere Klassifikationswerte als Betrieb A. Die Sensitivität erreicht für Betrieb B mit einer Mindestanzahl von 20 Instanzen je Klasse den höchsten Wert (47,3 %). Mit 200 Instanzen je Klasse beträgt die Sensitivität 46,9 %. Der Betrieb A₂₀ zeigt eine Sensitivität von 41,4 %, welche bei 200 Instanzen je Klasse auf 39,2 % fällt. Die Kappa Statistik läßt einen ähnlichen Trend wie die Sensitivität erkennen. Die Fehlerrate liegt für Betrieb B zwischen 14,2 % und 17,0 %. Für den Betrieb A erreicht die Fehlerrate höhere Werte, die bei A₂₀ 16,8 %, bei A₁₀₀ 18,9 % und bei A₂₀₀

19,2 % beträgt. Die Anzahl der Blätter und Knoten sinkt mit steigender Mindestanzahl Instanzen je Klasse und ist bei einem Minimum von 200 Instanzen je Klasse bei beiden Datensätzen mit 8 Blättern und 15 Knoten identisch. Die Klassifikationsgenauigkeit und die Spezifität schwanken zwischen den Einstellungen kaum.

Tabelle 2: Klassifikationsparameter der Betriebe A und B klassifiziert mit variierender Mindestanzahl Instanzen je Klasse.

Daten- satz ¹⁾	Klassifikations- genauigkeit	Sensitivität	Spezifität	Fehlerrate	Kappa Statistik	Anzahl Blätter	Anzahl Knoten
	%	%	%	%	%		
A ₂₀	85,4	41,4	97,7	16,8	47,7	26	51
B ₂₀	87,0	47,3	97,9	14,2	53,9	28	55
A ₁₀₀	84,9	40,3	97,4	18,9	46,0	14	27
B ₁₀₀	86,8	46,8	97,8	15,3	53,2	12	23
A ₂₀₀	84,7	39,2	97,4	19,2	44,9	8	15
B ₂₀₀	86,5	46,9	97,4	17,0	52,5	8	15

¹⁾ A = Datensatz A ($n = 14.897$)

B = Datensatz B ($n = 21.818$)

20, 100, 200 = Minimum von 20, 100 oder 200 Instanzen je Klasse

Die reduzierten Datensätze (A' und B') weisen erheblich höhere Klassifikationsparameter auf, wobei auch hier der Betrieb B' bessere Werte erzielt als A' (Tabelle 3). Für Datensatz B'₂₀ und B'₁₀₀ beträgt die Sensitivität 87,5 %, die auf 86,1 % für Betrieb B'₂₀₀ abfällt. Der Betrieb A'₂₀ erreicht eine Sensitivität von 74,8 %, die leicht absinkt und für A'₁₀₀ bei 74,2 % und für A'₂₀₀ bei 69,4 % liegt. Die Fehlerraten sind für beide Betriebe unabhängig von der Mindestanzahl Instanzen je Klasse sehr niedrig (10,1 bis 15,0 %). Die Klassifikationsgenauigkeiten sowie die Spezifitäten erreichen sehr hohe Werte für die geprüften Varianten. Die Anzahl der Blätter und Knoten ist um so höher, je kleiner die Mindestanzahl Instanzen je Klasse gewählt wurde. Insgesamt zeigt A' jedoch kleinere Bäume als B'.

Tabelle 3: Klassifikationsparameter der reduzierten Datensätze A' und B', klassifiziert mit variierender Mindestanzahl Instanzen je Klasse.

Daten- satz ¹⁾	Klassifikations- genauigkeit	Sensitivität	Spezifität	Fehlerrate	Kappa Statistik	Anzahl Blätter	Anzahl Knoten
	%	%	%	%	%		
A' ₂₀	93,3	74,8	97,9	10,1	77,6	16	31
B' ₂₀	95,0	87,5	96,9	12,3	84,4	23	45
A' ₁₀₀	92,2	74,2	96,7	15,0	74,5	7	13
B' ₁₀₀	94,4	87,5	96,1	14,9	82,8	9	17
A' ₂₀₀	92,2	69,4	97,9	10,8	73,5	3	5
B' ₂₀₀	94,4	86,1	96,6	13,6	82,8	5	9

¹⁾ A' = Datensatz A' ($n = 7.057$)

B' = Datensatz B' ($n = 12.149$)

20, 100, 200 = Minimum von 20, 100 oder 200 Instanzen je Klasse

4 Diskussion

In den untersuchten Daten erzielte der Betrieb B höhere Sensitivitäten und niedrigere Fehler-raten unabhängig von der Mindestanzahl Instanzen je Klasse als Betrieb A. Dieses kann durch eine einheitlichere Strategie hinsichtlich der Entscheidung, ob Sauen gemerzt werden oder nicht, erklärt werden. Allerdings sind die Unterschiede in den Evaluierungsparametern nur gering. Die generierten Bäume beschreiben jeweils sinnvolle Entscheidungsregeln (nicht abge-bildet).

Das Auffinden von Regelmäßigkeiten und das Abbilden von logischen Entscheidungen ver-besserte sich, wenn die Daten von Betrieb A und B um die Beobachtungen bereinigt wurden, in denen Sauen aus für den Entscheidungsbaum-Algorithmus nicht eindeutigen Gründen ab-gegangen sind, wie z. B. aggressive Sauen. Der Algorithmus kann diese Gründe nicht zuord-nen, da für diese Merkmale keine Informationen in vorangegangenen Würfen vorhanden sind. Die Klassifikationsparameter waren für die bereinigten Datensätze A' und B' erwartungsge-mäß höher, weil für den Algorithmus nur eindeutige Abgangsgründe in den reduzierten Da-tensätzen vorhanden waren. Die Bäume fielen ebenfalls kleiner aus, da eindeutigere Entschei-dungsregeln und kleinere Datensätze für die Klassifikation zur Verfügung standen.

Durch eine geringe Mindestanzahl von Instanzen je Klasse wurde der Entscheidungsbaum stark verfeinert und die Klassifikationsgenauigkeit stieg an. Es muß jedoch berücksichtigt werden, daß die Auswahl einer geringen Mindestanzahl Instanzen je Klasse zur einer starken Verzweigung führt und die Bäume sehr komplex und dadurch schwerer zu interpretieren sind. Durch eine zu starke Verfeinerung des Baumes kann es an den Astenden zu nicht erklärbaren Aussagen kommen, weil die Klassenbesetzung zu gering ist und keine allgemeingültigen Zu-sammenhänge berücksichtigt werden. Die optimale Baumgröße, das heißt, das richtige Ver-hältnis von Mindestanzahl Instanzen je Klasse zu der Gesamtanzahl der Instanzen des Daten-satzes ist entscheidend, damit ein übersichtlicher Baum mit einer möglichst hohen Klassifika-tionsgenauigkeit erzielt wird. Zu ähnliche Erkenntnissen kamen KIRCHNER et al. (2003) durch Analysen von simuliertem Sauenherden-Datenmaterial. Die Optimierung ist vor allem für die Betriebsleiter oder die Beratung wichtig, weil nur dann die wesentlichen Schwachstellen und/oder Entscheidungsregeln in den Bäumen richtig interpretiert werden und entsprechende Maßnahmen für Managementveränderungen gegeben werden können.

5 Schlußfolgerung

Die grafische Darstellung des Entscheidungsbaums ermöglicht es, Entscheidungsregeln des Betriebsleiters abzubilden, die später genutzt werden können, um das Management in der Fer-kelproduktion zu überprüfen. Durch den horizontalen und vertikalen Vergleich von Betrieben, die in Produktionsstruktur und Management vergleichbar sind, können Schwachpunkte in den Entscheidungsregeln der Landwirte aufgedeckt werden. Diese Erkenntnisse ermöglichen es, daß die Beratung nachhaltig gesteigert werden kann. Vorstellbar ist eine Integration des Ent-scheidungsbaum-Verfahrens in ein computergestütztes Management-Informationssystem.

6 Literatur

- KIRCHNER, K., TÖLLE, K.-H., KRIETTER, J. (2003): The analysis of simulated sow herd datasets using decision tree technique. Submitted for Computers and Electronics in Agriculture.
- QUINLAN, J. R. (1993): C4.5: Programs for machine learning, Morgan Kaufman, San Fran-cisco, USA
- WEKA 3-2-3, (2000): <http://www.cs.waikato.ac.nz/ml/weka/>
- WITTEN, I. H., FRANK, E. (2000): Data Mining—practical machine learning tools and tech-niques with Java implementation. Morgan Kaufmann.