Analyse von simulierten Daten der Ferkelproduktion mit Hilfe von Entscheidungsbaum-Verfahren

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

Abstract

The aim of the study was to investigate the ability of machine learning techniques for analysing practical problems in pig production herds. Farmers have to decide routinely whether to cull a sow and replace it with a gilt. The decision tree method was used to analyse the available information about the farmers' decision considering sow replacements. Therefore, a Monte Carlo simulation was conducted that modelled a pig production herd. Three different production levels were simulated with a low, medium, and high production level. The calculation included for each sow and each production cycle a set of fertility parameters. The C4.5-algorithm (WEKA) was applied to the simulated data sets. It classified the attributes which were relevant (10-fold cross-validation) for the decision about a sows' replacement and produced graphical trees. The classification of the simulated data sets was done with a confidence interval of 25% for pruning trees. The sensitivity was 51.5% to 68.6%, the specificity reached a value from 97.7 to 98.1%, and the error rate was between 9.1 and 17.0%.

1 Einleitung

Mit Data Mining-Methoden können durch Anwendung von überwachten Lernverfahren, insbesondere Entscheidungsbaum-Verfahren aus dem Bereich des Maschinellen Lernens, Entscheidungsregeln und Zusammenhänge in großen Datenmengen ermittelt werden. Data Mining ist die Anwendung effizienter Algorithmen, die in großen Datenmengen enthaltene Muster herausfiltern (FAYYARD, et al. 1996).

Das Ziel dieser Studie ist es, mit Hilfe dieser Methoden komplexe Fragestellungen aus der Ferkelproduktion zu analysieren. Am Beispiel der Nutzungsdauer von Sauen wird untersucht, inwieweit mit dem Entscheidungsbaum-Verfahren die Entscheidungsfindungen von Landwirten abgebildet werden können. Um die Effizienz des Entscheidungsbaum-Verfahrens zu analysieren, werden mit einer Monte Carlo Simulation verschiedene praxisnahe Datensätze von Ferkelerzeugerbetrieben mit drei variierenden Leistungsniveaus (niedrig, mittel, hoch) generiert.

2 Material und Methoden

Das simulierte Datenmaterial besteht aus präzisen Einzeltierinformationen mit folgenden Parametern: individuelle Wurfnummer, Absetz-Brunst-Intervall, Anzahl des Umrauschens, gesamt geborene, tot geborene, lebend geborene, abgesetzte Ferkel je Wurf. Nach jeder Abferkelung findet eine Selektion der Sauen nach folgenden Kriterien statt: (1) Fruchtbarkeitsprobleme, (2) klinische Probleme, (3) geringe Leistung und (4) zu hohes Alter. Die Datensätze sind jeweils mit einer Sauenanzahl von 500 Tieren, eine Anzahl von zehn aufeinanderfolgenden Produktionszyklen und einer daraus resultierenden Gesamtzahl von 5 000 Instanzen simuliert.

Zur Analyse der simulierten Daten wird das Entscheidungsbaum-Verfahren angewendet. Der Algorithmus „lernt" anhand vorhandener, bestimmten Klassen zugeordneter Trainingsobjekte, und ordnet neue unbekannte Instanzen aufgrund ihrer Werte den entsprechenden Klassen zu.

Ein Entscheidungsbaum besteht aus einer Wurzel, den nachfolgenden inneren Knoten, welche die Attribute vertreten und den abschließenden Blättern des Baumes, die die Klassen darstellen. Die Äste repräsentieren einen Test auf das jeweilige Attribut des Vaterknotens.
Die Analyse der simulierten Daten erfolgte mit dem Entscheidunsbaum-Verfahren weka.classfiers.j48.J48 des Open Source-Programmpaketes WEKA 3-0 (2000, University of Waikato, Environment for Knowledge Anlaysis). Diesem Klassifizierungsverfahren liegt der C4.5-Algorithmus zugrunde, welcher von QUINLAN (1993) entwickelt wurde. Der C4.5-Algorithmus generiert die Entscheidungsbäume nach dem Top-Down-Verfahren, also von der Wurzel ausgehend. Eine neue Instanz wird klassifiziert, indem der Baum von der Wurzel bis zum Blatt durchlaufen wird. Die Anordnung der Attribute innerhalb des Baumes erfolgt über die Berechnung des gain ratios für jedes Attribut, so daß die Attribute in Abhängigkeit ihres Informationsgehaltes im Baum rangiert werden.
Die drei simulierten Datensätze enthalten die wichtigsten Fruchtbarkeitsparameter von Sauen, sowie die Entscheidung, ob die Sau gemerzt und durch eine Jungsau ersetzt wird oder im Bestand bleibt. Das jeweilige Ergebnis des Zielattributes, welches die Blätter des Baumes bildet, heißt „Abgang“ oder „Bestand“. Der C4.5-Algorithmus bildet also in dieser Untersuchung einen binären Entscheidungsbaum. Die Klassifizierung der Attribute erfolgt mit einem Konfidenzintervall von 25 %, welches zu einem fehlerreduzierten Beschneiden der Bäume führt.
Die Prüfung der Güte der Klassifizierung erfolgt über die Berechnung der stratifizierten zehnfachen Kreuzvalidierung (WEISS AND KULIKOWSKI, 1991), in welcher die Daten in zehn Teilmengen gleicher Größe eingeteilt werden. Pro Klassifizierungsdurchlauf wird eine Menge zum Testen und der Rest für das Training verwendet. Dieses Verfahren wird zehnmal durchlaufen und die Fehlerraten anschließend gemittelt. Durch die Stratifikation erfolgt eine Verringerung der Varianz. Die Genauigkeit des Klassifizierungsalgorithmus wird durch folgende Evaluierungsparameter dargestellt: Die Sensitivität beschreibt den Anteil der korrekt klassifizierten, gemerzten Sauen im Verhältnis zu allen gemerzten Sauen. Die Spezifität ist ein Merkmal für die Einschätzung des Klassifizierungsvermögens des Algorithmus der im Bestand gebliebenen Sauen. Sie gibt den Anteil der korrekt klassifizierten, im Bestand gebliebenen Sauen im Verhältnis zu allen im Bestand gebliebenen Sauen an. 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 sich in der Klasse der gemerzten Sauen befinden.

Sensitivität = wahr positiv / (wahr positiv + falsch negativ) * 100

Spezifität = wahr negativ / (wahr negativ + falsch positiv) * 100

Fehlerrate = falsch positiv / (falsch positiv + wahr positiv) * 100

3 Ergebnisse

Der C4.5-Algorithmus klassifiziert die drei divergierenden Datensätze in unterschiedlicher Weise in Abhängigkeit von dem jeweiligen Leistungsniveau der Herde sowie auch der Mindestanzahl Instanzen je Klasse. Deshalb differieren sowohl die Größe der Entscheidungsbäume als auch die Rangierung der Attribute innerhalb der Bäume. Bei einer Mindestanzahl von beispielsweise 100 Instanzen je Klasse entstehen die in Tabelle 1 dargestellten Schätzfehler. Bei den Betrieben mit einem schlechten und einem mittleren Leistungsniveau ist die gleiche Anzahl von Knoten und Blättern zu erkennen, jedoch differieren die Teilungswerte der Äste. Der generierte Baum des guten Betriebes ist beispielhaft in Abbildung 1 dargestellt. Er besteht aus 13 Knoten und sieben Blättern. Jeder

Weg von der Wurzel bis zum einzelnen Blatt repräsentiert eine Entscheidung eines Landwirtes.

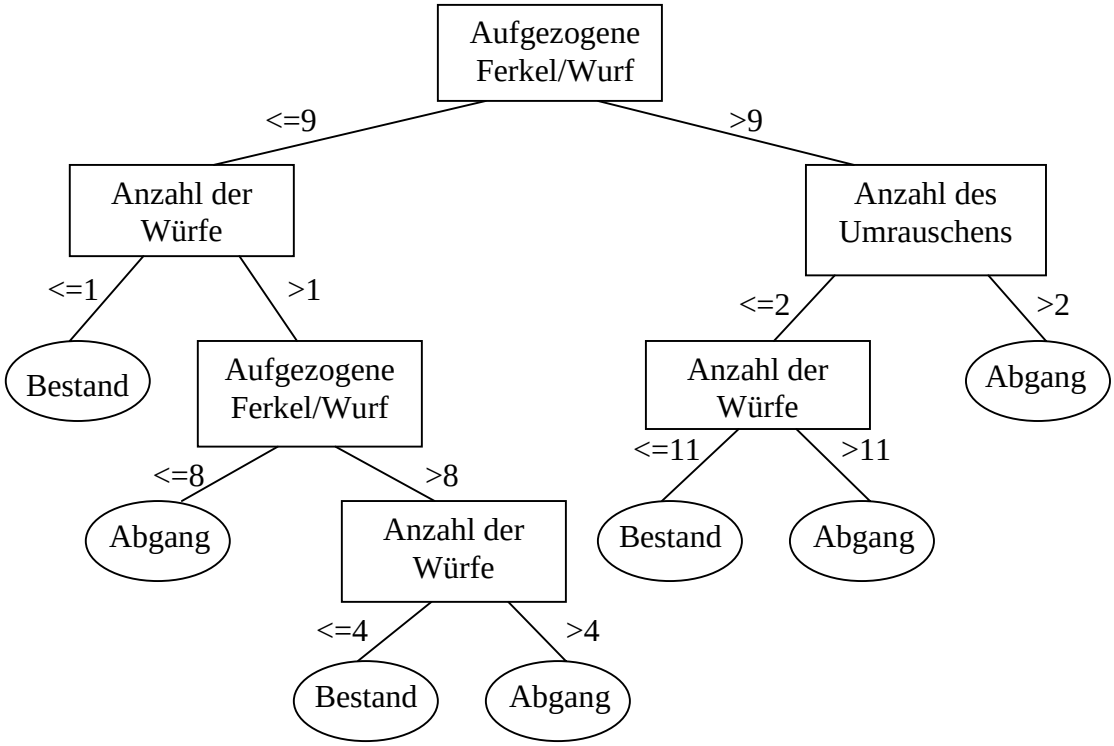

Abbildung 1: Entscheidungsbaum eines simulierten Ferkelerzeugerbetriebes mit hohem Leistungsniveau.

Die Evaluierungsparameter, die die Güte des Klassifzierungsalgorithmus charakterisieren, schwanken je nach Leistungsstand der Sauenherde. Bei einer Mindestanzahl von 100 Instanzen je Klasse wird die beste Klassifizierung für den Betrieb mit dem geringsten Leistungsniveau erzielt. Wie in Tabelle 1 zu sehen ist, liegt die Sensitivität bei 68,6 %, die Spezifität beträgt 98,1 % und die Fehlerrate ist mit nur 9,1 % sehr gering. Der mittlere und der gute Betrieb erreichen vergleichbare Spezifitäten und Fehlerraten, jedoch unterscheiden sich die Bäume hinsichtlich ihrer Größe. Die Sensitivität ist jedoch bei dem mittleren Betrieb mit 55,1 % besser als bei dem guten Betrieb (51,5 %).

Bei einer Veränderung der Mindestanzahl Instanzen je Klasse kommt es zu einer Veränderung der Schätzparameter. Die jeweils besten Schätzparameter werden mit einer Mindestanzahl von zwei Instanzen je Klasse erzielt (Abbildung 2). Durch die Zunahme der Mindestanzahl Instanzen je Klasse sinkt die Sensitivität ab, wobei ab 100 Instanzen eine starke Abnahme erfolgt. Bei 200 Instanzen je Klasse nimmt die Sensitivität wieder leicht zu, jedoch bei 300 Instanzen wird ein sehr deutlicher Rückgang sichtbar. Die Spezifität verringert sich insgesamt langsamer. Erst bei 200 Instanzen ist ein Abfall sichtbar, der wieder in eine Zunahme übergeht. Die Fehlerrate nimmt langsam zu und steigt ab 200 Instanzen deutlich an.

Tabelle 1: Darstellung der Evaluierungsparameter bei einer Mindestanzahl Instanzen je Klasse von 100.

Sauenanzahl = 500, Produktionszyklen = 10, n = 5 000

Leistungs-niveau	Sensitivität (in %)	Spezifität (in %)	Fehlerrate (in %)	Anzahl der Blätter	Anzahl der Knoten
Niedrig	68,6	98,1	9,1	4	7
Mittel	55,1	97,7	17,0	4	7
Hoch	51,5	97,8	16,0	7	13

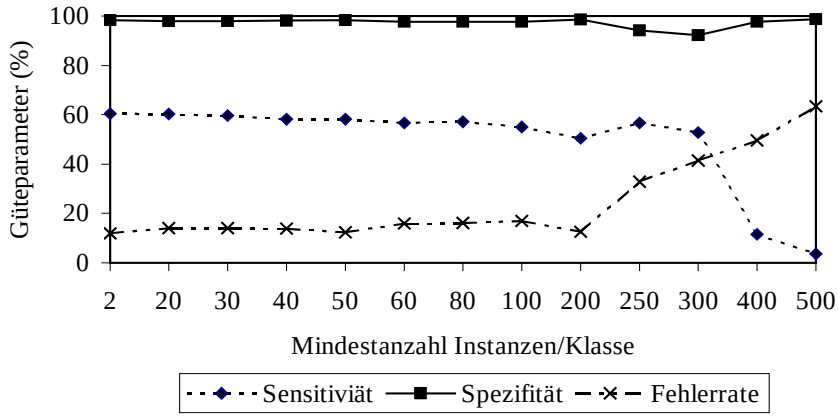

Abbildung 2: Verlauf der Klassifikationsgüteparameter in Abhängigkeit von der Mindestanzahl Instanzen je Klasse am Beispiel eines simulierten Datensatzes eines mittleren Leistungsniveaus.

4 Diskussion

Zur Einschätzung der Klassifikationsgüte der gemerzten Sauen müssen die Sensitivität und die Fehlerrate betrachtet werden. Da der Anteil der gemerzten Sauen geringer ist als der im Bestand gebliebenen Tiere, sind diese beiden Schätzparameter von entscheidender Aussage über die Fähigkeit des C4.5-Algorithmus. Der Betrieb mit der geringsten Leistung wird am besten klassifiziert, weil aufgrund seiner geringen Leistung der Hauptteil der Sauen aus eindeutigen Gründen, wie beispielsweise geringe Leistung, abgeht. Durch eine geringe Mindestanzahl von Instanzen je Klasse wird der Entscheidungsbaum stark verfeinert und die Klassifikationsgenauigkeit steigt an. Es muß aber auch berücksichtigt werden, daß die Auswahl einer geringen Mindestanzahl Instanzen je Klasse zur 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 Zusammenhänge berücksichtigt werden. Die optimale Baumgröße, das heißt, das richtige Verhältnis von Mindestanzahl Instanzen je Klasse, und die Gesamtanzahl der Instanzen des Datensatzes ist entscheidend, damit ein übersichtlicher Baum mit einer möglichst großen Klassifikationsgenauigkeit erzielt wird.

5 Schlußfolgerung

Durch Anwendung des C4.5-Algorithmus auf simulierte Daten der Ferkelproduktion kann die Entscheidung über den Abgangszeitpunkt von Sauen gut interpretiert werden. Die grafische Darstellung des Entscheidungsbaums ermöglicht es, Entscheidungsregeln des Landwirtes abzubilden, die später genutzt werden können, um das Management in der Ferkelproduktion zu überprüfen.

6 Literatur

FAYYARD, U. M.; PIATESTSKY-SHAPIRO, G.; SMYTH, P. (1996): From Data Mining in Knowledge Discovery. In: Fayyard, U. M.; Piatestsky-Shapiro, G.; Smyth, P.; Uthurusamy, R. (eds.). Advantage in Knowledge Discovery and Data Mining. AAAI/MIT Press, Menlo Park, 1-24

QUINLAN, J. R. (1993): C4.5: Programs for machine learning, Morgan Kaufman, San Francisco, USA

WEKA 3-0, 2000: http://www.cs.waikato.ac.nz/ml/weka/

WEISS, S. M.; KULIKOWSKI, C. A. (1991): Computer Systems that Learn, Morgan Kaufmann, San Mateo, USA