Artificial neural networks for genome-enabled prediction in cattle: Potential and limitations

vorgelegt von: M.Sc. agr. Anita Ehret
Institut für Tierzucht und Tierhaltung der Christian-Albrechts-Universität zu Kiel
Erster Berichterstatter: Prof. Dr. Georg Thaller

Das Ziel der vorliegenden Arbeit war es, unter Berücksichtigung komplexer genetischer Zusammenhänge eine geeignete und breit anwendbare Methode zur genomischen Leistungsvorhersage in der Tier- und Pflanzenzucht zu finden und auf ihre praktische Umsetzung hin zu untersuchen. Eine vielversprechende Methode für diese Aufgabe stellen Künstliche Neuronale Netze (KNN) dar, die zu den Verfahren des maschinellen Lernens zählen.

Im **ersten** und **zweiten Kapitel** dieser Arbeit wurde ein umfassender Literaturüberblick über die Methode und ihre Anwendungen im Bereich der Pflanzen- und Tierzucht gegeben. Verschiedene Netzwerktypen und -topologien, sowie Lernalgorithmen und Datenvoraussetzungen wurden analysiert und im Hinblick auf die Anwendung für genomisch basierte Leistungsvorhersagen in der Tierzucht bewertet. Bei der Vorhersage von unbekannten Phänotypen spielen Aspekte wie die Aufbereitung der Daten, die Anpassung der Parameter des Lernalgorithmus, als auch die optimale Auswahl der Netzwerktopologie für den erfolgreichen Einsatz von KNNs eine große Rolle. Bisher veröffentlichte Studien belegen, dass die Überlegenheit von KNNs gegenüber anderen Methoden stark von der jeweiligen Datenstruktur, der Tierart, dem Zielmerkmal, den Umweltfaktoren und der Stichprobengröße, sowie von der genetischen Verwandtschaft zwischen den Individuen in den Trainings- und Testdatensätzen bestimmt wird. Im **dritten Kapitel** wurden die Erkenntnisse aus der Literaturrecherche praktisch umgesetzt. Dafür wurde ein eigenes Programm in C++ geschrieben, welches auch die Auswertung von größeren Datensätzen ermöglichte. Ein regularisierter Back-Propagation-Algorithmus wurde als Lernregel für ein mehrschichtiges KNN eingesetzt. Um die Haupteinflussfaktoren auf die Vorhersagegüte von drei Milchmerkmalen (Milchleistung, Fett-%, Eiweiß-%) erfassen zu können, wurden drei verschiedene Datensätze ausgewertet. Es wurden Daten von 3341 Fleckvieh Bullen, 2303 Holstein Friesian Bullen und 777 Holstein Friesian Kühen verwendet. Alle Tiere wurden mit einem 50k SNP-Array genotypisiert. In der Analyse wurden verschiedene nicht-lineare Netzwerktopologien und unterschiedliche Dateneingabestrukturen hinsichtlich der Vorhersagegüte untersucht. Die Ergebnisse zeigten, dass dimensions-reduzierende Maßnahmen zu wesentlich genaueren und konsistenteren Vorhersagegenauigkeiten in den untersuchten Merkmalen führten und dass der Einsatz der gesamten genetischen Information dessen Vorhersagekraft verminderte. Zudem wiesen die Ergebnisse darauf hin, dass das Potential von KNNs besonders im Rahmen der Vorhersage von Merkmalen zum Tragen kommt, bei denen auch nicht-additive Genwirkungen eine Rolle spielen (z.B. funktionelle Merkmale). Auf diese Ergebnisse aufbauend behandelt das **vierte Kapitel** die Vorhersage des Ketose-Risikos bei Milchkühen, da dieses multifaktoriell bedingt ist und die Vorhersage mit linearen Methoden nur bedingt möglich ist. In der Analyse wurden mittels eines KNN-Ansatzes mehrere unabhängige Variablen und Variablenkombinationen simultan auf ihre Vorhersageeignung für das Zielmerkmal Ketose-Risiko getestet. Unter anderem wurden verschiedene Metabolitenkonzentrationen in der Milch, SNP-Genotypen, sowie Milchleistungsmerkmale bezüglich ihres Einflusses auf die Genauigkeit der Vorhersage hin evaluiert. In allen Vorhersagemodellen wurde die in Milchproben gemessene Betahydroxybuttersäure (BHBA)-Konzentration als Biomarker für das Risiko einer Kuh, an einer subklinischen Ketose zu erkranken, verwendet. Dabei konnten durchschnittliche Korrelationen zwischen beobachteten und vorhergesagten Merkmalswerten von bis zu 0,643 erzielt werden, wenn Kombinationen von Stoffwechselinformationen und Milchleistungsdaten für die Vorhersage verwendet wurden. Stoffwechsel-, sowie milchleistungsbasierte Modelle erzielten eine höhere Vorhersagegenauigkeit als Modelle, die genetische Informationen verwendeten. Die Vorhersage von niedrig erblichen, komplexen Merkmalen mittels KNNs ist somit vielversprechend. Bezugnehmend auf die Ergebnisse des dritten Kapitels wurde im **fünften Kapitel** die Eignung einer sogenannten „Extreme Learning Machine“ (ELM) zur Vorhersage von Milchmerkmalen untersucht. Es handelt sich hierbei um eine schnelle Lernarchitektur für KNNs, die das sehr parameterintensive Lernen ersetzt und somit, im Gegensatz zu einem Lernalgorithmus, die Nutzung der gesamten Marker Informationen gewährleistet. Experimentelle Ergebnisse zeigten, dass der ELM-Ansatz in der Lage ist, eine gute Vorhersagegüte zu erreichen, während der Rechenaufwand gering gehalten wird.