Feature selection and intelligent livestock management

Alsahaf, Ahmad

DOI:
10.33612/diss.145238079

Samenvatting

Computationele dierfokkerij is afhankelijk van genetisch-statistische modellen die zijn gericht op het schatten van fokwaarden, die vervolgens worden gebruikt om dieren te rangschikken op basis van hun genetische potentie. Moderne veeteeltsystemen verzamelen echter grote hoeveelheden gegevens gedurende het leven van een dier die niet direct geschikt zijn voor die statistische modellen, zoals het periodiek fenotype en omgevingswaarnemingen. In dit proefschrift onderzoeken we het potentieel van het benutten van die aanvullende gegevens om toekomstige fenotypevoorspelling in vee te verbeteren met behulp van machine learning methoden. Hiervoor geven we twee voorbeelden. In hoofdstuk 2 laten we zien dat random forest regressie beter presteert dan lineaire modellen bij het voorspellen van de slachtleeftijd bij varkens op basis van een mix van fenotypische, genetische en stamboominformatie. De voorspellingen worden gedaan vóór het begin van de mestfase van de varkens en kunnen daarom worden gebruikt om ze in uniforme groepen te zetten op basis van hun voorspelde groeipercentages. In dezelfde toepassing tonen we daarnaast aan dat analyse van het belang van kenmerken afgeleid van machine learning een uitsplitsing kan geven van de componenten die verantwoordelijk zijn voor de waargenomen fenotypische prestatie, waardoor een alternatief wordt geboden voor model-gebaseerde vaste en willekeurige effectschatting.

In hoofdstuk 3 gebruiken we een combinatie van RGB-D computervisie en gecontroleerd leren om de spiermassa van levende varkens in te schatten. Het doel van die toepassing is om de subjectieve beoordelingen van een menselijke operator op de boerderij te vervangen en om derhalve de stress voor de dieren te verminderen. De resultaten laten zien dat het voorgestelde systeem de beoordelingspatronen van de menselijke operator nauwkeurig kan nabootsen. Kortom, de twee voorgestelde voorbeelden leveren meer bewijs - in overeenstemming met recente bevindingen in de literatuur - dat veeteelt en managementpraktijken kunnen wor-

den verbeterd door middel van datagestuurde modellen.

In hoofdstuk 4 van het proefschrift stellen we een nieuw algoritme voor selectie van wrapper-features voor op basis van boomensembles en boosting. Functieselectie is een subdiscipline van machine learning die zich bezighoudt met het onderscheiden van relevante invoerfuncties van irrelevante en overtollige functies. Met de toename van de afmetingen van de dataset en een toenemende belangstelling voor modeltransparantie, wordt betrouwbare feature-selectie steeds nuttiger. Datasets in de omics-velden in het bijzonder, zoals die gebruikt worden in moleculaire veredeling en in gepersonaliseerde geneeskunde, kunnen aanzienlijk profiteren van feature-selectie. In die studiegebieden bevatten datasets vaak een groot aantal kenmerken en een klein aantal samples, wat een uitdaging vormt voor machine learning methoden die grote aantallen samples nodig hebben om hun parameters te leren, zoals diepe neurale netwerken.

Het algoritme dat we voorstellen, genaamd FeatBoost, gebruikt een iteratief proces van boosting, of herweging van monsters, en modelevaluaties om functies te selecteren die relevant zijn en niet overbodig voor elkaar. We evalueren de prestaties van het algoritme aan de hand van een aantal benchmarks, waaronder ReliefF, een op filters gebaseerde selectiemethode, en twee alternatieve op boomsamenstellingen gebaseerde methoden, Boruta en XGBoost-afgeleide functieclassificatie. FeatBoost presteert beter dan de concurrerende methoden op de meeste van de geteste datasets.