

**BỘ GIÁO DỤC
VÀ ĐÀO TẠO**

**VIỆN HÀN LÂM KHOA HỌC
VÀ CÔNG NGHỆ VIỆT NAM**

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ



NGUYỄN THÙY TRANG

LUẬN VĂN THẠC SĨ NGÀNH HỆ THỐNG THÔNG TIN

**NGHIÊN CỨU MỘT SỐ VẤN ĐỀ ẢNH HƯỞNG ĐẾN HIỆU SUẤT CỦA
HỆ THỐNG PHÂN LOẠI HÀNH VI BÒ**

Hà Nội - 2023

BỘ GIÁO DỤC
VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC
VÀ CÔNG NGHỆ VIỆT NAM

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ



Nguyễn Thùy Trang

**NGHIÊN CỨU MỘT SỐ VẤN ĐỀ ẢNH HƯỞNG ĐẾN HIỆU SUẤT CỦA
HỆ THỐNG PHÂN LOẠI HÀNH VI BÒ**

LUẬN VĂN THẠC SĨ HỆ THỐNG THÔNG TIN
Mã số: 8480104

NGƯỜI HƯỚNG DẪN KHOA HỌC:

1. TS. Trần Đức Nghĩa

A handwritten signature in blue ink, likely belonging to TS. Trần Đức Nghĩa.

2. TS. Hồ Thị Phương

A handwritten signature in blue ink, likely belonging to TS. Hồ Thị Phương.

Hà Nội - 2023

LỜI CAM ĐOAN

Tôi xin cam đoan đề tài nghiên cứu trong luận văn này là công trình nghiên cứu của tôi dựa trên những tài liệu, số liệu do chính tôi tự tìm hiểu và nghiên cứu. Chính vì vậy, các kết quả nghiên cứu đảm bảo trung thực và khách quan nhất. Đồng thời, kết quả này chưa từng xuất hiện trong bất cứ một nghiên cứu nào. Các số liệu, kết quả nêu trong luận văn là trung thực nếu sai tôi hoàn chịu trách nhiệm trước pháp luật.

Hà Nội, ngày 1 tháng 12 năm 2023

HỌC VIÊN

A handwritten signature in blue ink that reads "Trang". The signature is written in a cursive style and is positioned over a faint, light-colored grid background.

Nguyễn Thùy Trang


LỜI CẢM ƠN

Luận văn được thực hiện tại Học viện Khoa học và Công nghệ - Viện Hàn lâm Khoa học và Công nghệ Việt Nam, dưới sự hướng dẫn của TS. Trần Đức Nghĩa và TS. Hồ Thị Phương. Tôi xin bày tỏ lòng biết ơn sâu sắc đến thầy cô về định hướng khoa học, thầy cô đã động viên, trao đổi nhiều kiến thức và chỉ bảo tôi vượt qua những khó khăn để hoàn thành luận văn này. Tôi trân trọng cảm ơn Học viện Khoa học và Công nghệ và Viện Công nghệ thông tin - Viện Hàn lâm Khoa học và Công nghệ Việt Nam đã tạo điều kiện thuận lợi cho tôi trong suốt quá trình nghiên cứu và hoàn thiện luận văn.

Cuối cùng tôi xin gửi lời cảm ơn sâu sắc đến gia đình, bạn bè và những người đã luôn ủng hộ và giúp đỡ tôi về mọi mặt để tôi yên tâm học tập thật tốt.

Hà Nội, ngày 1 tháng 12 năm 2023

HỌC VIÊN



Nguyễn Thùy Trang

MỤC LỤC

LỜI CAM ĐOAN	3
LỜI CẢM ƠN	4
MỤC LỤC	6
Mở đầu	10
Chương 1: Tổng quan về chăn nuôi chính xác	12
1.1 Chăn nuôi chính xác	12
1.1.1 Một số ứng dụng của PLF	14
1.1.2 Cảm biến và xử lý thông tin	17
1.1.3 Động lực thay đổi theo hướng chăn nuôi chính xác.....	19
1.2 Nhu cầu tương lai về hệ thống sản xuất sữa.....	21
1.2.1 Hiệu suất và phúc lợi của bò sữa	22
1.2.2 Hệ thống cảm biến hỗ trợ hoạt động chăn nuôi bò sữa.....	24
1.2.3 Nghiên cứu khái niệm hệ thống cảm biến trong ngành sữa ...	26
1.3 Cảm biến trong nghiên cứu sữa chính xác	29
1.3.1 Gia tốc kế để phân loại hành vi.....	29
1.3.2 GPS để phân loại vị trí và hành vi	32
1.3.3 Các cảm biến và công cụ khác để phân loại hành vi.....	35
Chương 2: Một số phương pháp học máy	39
2.1. Giới thiệu	39
2.2. Bộ tính năng cho máy học	39
2.3. Học máy có giám sát và không giám sát.....	41
2.4. Thuật toán học máy.....	42
2.4.1. Học máy dựa trên tập Luật	42
2.4.2. Phương pháp Naïve Bayes	44
2.4.3. Phương pháp Support Vector Machine (SVM).....	45
2.4.4. Phương pháp Hồi quy Logistic (Logistic regression)	48
2.4.5. Phương pháp tiếp cận học sâu (Deep Learning)	51

Chương 3: Phân loại hành vi bằng Random Forest và ảnh hưởng của độ lệch vị trí cảm biến đến hiệu suất phân loại hành vi	55
3.1. Vật liệu và phương pháp	56
3.1.1. Phân loại hành vi	56
3.1.2. Phân tích chuyên động.....	56
3.1.3. Phân đoạn và trích xuất đặc trưng	56
3.1.4. Dự đoán hành vi	57
3.2. Phân loại hành vi bằng Random Forest.....	57
3.2.1. Thuật toán.....	57
3.2.2. Các thuộc tính.....	60
3.3. Kết quả thực nghiệm	62
KẾT LUẬN.....	72
Những đóng góp của luận văn	72
Định hướng nghiên cứu trong tương lai.....	72
Tài liệu tham khảo	73

Mục lục hình ảnh

Hình 1: Hệ thống cho ăn tự động trong chăn nuôi bò [42]	12
Hình 2: Xu hướng nghiên cứu được công bố về PLF trong giai đoạn 2000-2018 sử dụng từ khóa “chăn nuôi chăn nuôi chính xác”. Biểu đồ hiển thị tổng số lượt tìm kiếm là 234 ấn phẩm được sản xuất kể từ năm 2000 do Web of Science lập chỉ mục (Thomson Reuters, 2019)	14
Hình 3: Những thay đổi về lượng thức ăn ăn vào, thời gian cho ăn và tỷ lệ cho ăn trước và sau khi bò được chẩn đoán (ngày 0) bị hôi chân. Chuyển thể từ González et al. (2008)	23
Hình 4: Ví dụ về cảm biến ghi âm và áp suất được sử dụng để phân loại hành vi của động vật nhai lại. A = máy ghi IGER (áp suất); B = Rumiwatch (áp suất); C = micro gắn trên đầu; D = micro gắn trên đầu và thiết bị ghi âm (b và a tương ứng)	35
Hình 5: SVM tìm dòng tốt nhất phân tách hai lớp.....	45
Hình 6: Ví dụ về siêu phẳng trong SVM.....	46
Hình 7: Siêu phẳng phân chia lề xa nhất	47
Hình 8: Minh họa đồ thị của hàm Logistic khi t thuộc (-6,6).....	49
Hình 9: Mạng RNN có vòng lặp.....	51
Hình 10: Mô đun lặp lại trong RNN.....	52
Hình 11: Mô đun lặp lại trong một LSTM.....	53
Hình 12: Dữ liệu 1CowData_2sensor phụ vụ đánh giá.....	63
Hình 13a: Xác suất của 4 hành vi với 7 dataset với thiết bị đeo trên cổ.....	63
Hình 13b: Xác suất của 4 hành vi với 7 dataset với thiết bị đeo trên cổ và chân.....	64
Hình 13c: Recall 4 hành vi của 7 dataset chỉ dùng cổ.....	64
Hình 14a: Accuracy 4 hành vi của 7 dataset dùng chân + cổ.....	66
Hình 14b: Precision 4 hành vi của 7 dataset dùng chân + cổ.....	67
Hình 14c: Recall 4 hành vi của 7 dataset dùng chân + cổ	68

Mục lục Bảng biểu

Bảng 1: Số liệu về độ chính xác theo chiều ngang được sử dụng để đo hiệu suất GPS theo định nghĩa của Viện Điều hướng (1997).....	33
Bảng 2: Bình quân các chỉ số của cảm biến cổ.....	75
Bảng 3: Bình quân các chỉ số của cảm biến gắn chân và cổ.....	76

Mở đầu

Trong lĩnh vực chăn nuôi thì Việt Nam có nhiều hướng mũi nhọn như chăn nuôi lợn, gà và gần đây là lĩnh vực chăn nuôi bò lấy thịt và lấy sữa đang thu hút được nhiều sự quan tâm nhà sản xuất. Việc phát triển chăn nuôi bò đang đem lại nhiều lợi ích kinh tế cho các hộ nông dân, nông trường, nông trại trên cả nước [1]. Trước đây, việc chăn nuôi bò đã phát triển tại các nông trại của các hộ gia đình của nhiều quốc gia trên thế giới. Với các hộ gia đình chăn nuôi nhỏ lẻ từ 1-5 con thì người chăn nuôi rất dễ dàng có thể theo dõi được sức khỏe và tình trạng khác của bò. Tuy nhiên, với quy mô đàn lớn hơn đặc biệt với quy mô hàng ngàn con thì việc theo dõi thủ công gây rất nhiều khó khăn và tốn kém chi phí cho chủ trang trại. Từ đó, việc đầu tư công nghệ cao vào chăn nuôi được nhiều chủ trang trại quan tâm và hướng đến nhằm đem lại lợi ích kinh tế cho họ. Việc giám sát hoạt động và hành vi của bò tự động tại chuồng cũng như khi chăn thả là cần thiết [2,3]. Từ những dữ liệu này, giúp quá trình phân tích các thay đổi về mặt hành vi của vật nuôi có thể giúp chủ trang trại nắm bắt được tình trạng sức khỏe của vật nuôi và có hướng xử lý cho vật nuôi. Điều này giúp quá trình quản lý một trang trại chăn nuôi ngày càng tốt hơn [4,5].

Cho đến nay, hầu hết các công việc liên quan đến giám sát bò ăn cỏ đều quan tâm đến thời gian chăn thả và lượng cỏ ăn vào [1, 6, 7]. Phân loại hành vi của bò có thể được chia thành hai loại, đeo thiết bị và không đeo thiết bị. Giám sát bằng camera thuộc loại không đeo thiết bị. Phương pháp này thoải mái cho bò (vì không có gì để mặc) nhưng có nhiều hạn chế về độ chính xác và phạm vi triển khai, hạn chế (đối với bò ăn cỏ). Có nhiều loại cảm biến có thể được gắn vào bò trong cách tiếp cận của đeo thiết bị. Chẳng hạn như xây dựng hệ thống giám sát thông qua GPS, cảm biến nhiệt độ, gia tốc kế và cảm biến từ trường 3 trục.

Hệ thống giám sát dựa trên gia tốc kế có thể phát hiện và phân loại nhất quán các hành vi của bò [8, 9, 10]. Từ đó, hệ thống giám sát là một yếu tố đầu vào cần thiết cho việc phát triển một hệ thống hỗ trợ chăn nuôi. Hệ thống này sử dụng phương pháp học máy nhằm phân tích dữ liệu gia tốc giúp phân loại và đưa ra các quyết định. Đối với một vấn đề có thể được giải quyết bằng ML, dữ liệu là rất quan trọng. Để thu thập dữ liệu hoạt động của con bò bằng gia tốc kế thì có hai cách cơ bản. Gắn gia tốc kế vào cổ hoặc chân của bò. Sự lựa chọn vị trí gắn gia tốc kế chủ yếu phụ thuộc việc muốn phân tích hành vi nào của vật nuôi, chẳng hạn như dữ liệu gia tốc được gắn trên cổ sẽ tốt hơn để phân biệt hành vi ăn. Tuy nhiên, việc sử dụng gia tốc kế có thể dẫn đến một số hành vi phân loại nhầm lẫn nhau. Ví dụ: nếu chúng chứa các cử chỉ tương tự (hành vi đứng và hành vi ăn – với dữ liệu gia tốc kế gắn trên chân vật nuôi).

Vấn đề đặt ra là trong quá trình vận động của vật nuôi, các thiết bị có thể bị đeo lệch. Điều này gây ảnh hưởng đến độ chính xác của dữ liệu thu thập được từ thiết bị. Mục tiêu của luận văn hướng đến việc mô phỏng khi thiết bị bị đeo lệch sẽ gây ra ảnh hưởng ra sao đến kết quả phân loại dữ liệu. Trong luận văn trình bày việc thu thập dữ liệu cho 4 hoạt động của bò gồm: đứng, nằm, đi và ăn. Để phân loại các hành vi, luận văn sử dụng kỹ thuật Random Forest với số cây là 25 để phân loại cùng với việc đánh giá qua các chỉ số Accuracy, Precision và Recall cho tập huấn luyện và dữ liệu thử nghiệm. Luận văn xây dựng 7 dataset ứng với việc thiết bị bị đeo lệch 0, 4, 8, 12, 16, 24, và 28 độ. Từ đó, rút ra được mức độ ảnh hưởng của việc đeo lệch thiết bị ảnh hưởng đến quá trình phân loại dữ liệu.

Chính vì vậy, em lựa chọn đề tài nghiên cứu ***“Nghiên cứu một số vấn đề ảnh hưởng đến hiệu suất của hệ thống phân loại hành vi bò.”***

Chương 1: Tổng quan về chăn nuôi chính xác

1.1 Chăn nuôi chính xác

Quy trình chăn nuôi chính xác (Precision livestock farming-PLF) [5,11, 12] là việc sử dụng các công nghệ tiên tiến để tối ưu hóa sự đóng góp vào sản phẩm của mỗi con vật. Thông qua việc tiếp cận đến từng cá thể vật nuôi, người nông dân có thể thu lại được kết quả tốt hơn trong chăn nuôi. Các kết quả có thể là định lượng, định tính hoặc giải quyết tính bền vững trong chăn nuôi, hoặc đơn giản là giảm lượng nhân công cần thiết. Đối với quá trình chăn nuôi, nếu là các động vật lớn như bò, lợn thì việc theo dõi tiến hành từng con. Còn đối với các động vật nhỏ như gà, vịt thì việc theo dõi tiến hành theo từng đàn.

Khái niệm nông nghiệp chính xác đã được xuất hiện cách hàng thập kỷ trong lĩnh vực cây trồng, ở đó công nghệ GPS giúp người dân giảm việc sử dụng thuốc diệt cỏ, trừ sâu và phân bón. Gần đây, nó cũng nhận được nhiều sự quan tâm trong ngành chăn nuôi. Cốt lõi của quy trình này là sử dụng các thiết bị thông minh vào hoạt động chăn nuôi và đưa ra các cảnh báo cần thiết cho chủ trang trại. Bằng cách ứng dụng các công nghệ chăn nuôi chính xác, người chăn nuôi có thể theo dõi tình trạng của từng cá thể vật nuôi trong trang trại như: nhiệt độ, dinh dưỡng, độ ẩm, hoặc thậm chí là tình trạng sức khỏe. Điều này giúp ích rất nhiều cho chủ trang trại trong việc phát hiện và chăm sóc kịp thời cho vật nuôi. Các thiết bị công nghệ thông minh có khả năng thu thập và cung cấp các thông tin cần thiết giúp giảm thiểu những tổn thất không mong muốn.

Đối với cấp độ từng cá thể vật nuôi, các chủ trang trại bò thịt, bò sữa và thịt lợn đã ứng dụng các cảm biến đeo nhằm cảnh báo sớm về bệnh và cải thiện các tiện ích trong chuồng trại cho phù hợp với cá thể vật nuôi. Họ cũng sử dụng các cảm biến có thể phát hiện việc thay đổi nhiệt độ, độ pH và âm thanh của từng cá thể nhằm phân tích hành vi và hoạt động của vật nuôi. Đối

với gia cầm, thì hiện tại chưa có thiết bị phục vụ cho từng cá thể nhưng được thay thế bằng các công nghệ gắn thẻ hồng ngoại và nhận dạng tần số vô tuyến để theo dõi hành vi của đàn gà. Từ những dữ liệu thu thập được, chủ trang trại có thể cải thiện hơn nữa chuồng nuôi nhốt và các vấn đề khác trong chăn nuôi.



Hình 1: Hệ thống cho ăn tự động trong chăn nuôi bò

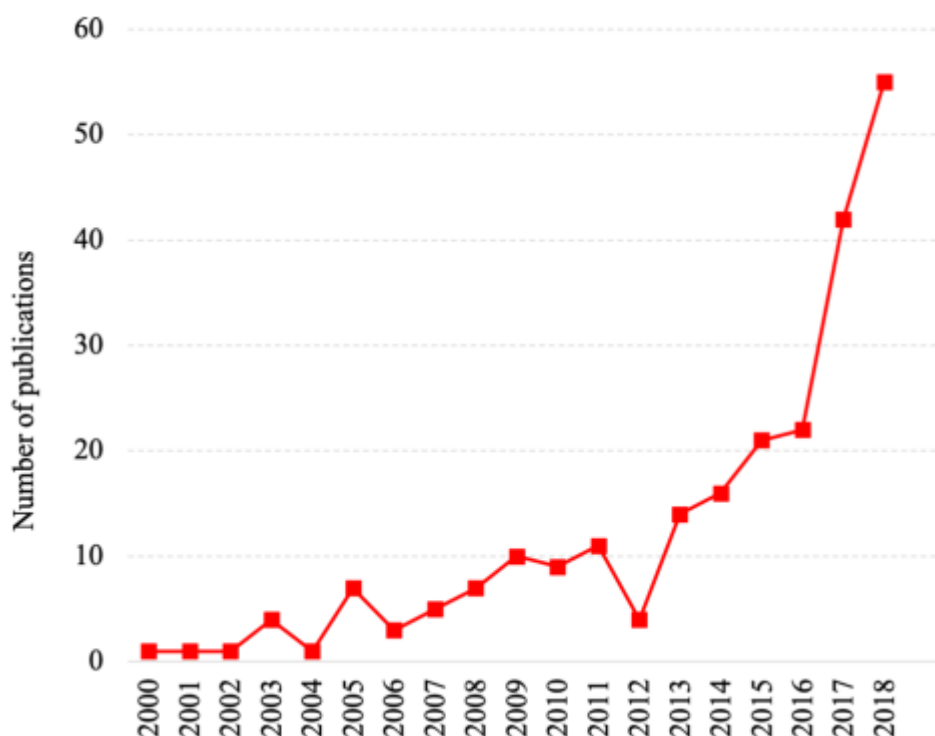
Ứng dụng PLF trong chăn nuôi bò sữa đã tăng nhiều trong những năm qua. Ở cấp độ cá thể, hoạt động của PLF thực sự tốt hơn vì đơn giản một con bò có giá trị cao hơn nhiều so với một con gà mái. Trong theo dõi hành vi chi tiết, quá trình theo dõi hoạt động của của vật nuôi đem lại nhiều dữ liệu có ích giúp cải thiện hiệu quả chăn nuôi hơn so với gia cầm. Đối với gia cầm, việc theo dõi được diễn ra theo hướng từng đàn hơn là theo từng cá thể riêng lẻ. Từ đó, chiến lược quản lý của chủ trang trại cũng thay đổi theo hướng phù hợp và chính xác hơn.

1.1.1 Một số ứng dụng của PLF

1.1.1.1 Ứng dụng trên đàn gà

Các dự án phát triển PLF trên chăn nuôi đàn gà đã được nghiên cứu nhiều tại các quốc gia lớn như Thụy Sĩ, Anh và các quốc gia Châu Âu [13, 14]. Ở trong các dự án này, những nhà nghiên cứu tiến hành các thí nghiệm cho các đàn gà nhằm xác định các yếu tố ảnh hưởng đến vật nuôi như ánh sáng, thức ăn, không khí, và mật độ của đàn gà. Các yếu tố hành vi của đàn gà như di chuyển từ chuồng ra khu vực sân chơi và khu vực ăn, hay khả năng đo đếm thời lượng ngủ, thời lượng gãi, mổ và giữ bụi trước khi về tổ. Các thông tin này được thu thập nhằm xác định tập tính của đàn gà và đưa ra các yếu tố đánh giá khách quan khi có những thay đổi từ phía các cá thể liên quan.

Ví dụ, việc những con gà mổ lông nhau trong đàn và các yếu tố gây ra hành vi này, các nhà nghiên cứu thấy rằng những con gà được thả tự do ít có xu hướng này hơn bị nhốt trong chuồng. Từ đó, họ có thể chọn lọc ra các đặc tính ảnh hưởng tới yếu tố này, chọn lọc các con gà gây ra vấn đề. Điều này sẽ cho phép các nhà chọn giống loại bỏ đi các đặc điểm không cần thiết của chúng trong quá trình nhân giống đàn gà. Từ việc thu thập dữ liệu cá thể cho thấy sau khi tiêm vắc-xin tiêu chuẩn, chuyển động của gà mái thay đổi và thay đổi này thường kéo dài từ 7 đến 8 ngày.



Hình 2: Xu hướng nghiên cứu được công bố về PLF trong giai đoạn 2000-2018 sử dụng từ khóa “chăn nuôi chính xác”. Biểu đồ hiển thị tổng số lượt tìm kiếm là 234 ấn phẩm được sản xuất kể từ năm 2000 do Web of Science lập chỉ mục (Thomson Reuters, 2019).

1.1.1.2 Ứng dụng trên chăn nuôi đàn heo

Ngành chăn nuôi sạch ngày càng nhận được sự quan tâm từ người tiêu dùng. Người tiêu dùng ngày càng có nhiều yêu cầu khắt khe hơn về thực phẩm tốt cho sức khỏe và hạn chế dư lượng kháng sinh trong thịt hay dư lượng thuốc trừ sâu trong rau quả. Điều này thúc đẩy ngành nông nghiệp chung cũng như ngành chăn nuôi nói riêng chuyển hướng đến thực phẩm an toàn. Để phát triển hệ thống chăn nuôi không kháng sinh toàn diện thì vấn đề dinh dưỡng chỉ là một yếu tố cơ bản, bên cạnh đó còn các yếu tố khác như chất lượng con giống, chuồng trại và trang thiết bị, khả năng quản lý trang trại, kiểm soát sức khỏe và các nguồn nhân lực cũng đang là vấn đề cấp thiết. Đối với chăn nuôi sạch, yếu tố cốt lõi là tập trung vào tăng cường khả năng

phòng bệnh, xác định chính xác và nhanh chóng mầm bệnh, các cá thể có thể nhiễm bệnh để chủ động ngăn chặn sự phát triển của mầm bệnh.

Quy trình chăn nuôi heo sạch không kháng sinh đã áp dụng dần các biện pháp của PLF. Các dữ liệu được thu thập của chuồng nuôi liên tục được cập nhật và phân tích theo thời gian thực. Chúng sẽ được kết nối với các thiết bị cho ăn nhằm đưa ra các quyết định cho ăn chính xác ứng với tình trạng sức khỏe của từng cá thể heo. Thức ăn cho heo cũng được phân loại và điều chỉnh theo khối lượng và nhu cầu của từng cá thể. Vì vậy, việc cho ăn cũng cần được thực hiện theo các mô hình và tiến hành đánh giá cũng như cải tiến nhằm phù hợp với nhu cầu của từng cá thể heo trong sản xuất thương mại.

1.1.1.3 Ứng dụng vào cho ăn tự động

Việc ứng dụng PLF vào cho ăn có nghĩa là cung cấp đầy đủ chất dinh dưỡng cho cá thể vật nuôi với số lượng cần thiết. Các phương pháp này áp dụng trong chăn nuôi heo và gia cầm giúp chủ trang trại giảm tình trạng lãng phí thức ăn và chi phí thức ăn. Đồng thời, PLF vẫn xử lý được các vấn đề về môi trường và sản xuất. Cho ăn chính xác là một phần quan trọng giúp giảm giá thành của sản phẩm từ thức ăn chăn và giảm thiểu nguồn nhân lực cần thiết phục vụ cho chăn nuôi.

Hệ thống PLF sẽ cải tiến một số phần trong quy trình chăn nuôi từ thủ công sang tự động thông qua khả năng sử dụng các công nghệ theo dõi vật nuôi, quan sát cách vật nuôi ăn và phát triển. Tùy từng trường hợp, theo mô hình của hệ thống sẽ lựa chọn công thức để trộn thức ăn phù hợp theo các dữ liệu đã được lập trình sẵn. Việc phân phối thức ăn cũng là tự động, giảm thiểu vai trò của nhân công trong hệ thống. Từ đó, vai trò của các nhà chăn nuôi và chuyên gia dinh dưỡng sẽ thay đổi. Hệ thống cho ăn chính xác được phát triển dựa vào các thông tin thu thập từ lượng thức ăn tiêu thụ và hành vi của vật nuôi. Các chuyên gia sẽ thêm và các yếu tố cảm biến khác, các mô hình để ước tính về mức độ dinh dưỡng (17) nhằm xây dựng công thức chính xác hơn.

Từ đó, hệ thống cho ăn chính xác giúp các nhà sản xuất cung cấp đúng và vừa đủ các chất dinh dưỡng cần thiết.

Hệ thống thức ăn chuẩn xác này còn tăng việc tự động hóa trong sản xuất, sử dụng công nghệ theo dõi và quan sát hành vi ăn và quá trình sinh trưởng. Máy tính sẽ là người quyết định dựa vào các thông tin thu thập được, và nó có thể sẽ trộn 2-3 loại thức ăn khác nhau trong một số trường hợp. Việc cho ăn cũng sẽ được tự động hoá, và nhà sản xuất và nhà dinh dưỡng khi đó sẽ có những vai trò mới trong sản xuất. Việc thay đổi phương pháp sản xuất, như việc tiến tới sản xuất không sử dụng kháng sinh, sẽ mang đến nhiều thử thách về hiệu suất thức ăn cho nhà sản xuất gia cầm và có thể hướng đến sử dụng nguyên liệu mới. Phương pháp dinh dưỡng chuẩn xác trong chăn nuôi lợn và gia cầm giúp nhà sản xuất tránh lãng phí các chất dinh dưỡng dư thừa và giảm chi phí liên quan đến thức ăn, đồng thời giải quyết phần nào những vấn đề môi trường và những khó khăn trong sản xuất.

Các nghiên cứu về dinh dưỡng, mô hình toán học, các hệ thống cần thiết của PLF trong cho ăn chính xác được nghiên cứu trong lĩnh vực chăn nuôi heo. Công nghệ dinh dưỡng chuẩn xác đã mang lại nhiều lợi ích bao gồm giảm lượng thức ăn được sử dụng và chi phí chăn nuôi khoảng 8% chi phí cho ăn và sản xuất. Điều này cũng giúp dự đoán chính xác về thời gian xuất chuồng và thiết lập các kế hoạch phù hợp hơn và cải thiện môi trường xung quanh cũng như giúp giảm áp lực lên môi trường. Hướng tiếp cận dinh dưỡng chính xác này còn giúp giảm bài tiết nitơ ít nhất 30% và giảm lượng phốt pho, phát hiện sớm dịch bệnh là một lợi ích khác mà người chăn nuôi sẽ nhận được.

1.1.2 Cảm biến và xử lý thông tin

Một hệ thống PLF có thể được cấu trúc sao cho người điều hành trang trại nhận được thường xuyên phản hồi từ các cảm biến dựa trên động vật hoặc cơ sở hạ tầng. Thông tin này có thể được sử dụng để đưa ra các

quyết định quản lý dài hạn hoặc thường xuyên hơn được thực hiện bằng tay hoặc tự động. Đo hoạt động của bò sữa bằng máy đếm bước chân. Mục đích của việc phát hiện động dục là một ví dụ về tự động hóa một phần một tập hợp con của hệ thống bò sữa. Điều này đã dẫn đến tỷ lệ thụ thai cao hơn và cải thiện trang trại hiệu quả [15-20].

Các lĩnh vực khác cũng được hưởng lợi. Một số nhà sản xuất trong lĩnh vực gia cầm sử dụng hệ thống cho ăn chính xác nơi việc phân bổ thức ăn có thể được điều chỉnh để đáp ứng mục tiêu tăng trưởng của từng loài chim. Trong tương lai, có thể có khả năng bổ sung thêm chiều cho các hệ thống này. Dữ liệu thu được từ người khác cảm biến hoặc internet có thể thêm ngữ cảnh và làm phong phú thêm quy trình quản lý bằng thực tế thông tin thời gian về hành vi của các loài cùng loài, vi khí hậu, gánh nặng ký sinh trùng, truyền bệnh cũng như thông tin thị trường, tất cả những điều này sẽ cho phép các nhà điều hành phản hồi ngay lập tức. Trên thực tế, hiện nay có rất nhiều công cụ và kỹ thuật đang được phát triển cho hầu hết các loài và cho nhiều mục đích khác nhau. Chúng bao gồm các công cụ để đo tự động tổng lượng thức ăn và nước uống của vật nuôi, tình trạng sức khỏe, tương tác cụ thể, bệnh tật, chuyển dạ, tư thế, nhiệt độ và âm thanh của gia súc cũng như ảnh hưởng của các yếu tố này đối với PLF.

Để nắm bắt thông tin hành vi, một bộ cảm biến và công nghệ riêng biệt được yêu cầu cho các nhiệm vụ cụ thể, việc lựa chọn chúng sẽ phụ thuộc vào một số biến. Ví dụ: loài được đề cập, thước đo mong muốn để đo lường và định lượng (ví dụ như tư thế động vật và tần suất các tư thế), hệ thống chuồng trại (trong nhà, nuôi thả), quy mô đàn, cách bố trí trang trại, kết nối internet, nhân viên cơ hội đào tạo và kết quả mong muốn. Các cảm biến này cần phải chịu được các áp lực từ môi trường, có độ chính xác cao và phải được tích hợp vào các hệ thống có khả năng hỗ trợ tốt cũng như dễ dàng cung cấp các thông tin đến với người dùng nhằm giúp họ đưa ra được các quyết định hợp

lý. Lý tưởng nhất là các hệ thống PLF phải đủ linh hoạt để đa chức năng (ví dụ như gia súc kể 3 trục gắn vào cổ gia súc có thể được lập trình để ghi lại hoạt động nhai lại, cho ăn và tư thế) nhưng một số hệ thống sẽ cụ thể cho một mục đích cụ thể. Ví dụ: Tư thế tự động dựa trên hình ảnh của hệ thống phân loại sẽ được thiết kế đặc biệt để phân biệt giữa con vật bị khập khiễng và những con không bị.

1.1.3 Động lực thay đổi theo hướng chăn nuôi chính xác

Tiện ích của bất kỳ hệ thống PLF nào là nền tảng cho mức độ áp dụng giữa các nông dân và người thực hành. Ngoài các yếu tố về công nghệ còn các yếu tố khác như ưu đãi của chính phủ hay khả năng thu thập dữ liệu chính xác của đàn gia súc. Ở đó, có thể là mong muốn tăng quy mô đàn vì lý do lợi nhuận hoặc hiệu quả, hoặc các nhà quản lý có thể thấy trước việc giảm lao động sẵn có và những người vận hành có tay nghề cao. Dữ liệu được thu thập tại các trang trại cừu ở Anh và xứ Wales cho thấy kiến thức về CNTT, khả năng sử dụng các thiết bị thông minh, thời gian quản lý đàn cừu hay nhu cầu về tăng cường sản xuất có liên quan rất lớn đến việc nhận dạng qua thiết bị điện tử.

Một hệ thống PLF đầy đủ nhìn chung sẽ gây nhiều khó khăn cho nông dân vì giá cả, công nghệ, kiến thức và khả năng áp dụng. Vì vậy, đôi khi để cải thiện hiệu quả có thể sử dụng PLF ở một khu vực nhất định, một phần hoặc tự động hóa một bộ phận của hệ thống nhằm đạt hiệu quả tối ưu và đạt được mong muốn của chủ trang trại. Họ nhận thấy rằng quyết định đầu tư vào một hệ thống tự động hệ thống phát hiện bệnh khập khiễng phụ thuộc đáng kể vào tầm quan trọng của người nông dân đến sự khập khiễng (quan trọng hơn = tiện ích được nhận thức lớn hơn của hệ thống), khoảng thời gian giữa lúc sinh bê và lần thụ tinh đầu tiên ở trang trại của họ (khoảng thời gian tăng = tăng nhận thức được tiện ích của hệ thống) và liệu nông dân đã áp dụng biện

pháp động dục hay chưa hệ thống phát hiện (đã được áp dụng = tiện ích nhận thức của hệ thống tăng lên).

Điều thú vị là, một khi cái giá phải trả cho việc đi khắp khiêng đã được thông báo tới người nông dân ở vùng này nghiên cứu, tiện ích được nhận thấy của thiết bị phát hiện bệnh khập khiễng đã tăng lên đáng kể nhưng điều này chỉ đúng với những nông dân đã sử dụng thiết bị phát hiện động dục. Rõ ràng, cần phải làm nhiều hơn nữa để thông báo cho nông dân về tác động kinh tế của một số rối loạn sản xuất, đặc biệt là những vấn đề mà dường như không có hậu quả tài chính trực tiếp.

Khi áp lực của thị trường đối với nhà sản xuất với vấn đề nâng cao hiệu quả, giảm việc sử dụng thuốc kháng sinh và đảm bảo phúc lợi động vật, có những cơ hội rõ ràng để nông dân phát triển và áp dụng công nghệ PLF vào quản lý ứng hộ. Ưu điểm rõ ràng của việc sử dụng công nghệ PLF là khả năng tiết kiệm lao động. Đúng vậy, với hệ thống PLF khả năng tiết kiệm sức lao động và giảm nhân công là lý do quan trọng nhất khiến họ quyết tâm đầu tư vào hệ thống cảm biến.

Một trong những ứng dụng phổ biến nhất của PLF trong ngành sữa là phát hiện động dục. Nếu không phát hiện động dục tự động, nông dân nên quan sát bò trong ít nhất 20 phút, ba lần mỗi ngày đối với các đợt nắng nóng. Nhưng với hệ thống giám sát sẽ tự động thu thập dữ liệu giúp thời gian và chi phí lao động. Hơn nữa, công nghệ PLF mục tiêu và khi được hỗ trợ bởi các thuật toán dự đoán mạnh mẽ có thể phân vùng và xác định chính xác vật nuôi cần được quan tâm sớm hơn so với người quan sát. Tuy nhiên, cần nhiều công việc hơn trong việc phát triển hệ thống vì khả năng chẩn đoán đã được chứng minh là kém hiệu quả hơn ý kiến chuyên môn.

Dù động lực áp dụng hoặc tích hợp các kỹ thuật PLF là gì thì ít nhất nó cũng có trong tương lai gần khó có khả năng những người vận hành con người sẽ bị loại bỏ hoàn toàn khỏi các hoạt động khác nhau các quy trình chi

phối hệ thống chăn nuôi. Cũng khó có khả năng PLF sẽ thay thế quản lý kém. Thay vào đó, tương lai của ngành chăn nuôi các hệ thống sản xuất có nhiều khả năng được hỗ trợ bởi các công cụ hỗ trợ quyết định sẽ nâng cao chất lượng sản phẩm chăn nuôi và tạo cơ hội cho các hệ thống năng suất và bền vững hơn.

1.2 Nhu cầu tương lai về hệ thống sản xuất sữa

Sản xuất sữa toàn cầu đã có sự tăng trưởng và thâm canh đáng kể trong 50 năm qua và nhu cầu về các sản phẩm sữa có thể sẽ tiếp tục tăng trong tương lai gần. Trên toàn cầu, sữa mức tiêu thụ dự kiến sẽ tăng trung bình 27% từ 87kg lên 119kg mỗi người vào khoảng 2067 phần lớn được thúc đẩy bởi áp lực dân số từ các đô thị lớn.

Đối với ngành sữa, mức tăng trưởng sản lượng này phần lớn sẽ được đáp ứng thông qua cải thiện lượng ăn vào và chất lượng của thức ăn được lựa chọn ứng với vật nuôi phù hợp. Điều này giúp hệ thống PLF với các quy trình lựa chọn thức ăn tự động, khả năng giám sát và thu thập dữ liệu hỗ trợ ra quyết định mong muốn đáp ứng được nhu cầu của người tiêu dùng. Đối với các trang trại sản xuất lớn thì các hệ thống quản lý chặt chẽ chiếm được ưu thế lớn. Áp lực bổ sung sẽ xảy ra với các hệ thống này bao gồm việc đảm bảo tính bền vững, sự chấp nhận về mặt đạo đức và các hệ thống đó thân thiện với môi trường.

Ví dụ, hệ thống chăn nuôi bò sữa được quản lý chặt chẽ thường có nghĩa là những con bò bị giam giữ trong chuồng trong phần lớn cuộc đời sản xuất sữa của chúng, thường là dẫn đến tiềm năng sản xuất lớn hơn mà còn nâng cao nhận thức của người tiêu dùng đối với phúc lợi của bò và sự an toàn của các sản phẩm sữa. Mặt khác, đồng cỏ các hệ thống dựa trên thường được người tiêu dùng và phúc lợi công nhận là tự nhiên hơn những cân nhắc có thể bị bỏ qua. Đối với các hệ thống dựa trên đồng cỏ, giá trị cũng có thể được tăng thêm cho các sản phẩm sữa bằng cách nêu bật

một số đặc điểm về chất lượng sữa các thuộc tính liên quan như hồ sơ axit béo. Tuy nhiên, các hệ thống phải đối mặt với những thách thức của riêng mình như điều kiện thời tiết thay đổi thất thường, khả năng theo dõi sinh lý và các chỉ số sức khỏe giảm theo thời gian tiếp xúc với nhân viên.

Cùng với những thách thức này là những cơ hội và ngành công nghiệp sữa đang nhanh chóng phát triển và sử dụng các công nghệ sẵn có để hỗ trợ hàng ngày quản lý. Từ góc độ sản xuất, những nỗ lực hiện nay tập trung vào phát triển các cảm biến có thể hỗ trợ việc quản lý hàng ngày của từng cá nhân động vật. Hầu hết các nghiên cứu trước đây đều tập trung vào quản lý khả năng sinh sản và các vấn đề về vận động nhưng ngày càng có nhiều vấn đề hơn người ta đang chú ý đến các cảm biến có thể đo lường thức ăn ăn vào của từng cá thể và phát hiện bệnh.

1.2.1 Hiệu suất và phúc lợi của bò sữa

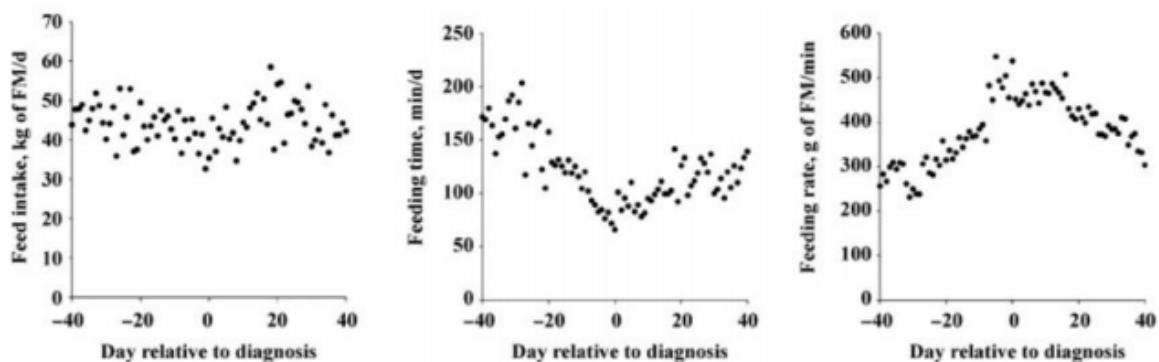
Chủ yếu do áp lực kinh tế, số lượng đàn bò sữa ở Anh đang tăng giảm nhanh chóng và tính đến tháng 1 năm 2019, tổng số 9.170 nhà sản xuất đã được ghi nhận ở Anh và xứ Wales; Ít hơn 33 so với tháng trước và ít hơn 177 so với tháng 1 - 2018 (AHDB Sữa, 2019a). Nhờ đó, số lượng bò mỗi đàn ngày càng tăng. Các quy mô đàn bò trung bình hiện tại ở Anh (dữ liệu năm 2018) ước tính là 148 con, tăng 2 con/đàn so với năm trước (2017) và nhiều hơn 32 con so với năm 2008. Sản lượng sữa trung bình trên mỗi con bò ở Anh năm 2018 ước tính là 7.825 lít. Năm 2008, con số này là 6.974 lít.

Mặc dù tăng sản xuất thường gắn liền với tình trạng sức khỏe giảm sút của bò sữa, nhưng đang có khá ít các nghiên cứu mối tương quan giữa việc tăng sản lượng với nguy cơ gây bệnh đối với vật nuôi. Một nghiên cứu cho rằng sản lượng cao hơn có liên quan đến cao hơn mắc bệnh viêm vú và khớp khiễng nhưng những người khác không tìm thấy mối quan hệ giữa tăng năng suất sữa và các bệnh như đẻ khó và viêm tử cung.

Theo hiểu biết của ứng viên, số lượng nhân viên hiện có trên mỗi con vật ở các trang trại bò sữa ở Anh vẫn chưa được biết nhưng hầu hết bằng chứng đều cho thấy rằng giải pháp kinh tế tốt nhất các trang trại đang hoạt động sử dụng ít lao động hơn trên mỗi con bò, với 25% trang trại hàng đầu ở Anh phân bổ 35h/bò mỗi năm so với 49 giờ/bò mỗi năm ở 25% đàn có thành tích thấp nhất. Một số bằng chứng cho thấy rằng khi số lượng bò trên mỗi đơn vị lao động tăng lên thì việc xác định kịp thời các bệnh như khập khiễng cũng tăng lên. Việc mở rộng quy mô đàn và khả năng chăm sóc từng cá thể trong đàn rất phức tạp. Một số người đã tìm thấy bằng chứng về việc giảm tình trạng nguy cơ bệnh khập khiễng ở các đàn lớn hơn nhưng tỷ lệ mắc bệnh trong đàn gia tăng các bệnh truyền nhiễm khác như bệnh Johne và bệnh lao bò.

Công nghệ PLF mang đến cơ hội đo lường khách quan những thay đổi hàng ngày trong hành vi của bò và cảnh báo người nông dân về sự cần thiết phải can thiệp khi cần thiết.

Ví dụ, hành vi cho ăn có thể mang lại cái nhìn đa chiều về sức khỏe và năng suất của bò. Trong một phân tích hồi cứu dữ liệu cho ăn riêng lẻ từ máy cấp liệu vi tính, họ phát hiện ra rằng bò sữa sau đó được chẩn đoán mắc bệnh ketosis cho thấy lượng thức ăn ăn vào hàng ngày giảm đáng kể (-10,4 kg chất tươi (FM)/d), thời gian cho ăn hàng ngày (-45,5 phút/d) và tốc độ cho ăn hàng ngày (-25,2 g FM/phút mỗi ngày), trung bình 3,6 ngày trước khi chẩn đoán. Bò được chẩn đoán mắc chứng vận động vắn đề là lượng ăn vào hàng ngày giảm (-1,57 kg FM/ngày), thời gian cho ăn hàng ngày (-19,1 phút/ngày) và tốc độ cho ăn tăng hàng ngày (+21,6 g FM/phút mỗi ngày) (Hình 3). Một số ấn phẩm đã báo cáo các mô hình tương tự về rối loạn sản xuất, và đây hiện là một lĩnh vực nghiên cứu rất mạnh mẽ trong khoa học sữa.



(González *et al.*, 2008)

Hình 3: Những thay đổi về lượng thức ăn ăn vào, thời gian cho ăn và tỷ lệ cho ăn trước và sau khi bò được chẩn đoán (ngày 0) bị hôi chân. Chuyên thể từ González *et al.* (2008).

1.2.2 Hệ thống cảm biến hỗ trợ hoạt động chăn nuôi bò sữa

Khi các hệ thống sản xuất sữa phát triển, cơ hội hỗ trợ quản lý bằng cách sử dụng cảm biến có rất nhiều công nghệ dựa trên và nông dân đã nhận ra lợi ích của một số công nghệ nhất định công cụ. Quả thực, có vẻ như khi số lượng bò trong mỗi trang trại tăng lên thì số lượng bò cũng tăng theo. mức độ áp dụng các công nghệ chính xác. Một cuộc khảo sát của Hà Lan cho thấy 39% các trang trại được khảo sát (n = 512) đã sử dụng hệ thống cảm biến và cảm biến về bệnh viêm vú và động dục phát hiện là phổ biến nhất. Hơn nữa, các trang trại có hệ thống cảm biến có ít số giờ lao động trên mỗi con bò nhưng năng suất mỗi con bò được đo bằng sản lượng sữa trên mỗi con bò, số lượng bò trên một ha và sản lượng sữa trên một ha không khác nhau giữa trang trại có và không có hệ thống cảm biến. Trong một nghiên cứu của Úc (n = 199 trang trại), trang trại với hơn 500 con bò (16,4%) được áp dụng độ chính xác cao hơn từ hai đến năm lần công nghệ so với các trang trại nhỏ hơn như hệ thống nhận diện bò bằng điện tử, phân loại tự động qua cổng ra vào và phần mềm quản lý đàn. Thực vậy, việc ứng dụng các công nghệ PLF vào trong các trang trại lớn có thể là do việc cần quản lý

số lượng lớn các cá thể bò. Việc này đòi hỏi chủ trang trại cần quyết định giữa việc đầu tư vào hệ thống hay chi trả chi phí nhân công đắt đỏ.

Điều quan trọng cơ bản là nông dân phải nhận thức được tiềm năng phản ứng sản xuất do việc áp dụng hệ thống cảm biến trước khi đầu tư được làm. Với các hệ thống vắt sữa, việc thực hiện tối ưu có thể đem lại các hiệu quả. Ví dụ, Steenveld và cộng sự nhận thấy SCC trung bình của đàn tăng (+12.000 tế bào/mL) tại các trang trại trong những năm sau khi đầu tư vào cảm biến hệ thống phát hiện viêm vú trong khi SCC trung bình giảm (-10.000 tế bào/mL) trong năm sau khi đầu tư vào các trang trại có hệ thống vắt sữa thông thường. Các tác giả kết luận rằng những tác động như vậy có thể được giải thích bởi các trang trại trải qua những thay đổi lớn khác vào hệ thống của họ chẳng hạn như đầu tư vào một hệ thống vắt sữa mới. Theo suy đoán, nó có thể là rằng nông dân đã có khả năng đưa ra các quyết định kịp thời, nhanh chóng và sáng suốt hơn nhờ vào hệ thống cảm biến. Cần phải thận trọng để đảm bảo rằng những lựa chọn sáng suốt được đưa ra trước khi đầu tư vào hệ thống cảm biến để không bị mất đi bất kỳ cải tiến nào về hiệu quả trong các khía cạnh khác của hệ thống sản xuất (hình 3b). Giảm lao động là một trong những lý do quan trọng nhất để đầu tư vào hệ thống cảm biến trong nghiên cứu đó. Các lợi ích kinh tế tổng thể của đầu tư có thể đến từ việc giảm lao động chi phí và không nhất thiết là từ những cải thiện về sức khỏe và hiệu suất.

Cho đến nay, phần lớn các hệ thống cảm biến được sử dụng trong trang trại có thể được mô tả như các hệ thống được mua như một phần của khoản đầu tư lớn hơn (ví dụ: độ dẫn điện của sữa mét trong AMS) hoặc những thứ được mua như một phần của chiến lược cụ thể để cải thiện hiệu suất ở một khu vực cụ thể (ví dụ: máy đếm bước chân để phát hiện động dục). Đứng trên góc độ nghiên cứu, các hệ thống cảm biến đã và đang được

phát triển nhằm hướng đến việc ứng dụng trong chăn nuôi tại các trang trại lớn.

Phần lớn các nghiên cứu liên quan đến bệnh viêm vú ($n = 31$) và khả năng sinh sản ($n = 41$) đều được thực hiện ở cấp độ hai (dữ liệu thu thập được sử dụng để giải thích thêm về tình trạng của một con bò, ví dụ như giảm hoạt động). Không tìm thấy ấn phẩm nào ở cấp độ ba (thông tin cảm biến được bổ sung với lời khuyên từ các nguồn bên ngoài, ví dụ: kinh tế) và bốn (nông dân hoặc hệ thống cảm biến đưa ra quyết định dựa trên thông tin ví dụ như gọi bác sĩ thú y). Người ta dự tính rằng trong thập kỷ tới, các cảm biến với khả năng giám sát các biến số hoạt động của bò sẽ được sử dụng, cho phép khả năng tiếp cận của nông dân với những thông tin mà trước đây không thể tiếp cận được. Cần nhiều hơn nữa để đảm bảo rằng nghiên cứu khái niệm được đưa lên các cấp độ cao hơn trong hệ thống phân cấp này.

1.2.3 Nghiên cứu khái niệm hệ thống cảm biến trong ngành sữa

Giảm chi phí lao động không phải là mục tiêu duy nhất của việc phát triển hệ thống cảm biến. Hệ thống cảm biến có tiềm năng không chỉ cung cấp các thông tin hữu ích nhanh chóng giúp giảm chi phí nhân lực mà còn giúp họ đưa ra các quyết định hợp lý hơn đối với từng cá thể vật nuôi.

Một nhóm nghiên cứu hiện đang phát triển việc sử dụng các cảm biến khác nhau để định lượng hành vi của bò sữa để phát hiện các bệnh và tình trạng khập khiễng, các biến số liên quan với thời điểm bắt đầu đẻ như nhiệt độ và thời gian nhai lại.

Mặc dù còn ở giai đoạn sơ khai, việc sử dụng hành vi như một đại diện cho sản xuất và phúc lợi quản lý có tiềm năng lớn và có thể sẽ có tính năng cao trong hệ thống cảm biến trong tương lai các ứng dụng. Hiện tại, cần nhiều hơn nữa để định lượng các hành vi mang lại nhiều tác động ảnh hưởng đến khả năng phát bệnh hoặc sự thay đổi của các hành vi theo thời gian. Hơn nữa,

tác động của các biến số khác lên hành vi cần phải được đánh giá đầy đủ. Ví dụ, trong hệ thống chăn nuôi bò trên các đồng cỏ, sự thay đổi về chất lượng cỏ có thể ảnh hưởng đến thời gian cần thiết cho việc chăn thả. Điều này có khả năng làm nhầm lẫn bất kỳ dấu hiệu nào về sự khởi đầu của một căn bệnh cụ thể.

Việc lựa chọn cảm biến thích hợp nhất cho nghiên cứu cũng rất quan trọng. Lựa chọn cảm biến sẽ phụ thuộc nhiều vào mục tiêu nghiên cứu (ví dụ hỗ trợ sinh sản, giảm sử dụng kháng sinh), hệ thống quản lý (ví dụ: bò nuôi so với chăn thả), các hành vi cần thực hiện được xác định (bị ảnh hưởng bởi mục tiêu nghiên cứu), khoảng thời gian lấy mẫu dữ liệu, dung lượng dữ liệu cảm biến và tuổi thọ pin cũng như vị trí của cảm biến trên động vật. Từ một nghiên cứu phối cảnh, các cảm biến được sử dụng để theo dõi hành vi cá nhân trong hầu hết các trường hợp là được thiết kế và xây dựng cho nhiệm vụ hiện tại và thường không mang tính thương mại sản phẩm có sẵn. Trong nghiên cứu về gia súc, các hệ thống cảm biến đã phát triển có nhiều cải tiến trong công nghệ cảm biến và nhận diện cho phép thu thập dữ liệu đơn chiều (GPS) phục vụ cho các công việc thống kê. Tuy nhiên, ít ấn phẩm hiện đang thảo luận về hiệu suất của các sản phẩm thương mại có sẵn.

Một mục tiêu nghiên cứu chính được triển khai gần đây là xác định hành vi của từng cá thể vật nuôi được thu thập bằng nhiều kỹ thuật phân loại. Trong này trường hợp, nhà phân tích được hưởng lợi từ thông tin hành vi có thể giải thích trực tiếp. Tuy nhiên, gần đây người ta đang chú ý đến việc sử dụng trực tiếp dữ liệu cảm biến thô. Một ví dụ: phân tích các hành vi của bò sữa trong không gian chuồng. Những phân tích này không đòi hỏi cao về hệ thống tiền xử lý dữ liệu. Từ đó, một phương tiện hiệu quả để thu được nhiều thông tin hơn về sự di chuyển của vật nuôi trong khi vẫn duy trì khả năng giám sát lâu dài sức khỏe.

Việc hiện thực hóa khả năng quản lý từng con vật một cách chính xác để một loạt các thông số quan trọng về thương mại và phúc lợi đã bùng phát ở một mức độ nào đó trong 10 năm qua. Trước khi tiến hành thảo luận về một loạt cảm biến dựa trên động vật hiện đang được sử dụng trong nghiên cứu bò sữa, cần tóm tắt đánh giá chính số liệu của các mô hình có sử dụng dữ liệu được thu thập từ các thiết bị cảm biến trong lĩnh vực phân loại hành vi.

1.3 Cảm biến trong nghiên cứu sữa chính xác

1.3.1 Gia tốc kế để phân loại hành vi

Gia tốc kế đo lực g và độ nghiêng theo ba trục (X , Y và Z) trước khoảng thời gian được lập trình. Những cảm biến này đã được sử dụng rộng rãi để theo dõi hành vi và trong thập kỷ qua, đã được chứng minh là rất có giá trị trong các nghiên cứu phân loại hành vi của cá thể khác nhau và các loài nuôi trong nhà. Vì với bò sữa mối tương quan giữa các hành vi của bò đến năng suất sữa với khả năng chăm sóc là đầy hứa hẹn. Ví dụ, mục tiêu của một số ứng dụng đầu tiên của gia tốc kế được trang bị cho động vật nhai lại nhằm xác định các thuộc tính của hành vi kiếm ăn.

Nghiên cứu trong 10 năm qua tập trung chủ yếu vào tư thế nằm và đứng hành vi của gia súc cũng như phân loại hành vi chung. Đến nay, người ta đã thừa nhận rõ ràng rằng những hành vi có thể nhận dạng này có liên quan chặt chẽ với các bệnh như khớp khiễng và viêm vú cũng như sự thoải mái chung của bò. Bây giờ có một tập trung mạnh vào việc sử dụng gia tốc kế để xác định nhiều hành vi của bò hiệu suất và liên quan đến phúc lợi, một số trong đó sẽ được thảo luận tiếp theo. Mặc dù không được sử dụng trong bất kỳ chương thử nghiệm nào ở đây, một cuộc thảo luận về tiện ích của gia tốc kế rất hữu ích vì việc sử dụng chúng trong nghiên cứu chăn nuôi đã tăng lên đáng kể kể từ năm 2009 và thực sự là kể từ khi bắt đầu cuộc ứng cử này.

1.3.1.1 Phân loại hành vi nằm và đứng

Để tự động nhận biết hành vi của bò là nằm hay đứng ta có thể sử dụng một gia tốc kế gắn trên chân bò. Nó thường được đặt ở vị trí sao cho trục X song song với mặt đất. Để phân tích hành vi, trục Y các giá trị được sử dụng để xác nhận các sự kiện nằm hoặc đứng và trục Z xác nhận nằm ngang (trái hoặc phải). Xác định số lần nằm và đứng từng con khoảng thời gian được thực hiện bởi bò đã được sử dụng rộng rãi để đánh giá mối liên hệ giữa chúng

khập khiễng và chấn thương ở chân và điều này thông tin có thể hữu ích cho việc đánh giá đàn và cũng cung cấp cho nông dân những thông tin tư vấn cải tiến hệ thống. Ví dụ, Westin và cộng sự thấy rằng khập khiễng bò nằm dài hơn 0,6 giờ/ngày so với bò không bị khập khiễng và vết thương ở chân liên quan đến thời gian nói dối ngắn hơn. Tần số nằm và đứng cũng đã được sử dụng là chỉ số sinh bê ở bò sữa. Các khoảng thời gian ghi nhật ký tối ưu cho những hành vi này trước đây đã được xác định là 30 giây với điều kiện là những lần nói dối rất ngắn, có thể mắc sai lầm đã bị xóa khỏi tập dữ liệu. Tuy nhiên, phần lớn các ấn phẩm đã sử dụng một khoảng thời gian lấy mẫu phút với hiệu suất tương đương với nghiên cứu đã sử dụng nhiều hơn khoảng thời gian lấy mẫu thường xuyên. Hầu hết các nghiên cứu sử dụng giá trị lực g được tạo ra bởi gia tốc kế với việc bổ sung các hằng số để tạo điều kiện xử lý dữ liệu.

1.3.1.2 Phân loại các hành vi quan trọng về mặt sinh học khác bằng gia tốc kế

Ngoài tư thế nằm và đứng, một số tác giả đã có thể thực hiện thành công xác định hành vi kiếm ăn và nhai lại của bò bằng thiết bị gắn trên đầu gia tốc kế gắn trên tai và gia tốc kế gắn trên cổ [21-27]. Tần số lấy mẫu cao hơn thường cần thiết cho hoạt động kiếm ăn và nhai lại. Cần phân biệt giữa sự thường xuyên của hành vi và hành vi có thể gây ra các dấu hiệu có thể thay đổi. Ví dụ, Nielsen đã sử dụng dây buộc gia tốc kế được gắn ở tần số lấy mẫu là 5 giây để thu thập dữ liệu về chặn thả và các sự kiện không được chặn thả từ gia súc. Sau đó, một thủ tục phân biệt tuyến tính đã được sử dụng để phân biệt chặn thả với không chặn thả và độ nhạy và độ đặc hiệu là 84% và 80% đã đạt được tương ứng. Tương tự, Watanabe và cộng sự (2008) lấy mẫu dữ liệu gia tốc kế trong khoảng thời gian 1 giây và sử dụng quy trình phân biệt để phân loại hoạt động ăn uống, nhai lại và nghỉ ngơi bằng cách sử dụng các tính năng được tính toán từ dữ liệu gia tốc thô

(giá trị trung bình, phương sai và hệ số biến thiên nghịch đảo). Tỷ lệ đúng phân loại có thể thay đổi theo các trục và tính năng được sử dụng nhưng hoạt động tốt nhất kết hợp đạt được phân loại chính xác >90%.

Sử dụng cổ và gắn trên đầu gia tốc kế, bộ phân loại ngưỡng đã được phát triển sử dụng các giá trị gia tốc. Chiến lược này tương đối đơn giản để thực hiện so với các thủ tục tính toán chuyên sâu hơn đòi hỏi phải xử lý trước dữ liệu rộng rãi, phát triển và sàng lọc thuật toán. Các quan sát thực tế về mặt hành vi sinh học liên quan đến nhau. Các dữ liệu được làm thủ công có thể là đầu vào giúp phân tích có được mô hình mẫu phục vụ cho các hoạt động phân loại đằng sau. Ngưỡng sau đó có thể được gán cho dữ liệu tăng tốc thể hiện ranh giới giữa từng hành vi và phân tích được thực hiện dựa trên khả năng phân biệt đối xử của các ngưỡng này đối với dữ liệu mới. Những cái này Các phương pháp đã được chứng minh là có khả năng phân biệt tốt giữa các hành vi như cho ăn và đứng, cũng như chần thả và nhai lại. Những kỹ thuật này có chung điểm chung với điều đó được thiết lập để phân biệt hành vi nằm và đứng bằng cách sử dụng chân gia tốc kế gắn trên. Cần nhiều hơn nữa để kiểm tra xem báo cáo phân loại thành công có đúng với dữ liệu được thu thập trong thời gian dài hơn không các giai đoạn mà chiều cao, mật độ đồng cỏ và sự thay đổi của động vật mang đến ảnh hưởng tới kết quả phân tích. Từ đó, cần thiết đảm bảo rằng tính chính xác của việc cho ăn và thời gian nhai lại là những thông tin hữu ích được sử dụng làm dữ liệu tham khảo cho các phân tích PLF. Hơn nữa, vì khoảng thời gian lấy mẫu khác nhau trong các nghiên cứu này (phạm vi = 4 Hz – 1 mẫu/30 giây), việc xem xét dữ liệu sẽ có giá trị để đo lường thành công của các nghiên cứu về khoảng thời gian lấy mẫu và kỹ thuật được thực hiện.

1.3.1.3 Tần số lấy mẫu gia tốc kế

Tần số lấy mẫu được lựa chọn cho dữ liệu gia tốc ở mức độ lớn phụ thuộc vào các hành vi cần được xác định, việc phân tích được thực hiện

sau đó, dung lượng của thiết bị hoặc hệ thống để lưu trữ dữ liệu và có lẽ cả pin hiệu quả tuổi thọ của thiết bị. Lợi ích của việc tối đa hóa tần suất lấy mẫu là tập dữ liệu có thể được lấy mẫu phụ để kiểm tra hiệu suất thuật toán trên một loạt các khoảng thời gian lấy mẫu.

Gần đây, tần số lấy mẫu lên tới 50 Hz (50 mẫu/s) đã được báo cáo và các nghiên cứu thường báo cáo tần suất lấy mẫu trong khoảng 1-20 Hz. Ảnh hưởng của khoảng thời gian lấy mẫu đến hiệu suất phân loại là không rõ ràng vì không có khoảng thời gian lấy mẫu tiêu chuẩn hoặc quy trình phân tích cho những dữ liệu đó.

Tuy nhiên, nhìn chung tần số lấy mẫu càng cao thì việc phân loại càng tốt hiệu suất đối với các hành vi xảy ra thường xuyên và trong thời gian dài. Hơn nữa, có nguy cơ là khoảng thời gian lấy mẫu không thường xuyên có thể dẫn đến mất dữ liệu thông tin hành vi và trình bày sai về tần suất và thời gian của hành vi do động vật đầu mỗi thực hiện. Công việc là cần thiết để đối chiếu hiệu suất của các chiến lược phân tích khác nhau (ví dụ: tính năng, thuật toán và phân tích thống kê được sử dụng) về phạm vi khoảng thời gian lấy mẫu được báo cáo trong dữ liệu để xác định các quy trình tốt nhất để phân loại hành vi của bò sữa bằng cách sử dụng gia tốc kế. Hơn nữa, hiệu suất của gia tốc kế liên quan đến vị trí trên động vật là cần thiết vì có điều này ảnh hưởng nhiều đến chất lượng của dữ liệu được thu thập cũng như khả năng áp dụng tại các trang trại.

1.3.2 GPS để phân loại vị trí và hành vi

GPS đã được sử dụng rộng rãi trong nghiên cứu động vật trong nhiều năm. Việc sử dụng GPS chủ yếu là để đánh giá vị trí không gian và sự di chuyển của động vật trong nghiên cứu về loài phân phối và tương tác. Một số lĩnh vực quan tâm đáng chú ý bao gồm nghiên cứu ảnh hưởng của khí hậu và bệnh khi di chuyển. Việc sử dụng động vật máy thu vệ tinh borne bắt đầu phát triển vào những năm 1990 phát triển từ việc sử dụng

chúng như các cảm biến độc lập cho đến ngày nay, nơi GPS thường được sử dụng kết hợp với các cảm biến khác như gia tốc kế, từ kế và con quay hồi chuyển để phân tích đa chiều về chuyển động của động vật.

1.3.2.1 Cấu hình GPS trong PLF

Giống như các hệ thống cảm biến khác, GPS có thể được cấu hình nhằm thu thập thông tin vật nuôi theo các mốc thời gian đã đặt trước. Tốc độ lấy mẫu dữ liệu thường được xác định bởi mục tiêu của nghiên cứu, về mặt lịch sử, thường tập trung vào việc biết đâu là con vật là gì, nó tương tác với cái gì và trong bao lâu. Gắn liền với vị trí địa lý hệ thống thông tin, dữ liệu GPS có thể được tăng cường để cung cấp nhiều chiều hơn cho phép các nhà phân tích tiếp cận sâu hơn vào hành vi của động vật đầu mối. Vị trí không gian, tốc độ di chuyển và độ cao là những ví dụ về các biến thường được thu thập có thể được đặt để đăng ký ở các khoảng thời gian cố định từ > 1 Hz đến một lần hàng giờ.

GPS cũng có khả năng thu thập dữ liệu theo các khoảng thời gian được đặt sẵn. Ví dụ, khi chuyển động vượt quá một mức nhất định ngưỡng xác định trước. Phương pháp thứ hai này có thể giúp tiết kiệm đáng kể pin mạnh mẽ hơn và do đó thời gian quan sát dài hơn. Các bản sửa lỗi có thể thay đổi trong dữ liệu GPS có thể cũng xảy ra khi các động vật tiêu điểm di chuyển vào và ra khỏi phạm vi vệ tinh và là một chủ đề thảo luận quan trọng trong dữ liệu sinh thái học về phong trào vì dữ liệu có thể đưa ra một thách thức bổ sung trong việc suy luận hành vi của động vật. Ít nhất là đối với động vật trên cạn, việc sử dụng kích hoạt (được xác định trước) đặc biệt đòi hỏi các máy thu GPS phải đã kiểm tra mức độ sửa lỗi vị trí của chúng (mặc dù điều này cũng cần thiết với tỷ lệ mẫu thông thường). Điều này là cần thiết để tối đa hóa cơ hội phân loại chính xác một trường hợp chuyển động từ một trường hợp không chuyển động từ động vật tiêu điểm.

1.3.2.2 Sửa lỗi GPS

Trong khi các máy thu GPS có thể thu thập thông tin chính xác dựa trên vị trí, một số vấn đề tồn tại có thể dẫn đến sự không nhất quán và đôi khi sự thay đổi lớn về cố định vị trí tuyệt đối của từng máy thu ở mỗi mẫu thứ n khi được đo theo điểm đã khảo sát. Những vấn đề này đã được mô tả chi tiết bởi một số tác giả chuẩn bị cho nghiên cứu về chuyển động của động vật và phải được tính đến nếu có ý nghĩa thông tin cần được thu thập.

Một số vấn đề chính có thể gây ra sai lệch vị trí và có sự thay đổi lớn các bản sửa lỗi bao gồm lỗi vệ tinh và máy thu, hiệu ứng khí quyển và lỗi đa đường gây ra bởi các vật thể lớn như các tòa nhà. Các phương pháp sàng lọc được sử dụng trước đây để loại bỏ các bản sửa lỗi GPS sai bằng nhiều chiến lược khác nhau. Nếu các sai số không được hiệu chỉnh có thể dẫn đến việc ước tính sai về khoảng cách di chuyển của vật nuôi và độ chính xác của kết quả được xác định cho chúng. Tốc độ lấy mẫu dữ liệu cũng có thể ảnh hưởng đến những ước tính này và việc lựa chọn điều này sẽ phụ thuộc vào loài được quan sát và độ phân giải cần thiết cho giám sát.

Ví dụ, tốc độ lấy mẫu là 10 giây trong một khu vực rộng 100 m² dẫn đến kết quả ước tính sai số dự đoán vị trí là 1% (Swain và cộng sự, 2008). Nếu quãng đường đi được là một nghiên cứu khách quan thì tốc độ lấy mẫu thường xuyên hơn cũng sẽ cung cấp kết quả chính xác hơn ước lượng khoảng cách di chuyển. Điều này là do các quỹ đạo đường thẳng giữa các mẫu sẽ dẫn đến sự không chắc chắn hơn về hoạt động của động vật [24].

Máy thu GPS thường được hỗ trợ với dữ liệu về dự kiến của chúng mức độ hiệu suất, các thước đo hiệu suất đáng chú ý nhất trong bối cảnh động vật nghiên cứu chuyển động có độ chính xác tuyệt đối theo chiều ngang

và chiều dọc, mặc dù những nghiên cứu khác tồn tại. Máy thu GPS dùng cho nghiên cứu mục đích cần được kiểm tra tính đồng nhất.

1.3.3 Các cảm biến và công cụ khác để phân loại hành vi

1.3.3.1 Cảm biến áp suất đầu nòng

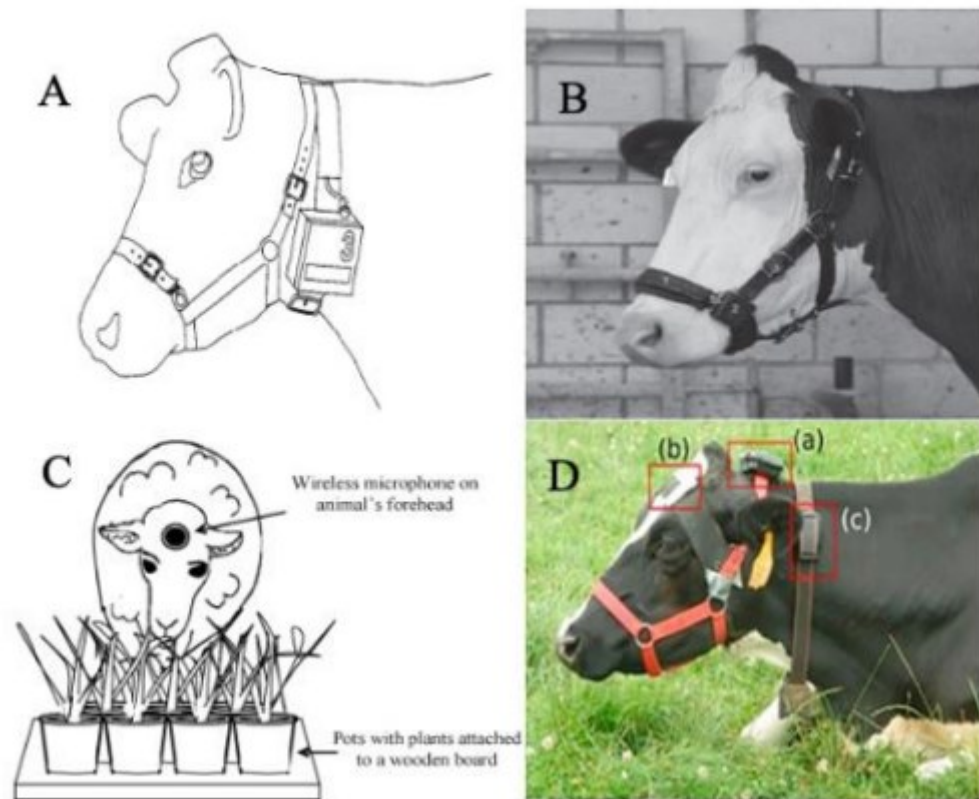
a) Máy ghi IGER

Trong khi gia tốc kế và GPS chiếm phần lớn nghiên cứu PLF hiện nay ở phân loại hành vi của gia súc, một số loại cảm biến khác cũng được đưa vào sử dụng và có thể sẽ xuất hiện nhiều hơn trong nghiên cứu trong tương lai. Cảm biến áp suất là một ví dụ như vậy. Một trong những phương pháp đầu tiên được sử dụng ở động vật nhai lại (cừu) được phát triển bởi Penning (1983) đã ghi lại chuyển động của hàm từ sự co giãn của một dây đeo mũi. Họ có thể phân loại việc chăn thả, nhai lại và chạy không tải mỗi phút nhưng thành công trong các đợt hành vi ngắn hơn là thấp. Phiên bản cập nhật của cảm biến này (máy ghi IGER) được phát triển bởi Rutter (1997) sử dụng máy vi tính hệ thống ghi âm chứ không phải là máy ghi băng cassette (Hình 4A). Xử lý hậu kỳ của dữ liệu sau đó được thực hiện bằng phần mềm riêng biệt. Mức độ tổng thể của sự thống nhất giữa các quan sát thủ công và các trường hợp được hệ thống phân loại là 91% cho việc ăn, nhai lại và các hành vi 'khác'.

b) Đồng hồ Rumi [25]

Gần đây hơn, một cảm biến áp suất dây đeo mũi có cấu hình thấp; Rumiwatch (Hình 4B), đã được phát triển để đo lường các thuộc tính liên quan đến việc cho ăn bò sữa (ví dụ như ăn nhai và cắn trước). Một lần nữa, công nghệ này sử dụng phần mềm riêng biệt để xử lý hậu kỳ. Ấn bản gần đây nhất để kiểm tra hiệu suất của hệ thống này nhận thấy rằng nó dự đoán số lượng về việc nhai lại, nhai lại và nhai lại rất tốt (dự đoán sai số < 0,10). Tuy

nhiên, các lỗi dự đoán đối với vết cắn trước và thời gian ở khả năng nắm bắt và ăn uống cao hơn (sai số dự đoán > 0,10).



Hình 4: Ví dụ về cảm biến ghi âm và áp suất được sử dụng để phân loại hành vi của động vật nhai lại. A = máy ghi IGER (áp suất); B = Rumiwatch (áp suất); C = micro gắn trên đầu; D = micrô gắn trên đầu và thiết bị ghi âm (b và a tương ứng).

1.3.3.2 Cảm biến âm thanh

Một phương pháp được thiết lập tốt khác để phân loại hành vi (đặc biệt là cho ăn) là giám sát âm thanh. Ungar và Rutter nhận thấy rằng việc sử dụng micro có thể thu được các kết quả phân loại tốt. Milone và cộng sự (2009) đã sử dụng micro không dây được gắn vào đầu cừu (Hình 4C) và mô hình Markov ẩn để tự động phân đoạn và phân loại các sự kiện nhai được tạo ra từ việc tiêu thụ hai loại khác nhau thức ăn thô xanh. Nhìn chung, các sự kiện nhai được xác định chính xác 82%. Hơn gần đây,

Vanrrell và cộng sự. (2018) đã sử dụng micro và thiết bị ghi âm gắn trên đầu (Hình 4D) và cách tiếp cận phân chia và phân loại hai giai đoạn của quá trình tìm kiếm thức ăn hành vi cư xử. Tính quy luật của các sự kiện được phát hiện bằng sự tự tương quan của âm thanh chữ ký được sử dụng để xác định ranh giới thời gian của các sự kiện. Thuộc tính cho ăn sau đó được phân loại bằng cách sử dụng các đặc điểm riêng cho từng loại hành vi như số lần tạm dừng (cao hơn trong thời gian nhai lại so với chăn thả). Số đo F là 0,89 và 0,93 đã đạt được tương ứng đối với chăn thả nhai lại và chăn thả gia súc.

1.3.3.3 Hệ thống cảm biến định vị cục bộ

Một kỹ thuật mới để đo vị trí của bò trong môi trường nuôi nhốt là thu hút được sự quan tâm vì cái nhìn sâu sắc mà nó có thể cung cấp về không gian-thời gian hành vi của bò. Cảm biến định vị cục bộ (LPS) về nguyên tắc rất giống với GPS. Nó có khả năng thu thập các thông tin về vị trí tuyệt đối của bò và cung cấp các mẫu thường xuyên để xác định khả năng di chuyển của vật nuôi trong chuồng.

Hệ thống bao gồm hai thành phần chính; máy thu tĩnh được đặt trong cảm biến chuồng trại và di động có thể được trang bị cho những con bò chuyển tiếp vị trí của chúng tại khoảng thời gian định trước cho các máy thu hồi tĩnh. Vị trí của bò trong chuồng là được tính toán bằng phép đo tam giác với mỗi máy thu tĩnh, cung cấp khả năng định vị theo tọa độ định dạng (x, y). Độ chính xác vị trí của cảm biến di động có thể được đo bằng cách sử dụng các phương pháp tương tự như các phương pháp được báo cáo trong Bảng 1 và những phân tích này được thực hiện bởi Barker và cộng sự (2018). Họ nhận thấy rằng độ chính xác về vị trí của hệ thống LPS không đồng nhất theo tư thế đứng hoặc nằm của bò. Để giảm thiểu một số vấn đề có thể phát sinh do dữ liệu vị trí đo được không chính xác, một chuyển động cửa sổ trung bình đã được áp dụng cho dữ liệu thô để loại bỏ các giá trị ngoại lệ.

Các hệ thống này có thể được kết hợp với các cảm biến chính xác khác để xác định chìa khóa hành vi trong toàn bộ môi trường ở. Kết hợp với hệ thống LPS, Barker và cộng sự (2018) đã sử dụng thuật toán cây quyết định để phân loại hành vi của các nhà ở bò sữa được trang bị máy đo gia tốc gắn trên cổ. Hệ thống LPS đã được sử dụng để bổ sung cho quy trình phân loại nhằm theo dõi hành vi ăn uống của bò khập khiễng và bò không khập khiễng. Diosdado và cộng sự. (2018) cũng sử dụng hệ thống tương tự để theo dõi mô hình sử dụng không gian của bò sữa trong chuồng. LPS có thể là một công cụ rất có giá trị để giám sát gia súc trong môi trường chuồng trại vì nó có thể cung cấp thông tin về hành vi trước đây không thể truy cập được bằng các cảm biến chính xác khác như GPS.

Bảng 1: Số liệu về độ chính xác theo chiều ngang được sử dụng để đo hiệu suất GPS theo định nghĩa của Viện Điều hướng (1997)

Measure	Formula	Probability	Definition
CEP (Circular error probability)	$0.62\sigma_y + 0.56\sigma_x$	50%	Radius of a circle centred at the true position containing 50% of the data points.
1 Sigma	$\sqrt{\sigma_x^2 + \sigma_y^2}$	68%	Square root of the average of the squared horizontal position errors.
R95	$R(0.62\sigma_y + 0.56\sigma_x)$	95%	Radius of a circle centred at the true position containing 95% of the data points. R = 2.08.
2 Sigma	$2\sqrt{\sigma_x^2 + \sigma_y^2}$	98%	Twice the value of 1 Sigma.

Chương 2: Một số phương pháp học máy

2.1. Giới thiệu

Khả năng thu thập dữ liệu lớn và thường phức tạp ở các trang trại hiện đang cung cấp tiềm năng ứng dụng kỹ thuật ML trong ngành sữa. Cơ hội khai thác dữ liệu đó để lấy thông tin mà trước đây không thể tiếp cận được là rất có khả năng giúp phát triển các công cụ ML hỗ trợ tại các trang trại. Dữ liệu này phần lớn xuất phát từ những cải tiến trong việc thu thập dữ liệu công nghệ. Tuy nhiên, cơ sở dữ liệu đã được thiết lập có thể bị khai thác bởi những kỹ thuật để tiết lộ những hiểu biết mới, ví dụ như về kiểu hình bệnh tật.

Ngành chăn nuôi bò sữa không phải là ngành duy nhất của ngành chăn nuôi được hưởng lợi từ những phương pháp này. Công việc đã được thực hiện để tự động xác định hành vi của cừu sử dụng gia tốc kế và thuật toán tổng hợp, phân loại thuộc tính chất lượng thịt bò sử dụng dữ liệu khối phổ và máy vectơ hỗ trợ thuật toán và xác định bệnh lợn bằng cách sử dụng môi trường dữ liệu cảm biến và thuật toán mạng thần kinh. Việc khai thác dữ liệu lớn ngày càng được cải thiện do sự xuất hiện của nền tảng điện toán đám mây mạnh mẽ cho phép phân tích dữ liệu thông lượng cao. Hệ thống hỗ trợ ra quyết định và quản lý bằng máy tính tại các trang trại bò sữa có đã được thảo luận trong nhiều năm.

2.2. Bộ tính năng cho máy học

Các quy trình ML thường liên quan đến việc đào tạo một mô hình trên một bộ tính năng được phát triển từ một tập dữ liệu thô được lấy mẫu. Các tính năng được trích xuất để tăng tính chiều của tập dữ liệu và có thể có nhiều dạng (ví dụ: nhị phân hoặc số). Bộ tính năng được sử dụng như một tập hợp các tham số đầu vào mà thuật toán ML có thể học hỏi. Trong các nghiên cứu sử dụng gia tốc kế để phân loại hành vi của động vật đầu mối, có một số

các tính năng đã được thiết lập có thể được trích xuất phổ biến trong nghiên cứu. Các tham số được quan tâm trong nghiên cứu như: trục gia tốc, giá trị trung bình, độ lệch, độ nhọn, độ lệch chuẩn, giá trị tối đa và tối thiểu được nghiên cứu trong chăn nuôi. Các tính năng khác được cũng được tính toán thường xuyên bao gồm gia tốc vector và gia tốc động tổng thể của vật thể. Chúng thường được sử dụng xác định chuyển động của động vật và tính toán mức tiêu hao năng lượng. Chúng đặc biệt hữu ích trong việc phân biệt giữa các hành vi có chuyển động thấp và cao.

Để tính ODBA và VeDBA, thành phần tĩnh (ST) của gia tốc và cần có chuyển động tổng thể của động vật (gia tốc động của cơ thể; DBA). Đến tính ST, giá trị trung bình đang chạy của kích thước cửa sổ cố định (ví dụ 5 giây) được sử dụng để làm mịn từng trục gia tốc (X, Y và Z). Tiếp theo, DBA được tính bằng cách trừ ST khỏi giá trị thô dữ liệu gia tốc. Từ đó, ODBA và VeDBA được tính như sau:

$$ODBA = DBA_X + DBA_Y + DBA_Z \quad (4)$$

$$VeDBA = \sqrt{DBA_X^2 + DBA_Y^2 + DBA_Z^2} \quad (5)$$

Các tính năng khác đôi khi được báo cáo cho dữ liệu tăng tốc và tổng hợp nghiên cứu và các tính năng tăng tốc đã chọn của họ. Khi các mô hình được đánh giá trên tập huấn luyện, chúng sẽ được áp dụng cho tập kiểm tra dữ liệu chưa từng thấy trước đó. Bộ kiểm tra có thể chỉ ra hiệu suất của mô hình trong các nhiệm vụ phân loại hoặc dự đoán trong tương lai, nhưng hiệu suất có thể bị suy giảm nếu tập kiểm tra chứa các trường hợp chưa từng thấy trong quá trình huấn luyện trước đó.

Lý do đằng sau ML là các mô hình có thể được cải thiện theo thời gian khi có nhiều dữ liệu hơn đại diện của không gian vấn đề được tích lũy. Các mô hình sau đó có thể được đo lường trên hiệu quả hoạt động bằng nhiều biện pháp khác nhau. Một phần, hiệu suất của bất kỳ mô hình

nào cũng có thể bị ảnh hưởng bởi các phương tiện lấy mẫu dữ liệu thô (ví dụ: tự động, thủ công), chất lượng tổng thể của nó (ví dụ tần số lấy mẫu, độ chính xác mẫu) và tính đa dạng của bộ đặc trưng được trích xuất. Khi các cảm biến như GPS được sử dụng để thu thập dữ liệu thô, tính năng đa dạng thường được lớn hơn, một phần vì bộ tính năng cho dữ liệu tăng tốc được phát triển và thử nghiệm tốt và cũng bởi vì dấu hiệu dữ liệu GPS về cơ bản khác với dấu hiệu gia tốc dữ liệu. Ví dụ về các bộ tính năng được trích xuất từ dữ liệu GPS. Tốc độ giữa GPS liên tiếp ví dụ: các bản sửa lỗi có thể được khai thác để trích xuất một số tính năng cơ bản tương tự như được trích xuất từ dữ liệu gia tốc như tốc độ trung bình, tốc độ tối đa và tối thiểu. Một tập hợp thêm các tính năng duy nhất cho GPS có thể được trích xuất từ tọa độ vị trí như như độ góc của chuyển động và hướng đi của con vật tiêu điểm.

Điều này không có nghĩa là không thể rút ra được nhiều đặc điểm đa dạng từ dữ liệu gia tốc. Tuy nhiên, mục tiêu là giảm thiểu độ phức tạp của kết quả mô hình để giảm thời gian tính toán cần thiết để xử lý dữ liệu thô trong tương lai. Điều này có thể đạt được bằng cách giảm kích thước của tập tính năng sao cho thuật toán học chỉ xem xét các tính năng có ý nghĩa lớn phục vụ cho phân loại hoặc dự đoán. Quá trình này được gọi là lựa chọn tính năng hoặc giảm kích thước.

2.3.Học máy có giám sát và không giám sát

Nhiệm vụ ML có thể được phân loại là được giám sát hoặc không được giám sát. Giám sát việc học yêu cầu các thuật toán ML được cung cấp một tập hợp các trường hợp đào tạo được gắn nhãn với biến kết quả tương ứng hoặc nhãn lớp. Thuật toán sau đó ánh xạ sự phụ thuộc của từng nhãn lớp vào tập tính năng và các mô hình được học theo cách giảm thiểu lỗi trên toàn bộ tập dữ liệu. Từ đó, trên một tập hợp các trường hợp thử nghiệm để dự đoán nhãn lớp tương ứng trước khi đánh giá hiệu suất của mô hình.

Trong các nhiệm vụ học không giám sát, dữ liệu được cung cấp cho thuật toán ML là không được gắn nhãn và do đó, mục tiêu là tìm kiếm những điểm tương đồng trong dữ liệu và nhóm mỗi trường hợp mới dựa trên việc có tồn tại một điểm chung cụ thể hay không. Học không giám sát có thể dẫn đến việc khám phá các mô hình ẩn và có hai kỹ thuật chính là phân cụm và liên kết; cái sau được sử dụng để tìm kiếm các quy tắc giải thích phần lớn tập dữ liệu. Một ví dụ là việc sử dụng khai thác luật kết hợp trong việc liên kết kiểu hình bệnh gia súc với các triệu chứng của bệnh nâng cao khả năng xác định bệnh. Mặt khác trước đây đã được sử dụng để xác định các hành vi khác biệt của gia súc. Sử dụng các nhóm hành vi riêng biệt từ giai đoạn học không giám sát, đã tiến hành thử và phân loại các nhóm này sử dụng kỹ thuật ML được giám sát, do đó tận dụng khả năng không được giám sát phân tích cụm để giảm số lượng các loại hành vi dư thừa. Các chuyên gia đã sử dụng mạng nơ-ron tích chập sâu để trích xuất và tìm hiểu một tập hợp các tính năng phân biệt đối xử từ hình ảnh mõm của gia súc. Bộ tính năng này sau đó được sử dụng để phân loại từng gia súc và mục tiêu của hệ thống như vậy là giảm nguy cơ nhận dạng sai gia súc.

Công việc thử nghiệm ở đây chỉ sử dụng các kỹ thuật ML được giám sát. Cho dù, việc sử dụng hạn chế các phương pháp không giám sát trong khoa học về sữa, đặc biệt là đối với gia súc nhận dạng hành vi, một số ví dụ đang nổi lên. Các ví dụ gần đây bao gồm sử dụng thuật toán phân cụm; k-means để phân nhóm dữ liệu gia súc (gia súc kể gắn trên cổ) để phân loại các sự kiện nhiệt và hành vi của bò sữa trong chuồng [23-25].

2.4. Thuật toán học máy

2.4.1. Học máy dựa trên tập Luật

Hệ thống tự động thực hiện các nhiệm vụ phân tích dựa trên một tập luật đã được tạo dựng từ trước từ các chuyên gia. Tập luật này cần được tạo

thủ công và chỉ rõ được các đặc tính của dữ liệu cần phân loại hoặc thực thi một nhiệm vụ cụ thể.

Các luật quyết định được xây dựng trên một cụm dữ liệu có phân bố chặt với tỷ lệ cao các phần tử có cùng một lớp sẽ có chất lượng tốt. Việc xác định các cụm thường được giải quyết bằng các giải thuật phân cụm dữ liệu. Tuy nhiên các giải thuật phân cụm thường gom các phần tử gần nhau vào cùng 1 cụm nhưng các phần tử này lại có thể thuộc về nhiều lớp, đây là đặc tính chung của phương thức học không giám sát. Vì vậy, việc sử dụng các kỹ thuật Phân cụm bán giám sát có thể tìm ra các cụm dữ liệu của cùng một lớp để khởi tạo các luật quyết định tốt cho các giải thuật học luật quyết định.

Dữ liệu huấn luyện thường bị thiếu sót, với sự xuất hiện của dữ liệu bị mất. Thông thường, trong giai đoạn tiền xử lý dữ liệu, các dữ liệu bị mất sẽ được thay thế bằng các dữ liệu được ước lượng được tính bởi các phương pháp suy đoán dữ liệu. Tiêu chí của các phương pháp suy đoán dữ liệu thường không dựa vào các phương pháp khai thác dữ liệu, vì thế việc thay thế giá trị bị mất bằng một giá trị suy đoán sẽ làm giảm hiệu quả của giải thuật phân lớp khi học trên tập dữ liệu huấn luyện đã được làm sạch. Do tính đặc thù của bản thân của các giải thuật Học luật, việc xử lý trực tiếp dữ liệu bị mất ngay trong quá trình học sẽ mang lại hiệu quả cao hơn. Giải thuật CN2 đã được vận dụng cách tiếp cận này trong xử lý trực tiếp dữ liệu bị mất và độ chính xác của giải thuật CN2 đã được cải thiện. Các giải thuật học luật khác nên được khảo sát và vận dụng cách tiếp cận này để nâng cao hiệu quả phân lớp.

Đối với bài toán phân lớp, việc mất cân bằng trong phân bố các lớp gây khó khăn cho việc tìm ra bộ phân lớp hiệu quả. Đã có những nghiên cứu về việc luật phân lớp kết hợp trong việc xử lý bài toán dữ liệu mất cân bằng này. Một hướng khác là điều chỉnh trực tiếp trên các giải thuật học luật. Đối với cách tiếp cận “học từng luật với tập huấn luyện suy giảm dần”, phần “Giải thuật học 1 luật (Induce_One_RuleA)” có thể xem xét thay đổi độ ưu tiên cho

các ứng viên luật dựa trên phân bố của của các lớp. Đối với cách tiếp cận “Học từng luật trên tập huấn luyện bao gồm các đối tượng chưa được bao phủ”, phần “giải thuật học 1 luật (Induce_One_RuleB)” cũng có thể xem xét thay đổi độ ưu tiên cho các ứng viên luật dựa trên phân bố của của các lớp. Tuy nhiên, cần có các nghiên cứu cụ thể để nâng cao tính hiệu quả.

2.4.2. Phương pháp Naïve Bayes

Bộ phân lớp quan điểm Naïve Bayes được xây dựng dựa trên lý thuyết Bayes về xác suất có điều kiện để phân lớp quan điểm:

$$P(c | d) = P(c) \cdot \frac{P(d | c)}{P(d)} \quad (3.1)$$

Mục tiêu là tìm được phân lớp c sao cho $P(c|d)$ là lớn nhất hay xác suất của dữ liệu d thuộc lớp c là lớn nhất.

Ta có thể nhận thấy từ công thức trên $P(d)$ không đóng vai trò gì trong việc quyết định phân lớp $c \square P(c|d)$ lớn nhất $\Leftrightarrow P(c) \cdot P(d|c)$ lớn nhất.

Để có thể xấp xỉ giá trị của $P(d|c)$, thuật toán Naïve Bayes giả sử rằng: các vector đặc trưng f_i của một dữ liệu khi đã biết phân lớp là độc lập với nhau. Từ đó ta có công thức:

$$\begin{aligned} P(c|d) \max &= \operatorname{argmax} P(c) \cdot P(f_1, f_2, \dots, f_n | c) \\ \Leftrightarrow P(c|d) \max &= \operatorname{argmax} P(c) \cdot \prod_{1 \leq i \leq n} P(f_i | c) \\ \Leftrightarrow P(c|d) \max &= \operatorname{argmax} (\log P(c) + \sum_{1 \leq i \leq n} \log P(f_i | c)) \end{aligned} \quad (3.2)$$

Trong đó f là các vector đặc trưng cho dữ liệu d .

Khi tiến hành huấn luyện, thuật toán sử dụng phương pháp xấp xỉ hợp lý cực đại MLE (Maximum Likelihood Estimation) để xấp xỉ $P(c)$ và $P(f_i|c)$ cùng thuật toán làm mịn add-one (add-one smoothing). Ta có:

$$P(c) = \frac{N_c}{N} \quad (3.3)$$

Trong đó N_c là số văn bản được phân loại vào lớp c ; N là tổng số văn bản trong tập huấn luyện.

$$P(f_i | c) = \frac{N_{cf_i}}{\sum_{f \in F} N_{cf}} \quad (3.4)$$

Trong đó N_{cf_i} là số lần xuất hiện của vector đặc trưng i trong dữ liệu thuộc phân lớp c .

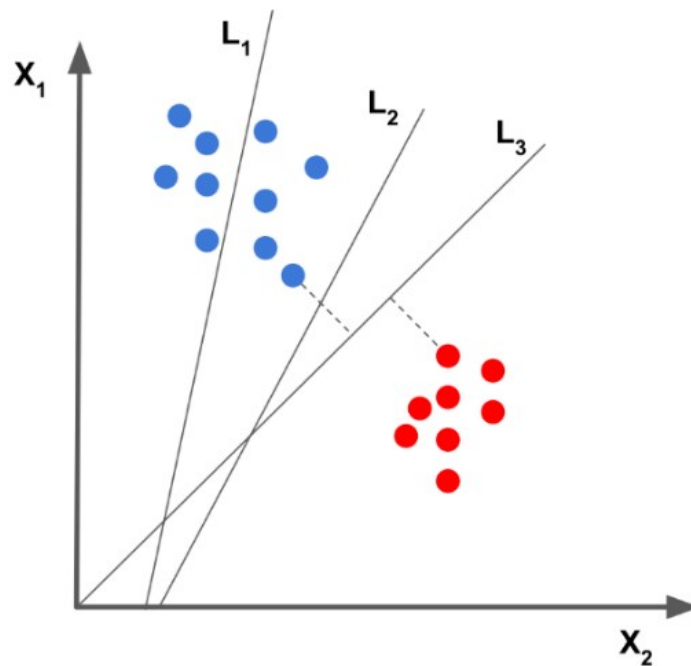
Trong nhiều bài toán còn phụ thuộc vào dữ liệu để lựa chọn các mô hình Naive Bayes. Bao gồm 3 mô hình được đưa ra dưới đây:

- a) **Gaussian:** Mô hình Gaussian giả định rằng các đối tượng địa lý tuân theo phân phối chuẩn. Điều này có nghĩa là nếu các bộ dự đoán nhận các giá trị liên tục thay vì rời rạc, thì mô hình giả định rằng các giá trị này được lấy mẫu từ phân phối Gaussian.
- b) **Multinomial:** Bộ phân lớp Naïve Bayes đa lớp được sử dụng khi dữ liệu được phân phối đa lớp. Nó chủ yếu được sử dụng cho các vấn đề phân lớp dữ liệu, nó có nghĩa là một dữ liệu cụ thể thuộc về danh mục nào như tích cực, tiêu cực, rất tích cực, tiêu cực hoặc trung tính. Trình phân lớp sử dụng tần suất từ cho các yếu tố dự đoán.
- c) **Bernoulli:** Bộ phân lớp Bernoulli hoạt động tương tự như bộ phân lớp Đa thức, nhưng các biến dự báo là các biến Booleans độc lập. Chẳng hạn như nếu một từ cụ thể có trong dữ liệu hay không. Mô hình này cũng nổi tiếng với các nhiệm vụ phân lớp dữ liệu.

2.4.3. Phương pháp Support Vector Machine (SVM)

Support vector machines (SVM) là phương pháp học có giám sát bao gồm phân tích dữ liệu và phát hiện mẫu, được sử dụng cho phân lớp và phân tích hồi quy. Thuật toán SVM được Vladimir Vapnik đề xuất vào năm 1995.

Cách dễ nhất để hiểu SVM là sử dụng một bài toán phân lớp nhị phân. Hai lớp được hiển thị bằng hai màu khác nhau. SVM tìm dòng tốt nhất phân tách hai lớp. Ta thấy dữ liệu được biểu diễn dưới dạng các chấm trên mặt phẳng 2D. Dữ liệu thuộc hai lớp khác nhau được biểu thị bằng màu sắc của các dấu chấm xanh và chấm đỏ. Đối với phân tích quan điểm, điều này sẽ là tích cực và tiêu cực. Một cách để học cách phân biệt giữa hai lớp là vẽ một đường phân chia không gian 2D thành hai phần. **Huấn luyện** hệ thống chỉ đơn giản là tìm dòng. Khi đã huấn luyện hệ thống (tức là đã tìm thấy đường thẳng), có thể biết liệu một điểm dữ liệu mới thuộc lớp màu xanh hay màu đỏ bằng cách chỉ cần kiểm tra xem nó nằm ở phía nào của đường thẳng.

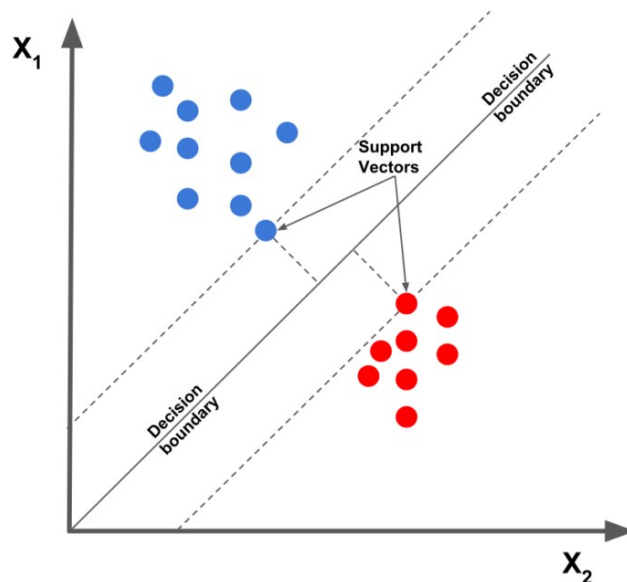


Hình 5: SVM tìm dòng tốt nhất phân tách hai lớp

Trong ví dụ hình 2.2 trên, rõ ràng là dòng L1 không phải là một lựa chọn tốt vì nó không tách biệt hai lớp. L2 và L3 đều tách biệt hai lớp, nhưng trực quan chúng ta biết L3 là lựa chọn tốt hơn L2 vì nó phân tách rõ ràng hơn hai lớp.

Ý tưởng chính của thuật toán này là cho trước một tập huấn luyện được biểu diễn trong không gian vector, trong đó mỗi dữ liệu là một điểm trong không gian n chiều và từ các dữ liệu huấn luyện ban đầu được gán nhãn sẽ

tìm ra một siêu phẳng phân lớp chính xác các dữ liệu



Hình 6: Ví dụ về siêu phẳng trong SVM

Chất lượng của siêu phẳng được quyết định bởi khoảng cách của điểm dữ liệu gần nhất của mỗi lớp đến mặt phẳng. Khoảng cách biên càng lớn thì mặt phẳng quyết định càng tốt, đồng thời việc phân lớp càng chính xác. Mục đích của thuật toán là tìm được khoảng cách biên lớn nhất để tạo ra kết quả phân lớp tốt

Trong ví dụ trên siêu phẳng tối ưu phân chia dữ liệu thành hai lớp màu xanh và màu đỏ. Các điểm gần nhất là các *vector* hỗ trợ được tô đậm. Hai bên của siêu phẳng là hai lề chứa các vector hỗ trợ – tức là các điểm dữ liệu gần siêu phẳng nhất. SVM thực chất là bài toán tối ưu, mục tiêu của thuật toán này là tìm được một không gian siêu phẳng khi Vector hỗ trợ có khoảng cách lớn nhất có thể từ ranh giới quyết định (tức là tách siêu phẳng) và hai lớp nằm trên các mặt khác nhau của siêu phẳng.

Xét một tập dữ liệu mẫu:

$$D = \{(x_1, y_1), \dots, (x_l, y_l)\}, x \in \mathbb{R}^n, y \in \{-1, 1\}$$

(3.5)

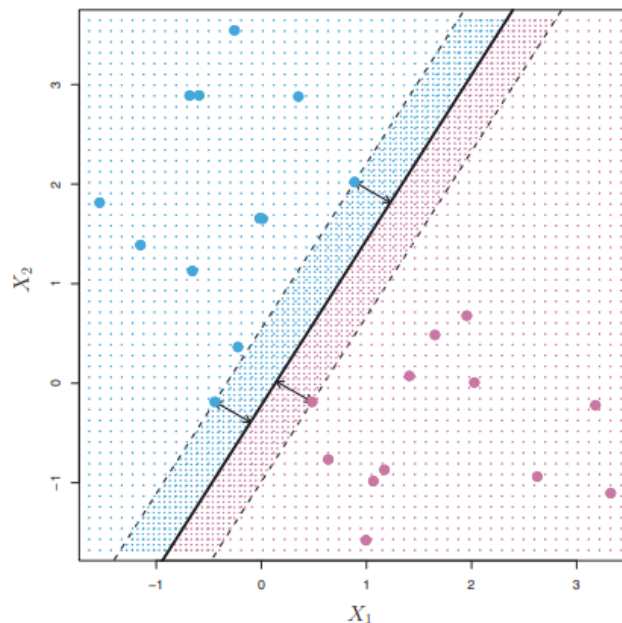
Trong đó x_i là một véc tơ đặc trưng hay một điểm (trong không gian n chiều $x \in \mathbb{R}^n$) biểu diễn tập mẫu d_i cặp (x_i, y_i) biểu diễn rằng với một vector

đặc trưng x_i thì được gán nhãn là y_i tương ứng trong đó $y \in \{-1, 1\}$ hay nói cách khác với tập mẫu d_i sẽ được gán nhãn cho trước là y_i . Ta có phương trình một siêu phẳng

$$f(x) = w \cdot x + b = 0 \quad (3.6)$$

Trong đó $w \cdot x$ là tích vô hướng giữa véc tơ x và véc tơ pháp tuyến $w \in \mathbb{R}^n$ được biểu diễn trong không gian n chiều, và $b \in \mathbb{R}$ là hệ số tự do.

Thực tế, các dữ liệu ban đầu có thể sinh ra vô số các siêu phẳng khác nhau để phân lớp dữ liệu tuy nhiên bài toán đặt ra là trong một không gian n chiều với các tập dữ liệu mẫu như vậy làm thế nào để tìm được một siêu phẳng luôn đảm bảo sự phân chia dữ liệu một cách tốt nhất, ta có thể hiểu một siêu phẳng tốt là một siêu phẳng mà khoảng cách từ các điểm dữ liệu được phân lớp gần nhất với siêu phẳng đó là lớn nhất. Phương trình chứa các điểm dữ liệu này được gọi là các lề, như vậy siêu phẳng tốt là siêu phẳng mà khoảng cách giữa nó và lề càng xa càng tốt.



Hình 7: Siêu phẳng phân chia lề xa nhất

2.4.4. Phương pháp Hồi quy Logistic (Logistic regression)

2.4.4.1. Mô hình Logistic

Đầu tiên, ta sẽ xem xét mô hình logistic như sau: Hãy xem xét một mô

hình có hai yếu tố dự đoán, x_1 và x_2 , và một biến phản hồi nhị phân Y , mà chúng tôi biểu thị $p = P(Y = 1)$. Giả định mối quan hệ tuyến tính giữa các biến dự đoán và tỷ lệ cược log của sự kiện $Y = 1$. Mối quan hệ tuyến tính này có thể được viết dưới dạng toán học sau (trong đó ℓ là tỷ lệ cược log, b là cơ số của logarit và β là các thông số của mô hình):

$$\ell = \ln \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \quad (3.7)$$

Có thể sử dụng lũy thừa để phục hồi tỷ lệ cược ta có:

$$\frac{p}{1-p} = b^{\beta_0 + \beta_1 x_1 + \beta_2 x_2} \quad (3.8)$$

Bằng thao tác đại số (chia cả tử và mẫu số cho $b^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}$) ta thu được xác suất $Y = 1$ là:

$$p = \frac{b^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}}{b^{\beta_0 + \beta_1 x_1 + \beta_2 x_2} + 1} = \frac{1}{1 + b^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}} = S_b(\beta_0 + \beta_1 x_1 + \beta_2 x_2) \quad (3.9)$$

Trong đó S_b là hàm sigmoid với cơ sở b . Công thức trên cho thấy rằng sau khi β_i được sửa, chúng ta có thể dễ dàng tính toán tỷ lệ cược $Y = 1$ cho một quan sát nhất định, hoặc xác suất $Y = 1$ cho một quan sát nhất định. Trường hợp sử dụng chính của mô hình logistic là đưa ra một quan sát (x_1, x_2) và ước tính xác suất p mà $Y = 1$. Trong hầu hết các ứng dụng, cơ số b của lôgarit thường được coi là e . Tuy nhiên, trong một số trường hợp, việc truyền đạt kết quả có thể dễ dàng hơn bằng cách làm việc trong cơ số 2 hoặc cơ số 10.

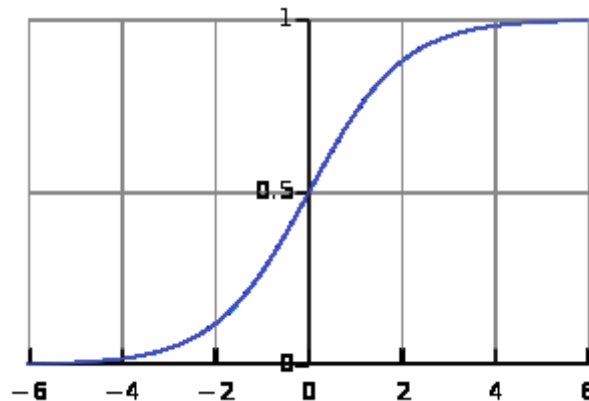
2.4.4.2. Hàm Logistic và các tỉ lệ

Một hàm Logistic chuẩn là một hàm Sigmoid, nhận bất kỳ đầu vào thực t và xuất ra giá trị từ 0 đến 1. Đối với logit, điều này có nghĩa là với bất kỳ tỉ

lệ logit đầu vào sẽ có đầu ra là xác suất. Hàm Logistic chuẩn được định nghĩa như sau:

$$\sigma(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}} \quad (3.11)$$

Minh họa, trường hợp t liên tục trong khoảng từ -6 đến 6 ta có hình 11.



Hình 8: Minh họa đồ thị của hàm Logistic khi t thuộc (-6,6)

Giả sử t là một hàm tuyến tính một biến x khi đó ta có:

$$t = \beta_0 + \beta_1 x \quad (3.12)$$

Và khi đó, ta có thể có hàm logistic tổng quát dạng $p: \mathbb{R} \rightarrow [0, 1]$ là:

$$p(x) = \sigma(t) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \quad (3.13)$$

Như trong mô hình Logistic, xác suất $p(x)$ được hiểu là phụ thuộc vào giá trị đầu ra của $Y = 1$ thay vì $Y = 0$. Rõ ràng, các biến có độ phản hồi Y_i là không giống nhau với các điểm dữ liệu X_i . Vì vậy ta có ma trận X và các tham số chia sẻ β .

Đặt $\beta_0 + \beta_1 x = W^T x$ khi đó ta sẽ có công thức tính xác suất cho hàm logistic tổng quát có dạng như sau:

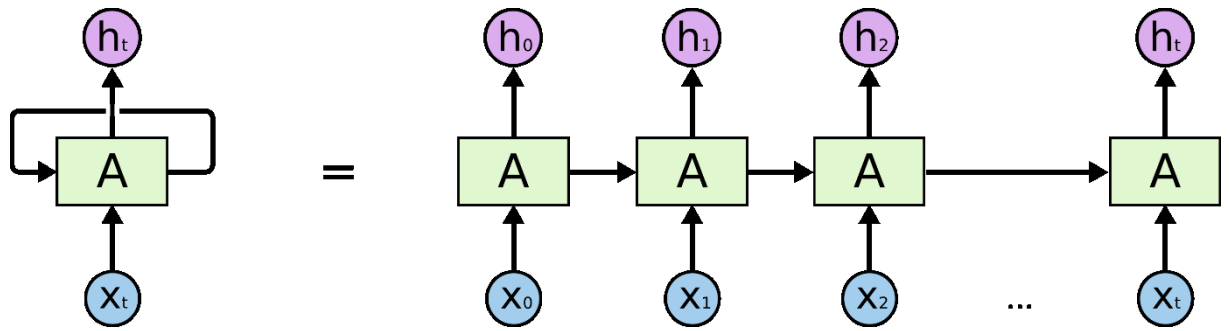
$$p(y = 1 | x : w) = \frac{1}{1 + e^{-w^T x}} \quad (3.14)$$

2.4.5. Phương pháp tiếp cận học sâu (Deep Learning)

Trong nhiều thập kỷ các phương pháp học máy được sử dụng trong các bài toán xử lý ngôn ngữ tự nhiên trong đó có bài toán phân tích quan điểm đã dựa trên các mô hình học máy truyền thống, được huấn luyện trên các đặc trưng với số chiều và độ thưa dữ liệu rất cao để khắc phục hạn chế này gần đây sử dụng các mạng nơ ron dựa trên các biểu diễn véc tơ dày đặc đã tạo ra kết quả vượt trội cho một số bài toán xử lý ngôn ngữ tự nhiên. Xu hướng này được tạo ra bởi sự thành công của việc nhúng các từ (Word Embedding) dưới dạng véc tơ và phương pháp học sâu cho phép học nhiều mức tự động các đặc trưng đại diện. Một số phương pháp học sâu được sử dụng cho bài toán xử lý ngôn ngữ tự nhiên cho kết quả đáng khích lệ như: Mạng nơ ron tích chập (Convolutional neural networks – CNN), mạng nơ ron hồi quy (Recurrent neural network – RNN) và mạng bộ nhớ dài ngắn (Long Short Term Memory network –LSTM).

2.4.5.1. Mạng Neural hồi quy RNN

Mỗi suy đoán hoặc quyết định của chúng ta được liên kết từ các thông tin phức tạp được tìm hiểu trước. Vấn đề càng phức tạp, lượng thông tin được tổng hợp để giải quyết càng lớn. Mạng RNN đưa ra các dự đoán dựa trên các thông tin đã có trong các vòng lặp chứa thông tin phía trước. Mạng RNN chứa các vòng lặp bên trong cho phép thông tin có thể lưu lại được



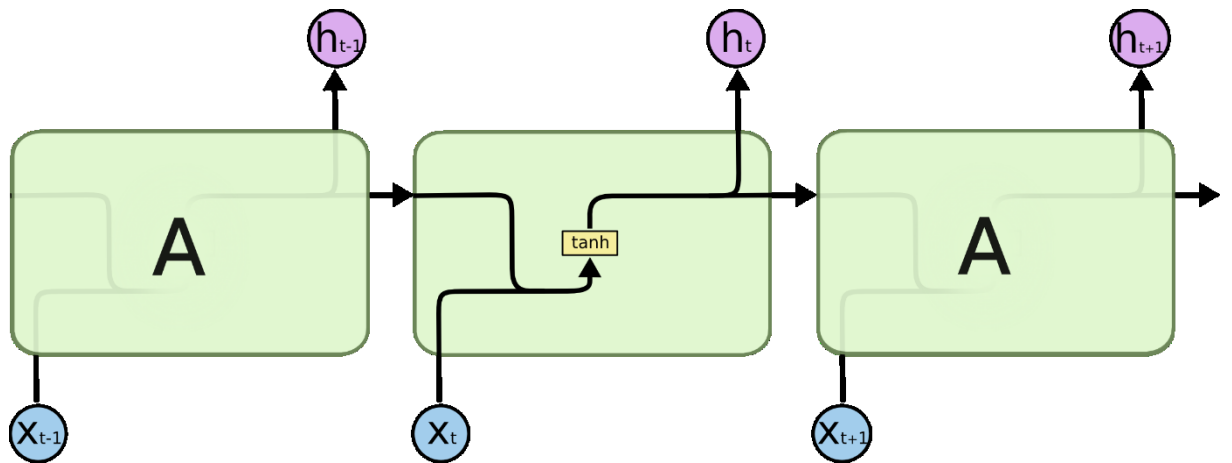
Hình 9: Mạng RNN có vòng lặp

Trong mô hình trên mô tả một đoạn của mạng RNN A với đầu vào là x_t và đầu ra là h_t . Một vòng lặp cho phép thông tin có thể được truyền từ bước này qua bước này qua bước khác của mạng nơ-ron. Có thể thấy kết quả của đầu ra h_1 nhận dữ liệu từ đầu vào x_1 và kết quả của nút phía trước hay nói cách khác đầu ra của nút này chính là đầu vào của nút ngay phía sau. Điều này đã mang lại sự thành công cho mạng RNN nhờ vào khả năng suy luận có cơ sở. Tuy nhiên, khả năng này tỏ ra không hiệu quả đối với các trường hợp phức tạp, đòi hỏi lượng thông tin cần tổng hợp cho suy đoán nhiều hơn.

Trong quá trình tiếp nhận những thông tin phản hồi, đôi khi chúng ta không hoàn toàn tiếp nhận hết tất cả lượng thông tin được gửi đến. Tuy nhiên, chúng ta có thể dự đoán nội dung thông tin một cách chính xác, điểm nổi bật của RNN chính là ý tưởng kết nối các thông tin phía trước để dự đoán cho hiện tại.

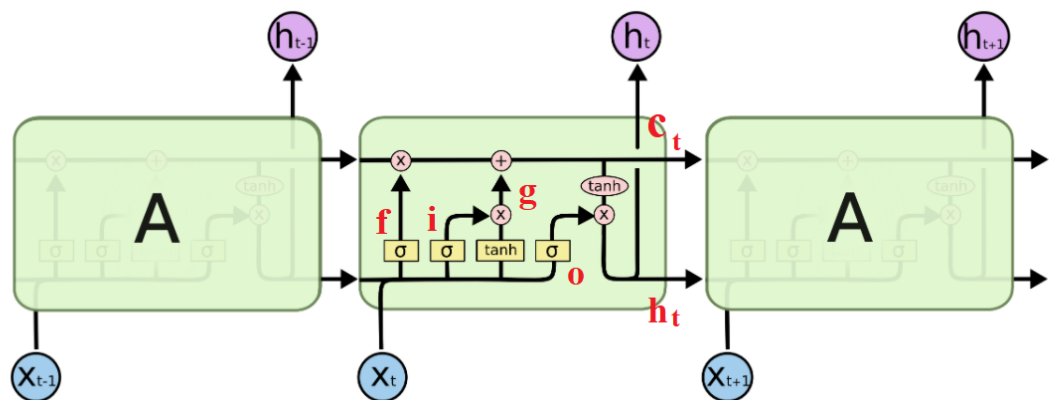
2.4.5.2. Mạng bộ nhớ thuật ngữ ngắn dài (Long Short-Term Memory)

Mạng LSTM là một dạng đặc biệt của RNN, nó có khả năng học được các phụ thuộc từ quá dài. LSTM được giới thiệu bởi Hochreiter & Schmidhuber (1997), nhưng được cải tiến và phổ biến rất. Chúng hoạt động cực kỳ hiệu quả trên nhiều bài toán khác nhau nên dần đã trở nên phổ biến như hiện nay.



Hình 10: Mạng RNN có vòng lặp

Kiến trúc dạng chuỗi các module của LSTM tương tự RNN tuy nhiên trong các module LSTM có cấu trúc phức tạp hơn, bao gồm 4 tầng tương tác với nhau theo một cấu trúc cụ thể với các cổng Forget Gate (f_t), Input Gate (i_t), Output Gate (o_t). Các cổng này quyết định việc thông tin nào sẽ được lưu trữ, xoá, chỉnh sửa và truyền đi



Hình 11: Mô đun lặp lại trong một LSTM

Trong sơ đồ trên, mỗi dòng mang một vector toàn bộ, từ đầu ra của một nút đến đầu vào của những nút khác. Các vòng tròn màu hồng đại diện cho

các phép toán theo chiều kim loại, như phép cộng vectơ, trong khi các hộp màu vàng là các lớp mạng nơ-ron đã học. Hợp nhất các dòng biểu thị sự móc nối, trong khi một dòng cho biết nội dung của nó đang được sao chép và các bản sao sẽ đến các vị trí khác nhau. Mô hình thiết kế của LSTM là một bảng mạch số, gồm các mạch logic và các phép toán logic

Chương 3: Phân loại hành vi bằng Random Forest và ảnh hưởng của độ lệch vị trí cảm biến đến hiệu suất phân loại hành vi

Thường rất khó để theo dõi hành vi của bò sữa vì việc chăn thả góp phần tỷ lệ đáng kể trong chế độ ăn của bò sữa và giảm sự tiếp xúc với bò. Chúng tôi trước đây đã phát triển một mô hình hành vi của bò sữa trên đồng cỏ đòi hỏi dữ liệu GPS chuyên đổi đến được thu thập từ gia súc để được phân chia thành các phân đoạn một khoảng thời gian cố định trước khi phân loại hành vi thành chăn thả, nghỉ ngơi hoặc đi bộ. Tuy nhiên, phân đoạn theo thời gian cố định gây ra vấn đề trong quá trình phân loại hành vi bởi vì ranh giới phân khúc có thể không được định vị chính xác tại điểm hành vi chuyển đổi, dẫn đến lỗi phân loại. Mục tiêu của công việc này là cố gắng khắc phục vấn đề này bằng cách điều chỉnh thống kê các dự đoán hành vi. Đây là đạt được bằng cách sử dụng mô hình Markov ẩn được đào tạo bằng 90 giờ dữ liệu được giám sát được thu thập từ một nhóm bò sữa được nghiên cứu trước đây.

Xác suất thống kê của các hành vi được bộ phân loại dự đoán là các hành vi thực sự (ẩn) được thể hiện bởi bò sữa và xác suất thay đổi các hành vi của chúng được phân loại nhằm đưa ra dự đoán. Sử dụng 51 giờ hành vi dữ liệu xác thực được gắn nhãn, hành vi chăn thả, nghỉ ngơi và đi lại của bò sữa Holstein (phân loại tổng thể độ chính xác = 0,85 (CI = 0,83 – 0,87) so với 0,94 (CI = 0,92 – 0,95)) cho riêng bộ phân loại và sau khi áp dụng mô hình Markov ẩn vào các hành vi được dự đoán tương ứng được thay đổi đáng kể. Để kiểm tra thêm các mô hình kết hợp của chúng tôi, cho ăn đậm, khỏe mạnh, cho con bú sớm (trung bình \pm SD; $43 \pm 20,9$ DIM) đồng cỏ nguyên thủy (n = 12) và đồng cỏ đa bội (n = 12) bò sữa dựa trên Holstein được gắn GPS GlobalSat DG-100 và được theo dõi cách ngày trong 10 ngày theo tỷ lệ thời gian dành cho việc chăn thả, nghỉ ngơi và đi dạo.

3.1. Vật liệu và phương pháp

3.1.1. Phân loại hành vi

Bước đầu tiên trong quá trình nhận dạng hành vi là tạo ra một loạt các nhãn hành vi được dự đoán (lượng khí thải) để đưa vào HMM. Việc phân loại hành vi được chia thành các giai đoạn:

- Phân tích chuyển động
- Phân đoạn và trích chọn đặc trưng
- Dự đoán hành vi.

3.1.2. Phân tích chuyển động

Phân tích chuyển động tính toán thông tin cơ bản cần thiết cho tính năng trích xuất như quãng đường bò đi được (m), tốc độ (m/s), gia tốc (m/s²) và góc quay (độ) giữa tọa độ liên kề và dữ liệu tốc độ. Trong này hoạt động, tốc độ lấy mẫu GPS được đặt thành 5 giây vì điều này được cho là mang lại kết quả tốt nhất cân bằng giữa năng lượng pin và độ phân giải chuyển động.

3.1.3. Phân đoạn và trích xuất đặc trưng

Sau khi phân tích chuyển động hoàn tất, toàn bộ tập dữ liệu được chia thành các đoạn có kích thước xác định trước. Ở đây, kích thước của mỗi phân đoạn được thiết lập sao cho mỗi phân đoạn chứa 32 chuyển động hoặc các trường hợp dữ liệu được thu thập liên tục (160 giây dữ liệu). Ví dụ, một máy thu GPS được thiết lập để thu thập tọa độ vị trí với tốc độ lấy mẫu là 5 giây sẽ thu thập khoảng 17.280 bản sửa lỗi vị trí trong 24 giờ mang lại 540 phân đoạn. Sau khi phân chia một số các tính năng có thể được trích xuất từ dữ liệu được lấy từ thông tin được thu thập trong phân tích chuyển động. Giai đoạn trích xuất đặc trưng thu thập thêm thông tin từ dữ liệu giúp xác định các lớp quyết định mà trong trường hợp này đang được chặn thả (trong đó bò đang đứng và đang tích cực ăn nguyên liệu thực vật hoặc đang tìm kiếm), nghỉ ngơi

(nơi bò đứng yên hoặc nằm và không biểu hiện thêm đặc điểm nào của chân thả) và đi bộ (nơi con bò di chuyển có mục đích thường với cường độ cao). Để phân loại, tổng cộng 13 đặc điểm được trích xuất từ mỗi phân đoạn trong tác phẩm này, ví dụ trong đó bao gồm tối thiểu, trung bình và tốc độ tối đa(m/s) và tốc độ thay đổi hướng trên mỗi đoạn.

3.1.4. Dự đoán hành vi

Cuối cùng, bằng cách sử dụng tất cả các đặc điểm được trích xuất, mỗi phân đoạn dữ liệu được phân loại thành chân thả, nghỉ ngơi hoặc đi dạo bằng cách sử dụng mô hình hành vi dựa trên quy tắc của bò. Đầu ra của quá trình phân loại trên thực tế là trình tự phát xạ sau đó được đưa vào HMM được mô tả trong phần tiếp theo.

3.2. Phân loại hành vi bằng Random Forest

Rừng ngẫu nhiên hoặc rừng quyết định ngẫu nhiên là một phương pháp học tập tổng hợp để phân loại, hồi quy và các nhiệm vụ khác hoạt động bằng cách xây dựng vô số cây quyết định tại thời điểm đào tạo. Đối với các nhiệm vụ phân loại, đầu ra của rừng ngẫu nhiên là lớp được hầu hết các cây chọn. Đối với các tác vụ hồi quy, giá trị dự đoán trung bình hoặc trung bình của từng cây sẽ được trả về. Rừng quyết định ngẫu nhiên điều chỉnh thói quen trang bị quá mức của cây quyết định cho tập huấn luyện của chúng. Rừng ngẫu nhiên thường hoạt động tốt hơn cây quyết định nhưng độ chính xác của chúng thấp hơn cây được tăng cường độ dốc. Tuy nhiên, đặc điểm dữ liệu có thể ảnh hưởng đến hiệu suất của chúng.

3.2.1. Thuật toán

a) Cây quyết định

Cây quyết định là một phương pháp phổ biến cho các nhiệm vụ học máy khác nhau. Cây học máy "tiến gần nhất đến việc đáp ứng các yêu cầu

phục vụ như một thủ tục có sẵn để khai thác dữ liệu". "Bởi vì nó bất biến khi chia tỷ lệ và nhiều phép biến đổi khác của các giá trị đối tượng, rất mạnh mẽ để bao gồm các đối tượng không liên quan và tạo ra các mô hình có thể kiểm tra được. Tuy nhiên, chúng hiếm khi chính xác". [28]

Đặc biệt, những cây được trồng rất sâu có xu hướng học các mẫu rất bất thường: chúng phù hợp quá mức với tập huấn luyện của chúng, tức là có độ lệch thấp nhưng phương sai rất cao. Rừng ngẫu nhiên là một cách lấy trung bình nhiều cây quyết định sâu, được huấn luyện trên các phần khác nhau của cùng một tập huấn luyện, với mục tiêu giảm phương sai. Điều này phải trả giá bằng sự gia tăng nhỏ về độ lệch và mất đi một số khả năng diễn giải, nhưng nhìn chung làm tăng đáng kể hiệu suất trong mô hình cuối cùng. [28] (trang 587-588).

b) Đóng gói

Thuật toán huấn luyện cho các khu rừng ngẫu nhiên áp dụng kỹ thuật chung về tổng hợp bootstrap hoặc đóng gói cho người học cây. Cho một tập huấn luyện $X = x_1, \dots, x_n$ với các phản hồi $Y = y_1, \dots, y_n$, đóng bao liên tục (B lần) chọn một mẫu ngẫu nhiên thay thế tập huấn luyện và ghép cây vào các mẫu này:

Với $b = 1, \dots, B$:

- Mẫu, có thay thế, n ví dụ huấn luyện từ X, Y ; gọi đây là X_b, Y_b .
- Huấn luyện cây phân loại hoặc cây hồi quy f_b trên X_b, Y_b .

Sau khi đào tạo, các dự đoán cho các mẫu không nhìn thấy x' có thể được thực hiện bằng cách lấy trung bình các dự đoán từ tất cả các cây hồi quy riêng lẻ trên x' :

$$\hat{f} = \frac{1}{B} \sum_{b=1}^B f_b(x')$$

hoặc bằng cách lấy phiếu đa số trong trường hợp cây phân loại.

Quy trình khởi động này dẫn đến hiệu suất mô hình tốt hơn vì nó làm giảm phương sai của mô hình mà không làm tăng độ lệch. Điều này có nghĩa là mặc dù dự đoán của một cây rất nhạy cảm với nhiễu trong tập huấn luyện của nó, nhưng giá trị trung bình của nhiều cây lại không như vậy, miễn là các cây đó không có mối tương quan. Đơn giản chỉ cần huấn luyện nhiều cây trên một tập huấn luyện sẽ tạo ra các cây có mối tương quan chặt chẽ (hoặc thậm chí cùng một cây nhiều lần, nếu thuật toán huấn luyện mang tính xác định); lấy mẫu bootstrap là một cách khử tương quan giữa các cây bằng cách hiển thị cho chúng các tập huấn luyện khác nhau.

Ngoài ra, ước tính độ không đảm bảo của dự đoán có thể được thực hiện dưới dạng độ lệch chuẩn của các dự đoán từ tất cả các cây hồi quy riêng lẻ trên x' :

$$\sigma = \sqrt{\frac{\sum_{b=1}^B (f_b(x') - \hat{f})^2}{B - 1}}.$$

Số lượng mẫu/cây (đại lượng B) là một tham số miễn phí. Thông thường, vài trăm đến vài nghìn cây được sử dụng, tùy thuộc vào quy mô và tính chất của tập huấn luyện. Có thể tìm thấy số lượng cây B tối ưu bằng cách xác thực chéo hoặc bằng cách quan sát lỗi ngoài bao đóng: lỗi dự đoán trung bình trên mỗi mẫu huấn luyện x_i , chỉ sử dụng những cây không có x_i trong mẫu bootstrap của chúng. Lỗi huấn luyện và kiểm tra có xu hướng chững lại sau khi một số cây đã phù hợp.

c) Từ đóng gói đến rừng cây ngẫu nhiên

Quy trình trên mô tả thuật toán đóng bao ban đầu cho cây. Rừng ngẫu nhiên cũng bao gồm một loại sơ đồ đóng bao khác: chúng sử dụng thuật toán học cây đã sửa đổi để chọn, tại mỗi phần phân chia ứng viên trong quá trình học, một tập hợp con ngẫu nhiên của các tính năng. Quá trình này đôi khi được gọi là "đóng bao tính năng". Lý do thực hiện điều này là do mối tương quan giữa các cây trong một mẫu bootstrap thông thường: nếu một hoặc một

vài đặc điểm là yếu tố dự báo rất mạnh cho biến phản hồi (đầu ra mục tiêu), thì các đặc điểm này sẽ được chọn trong nhiều cây B, khiến chúng bị ảnh hưởng. trở nên tương quan.

3.2.2. Các thuộc tính

a) Biến quan trọng

Rừng ngẫu nhiên có thể được sử dụng để xếp hạng tầm quan trọng của các biến trong bài toán hồi quy hoặc phân loại một cách tự nhiên. Kỹ thuật sau đây được mô tả trong bài báo gốc của Breiman và được triển khai trong gói R RandomForest. [29]

b) Hoán vị quan trọng

Bước đầu tiên trong việc đo lường tầm quan trọng của biến trong tập dữ liệu

$$\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$$

là để khớp một khu rừng ngẫu nhiên với dữ liệu. Trong quá trình điều chỉnh, lỗi ngoài túi của mỗi điểm dữ liệu được ghi lại và tính trung bình trên toàn nhóm (các lỗi trên bộ kiểm tra độc lập có thể được thay thế nếu không sử dụng tính năng đóng bao trong quá trình huấn luyện).

Để đo lường tầm quan trọng của tính năng thứ j sau khi đào tạo, các giá trị của tính năng thứ j được hoán vị trong các mẫu xuất túi và lỗi xuất túi một lần nữa được tính toán trên tập dữ liệu bị nhiễu loạn này. Điểm quan trọng đối với tính năng thứ j được tính bằng cách lấy trung bình chênh lệch về lỗi xuất túi trước và sau khi hoán vị trên tất cả các cây. Điểm số được chuẩn hóa bằng độ lệch chuẩn của những khác biệt này.

Các đặc điểm tạo ra giá trị lớn cho điểm này được xếp hạng quan trọng hơn các đặc điểm tạo ra giá trị nhỏ. Phương pháp xác định tầm quan trọng thay đổi này có một số nhược điểm:

- ✓ Đối với dữ liệu bao gồm các biến phân loại có số cấp khác nhau, các khu rừng ngẫu nhiên sẽ thiên về những thuộc tính có nhiều cấp hơn.
- ✓ Nếu dữ liệu chứa các nhóm đặc điểm tương quan có mức độ liên quan tương tự đối với đầu ra thì các nhóm nhỏ hơn sẽ được ưa chuộng hơn các nhóm lớn hơn.
- ✓ Ngoài ra, quy trình hoán vị có thể không xác định được các đặc điểm quan trọng khi có các đặc điểm thẳng hàng. Trong trường hợp này, việc hoán vị các nhóm đặc điểm tương quan với nhau là một giải pháp.

c) Mối quan hệ với K láng giềng gần nhất

Rừng ngẫu nhiên và K láng giềng gần nhất có nhiều điểm tương đồng với nhau. Chúng đều sử dụng một sơ đồ vùng lân cận có trọng số. Trong đó, các mô hình được xây dựng từ một tập huấn luyện

$$\{(x_i, y_i)\}_{i=1}^n$$

Với phán đoán \hat{y} cho mỗi điểm mới x' được xác định bởi các láng giềng gần nhất của nó. Công thức tính:

$$\hat{y} = \sum_{i=1}^n W(x_i, x') y_i.$$

Trong đó, $W(x, x')$ là có trọng số không âm của điểm được huấn luyện thứ i liên quan đến x' trong cùng cây. Cho từng phần x' , trọng số cho điểm x_i phải có tổng bằng 1. Hàm trọng số được thực hiện như sau:

- ✓ Trong k-NN, các trọng số $W(x, x') = 1/k$ nếu x_i là 1 trong k láng giềng gần nhất và bằng 0 nếu ngược lại.
- ✓ Trong 1 cây, các trọng số $W(x, x') = 1/k'$ nếu x_i là 1 trong k' láng giềng gần nhất và bằng 0 nếu ngược lại.

Từ đó, một phán đoán trung bình của rừng là một tập m cây với tập trọng số W_j và quyết định là:

$$\hat{y} = \frac{1}{m} \sum_{j=1}^m \sum_{i=1}^n W_j(x_i, x') y_i = \sum_{i=1}^n \left(\frac{1}{m} \sum_{j=1}^m W_j(x_i, x') \right) y_i.$$

Điều này cho thấy rằng toàn bộ khu rừng lại là một sơ đồ lân cận có trọng số, với trọng số trung bình là trọng số của từng cây riêng lẻ. Những người lân cận của x' theo cách hiểu này là những điểm x_i chia sẻ cùng một lá ở bất kỳ cây j . Theo cách này, lân cận của x' phụ thuộc một cách phức tạp vào cấu trúc của cây và do đó phụ thuộc vào cấu trúc của tập huấn luyện. Gán nhãn tự động so sánh.

3.3. Kết quả thực nghiệm

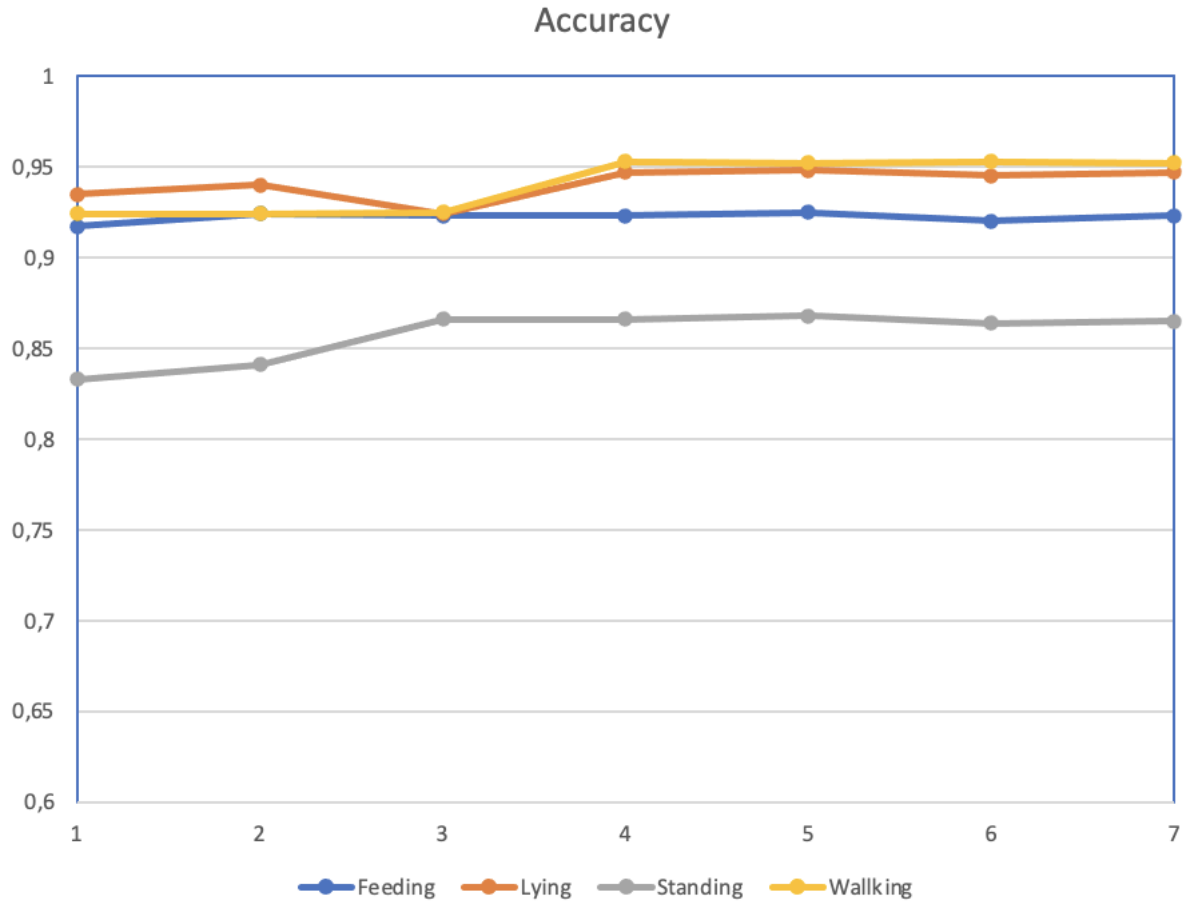
Thử nghiệm được thực hiện với mục đích đánh giá tác động đến hiệu suất phân loại khi gia tốc kế gắn trên cổ áo bị lệch. Chúng tôi đã so sánh hiệu suất của bộ phân loại đối với bảy bộ dữ liệu về độ chính xác, độ nhạy và độ chính xác.

- ✓ Tập dữ liệu gốc (dữ liệu thử nghiệm): DATA01
- ✓ 10% dữ liệu xoay 4 độ (mô phỏng 10% dữ liệu): DATA02
- ✓ 30% dữ liệu với góc xoay 4 độ (mô phỏng 30% dữ liệu): DATA03
- ✓ 50% dữ liệu với góc xoay 4 độ (mô phỏng 50% dữ liệu): DATA04
- ✓ 10% dữ liệu xoay 8 độ (mô phỏng 10% dữ liệu): DATA05
- ✓ 30% dữ liệu với góc xoay 8 độ (mô phỏng 30% dữ liệu): DATA06
- ✓ 50% dữ liệu với góc xoay 8 độ (mô phỏng 50% dữ liệu): DATA07
- ✓ Tác động làm lệch cảm biến đã được thử nghiệm trong hai trường hợp
- ✓ Hệ thống sử dụng cảm biến gắn trên vòng cổ (chỉ sử dụng dữ liệu gia tốc của cảm biến vòng cổ trên bò để phân loại hành vi). Kết quả được thể hiện trong Hình 16.
- ✓ Hệ thống sử dụng cảm biến gắn ở chân và vòng cổ – hệ thống được đề xuất (sử dụng dữ liệu gia tốc đồng bộ của cảm biến ở chân và vòng cổ trên bò để phân loại hành vi). Kết quả được thể hiện ở Hình 17

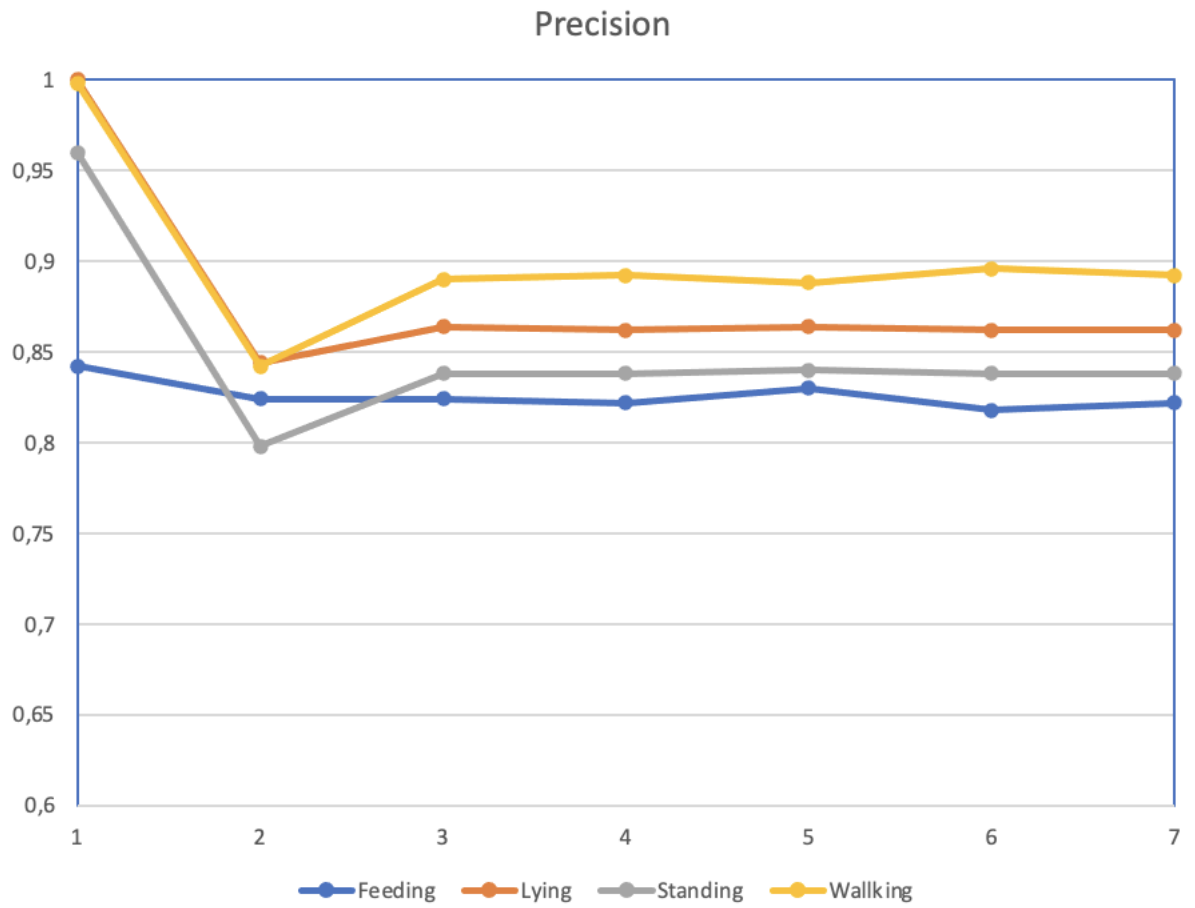
Dữ liệu thử nghiệm được thể hiện trong hình 12.

1	Feeding						Lying						Standing						Normal walking					
2	Acc in x (g)	Acc in y (g)	g(Acc in z)	g(Acc in x)	g(Acc in y)	g(Acc in z)	Acc in x (g)	Acc in y (g)	g(Acc in z)	g(Acc in x)	g(Acc in y)	g(Acc in z)	Acc in x (g)	Acc in y (g)	g(Acc in z)	g(Acc in x)	Acc in y (g)	g(Acc in z)	g(Acc in x)	Acc in y (g)	g(Acc in z)			
3	-962.79365	-202.825	34.46032	-640.182	386.3636	-718.364	0	998.5417	-64.2083	-649.273	-333.455	-733.636	-938.273	353.2727	-251.545	-390.091	368	-792.727	-979.556	-106.389	-36	-931.182	28.27273	-277.909
4	.949	-257	37.25	-690.364	330.0909	-709.727	2.583333	998.1667	-66.1667	-647.182	-330.545	-737.545	-881.833	-388.667	-222.25	-428.818	558.0909	-738.091	-988.75	-17.8333	-133.417	-967	46.72727	-310.455
5	-958.45455	-233	-1.72727	-668.818	167.0909	-748.273	-0.72727	998	-68	-645.545	-333.364	-732.636	-934.455	-298.818	-197.818	-421.545	489.5455	-799.727	-1140.73	-111.818	-389.091	-941.667	48.75	-277.833
6	-959.58333	-220.333	-12	-663.917	142.0833	-779	-2.09091	1001.182	-67.6364	-644.636	-331.364	-733.182	-942.273	-300.636	-117.364	-483.083	544.9167	-752.333	-1164.75	-110.917	-121.917	-940.364	44.45455	-276.455
7	-958.45455	-231.545	0.909091	-677.727	106.8182	-766.364	-1.42857	996.8571	-66.1429	-644.182	-330.364	-731.455	-938.545	-288.909	22.72727	-419.727	372	-735	-1093.36	-167.364	-108.909	-966.909	18.18182	-351.091
8	-959.5	-232	-3.33333	-681.909	123.8182	-783.818	2.666667	1001.667	-66.5833	-640.083	-332.583	-738.417	-976.167	-286.5	99.58333	-391.364	553.1818	-788.091	-798.636	51.90909	-149.455	-941.727	88.54545	-293.273
9	-959.45455	-234.364	-81.2727	-664.727	171.4545	-774.091	5.5	998.8333	-66.75	-642.545	-335.273	-730.273	-1260	-102.636	-194.455	-393.455	462.1818	851.818	-993.727	-25	-197.273	-920.083	29.33333	-285.667
10	-995.81818	-230.545	-17.2727	-620.417	411.5833	-722	0.916667	992.5	-68.8333	-648.182	-338.273	-735.091	-957.545	-226.364	-161.636	-400	464.6364	-839.909	-1018.27	-147.273	-376.091	-976.909	77.36364	-292.909
11	-959.58333	-236.75	-32.1667	-535.455	465.0909	-766.364	4	997	-71.0833	-643.455	-334.909	-727.818	-958.364	-218.818	-160	-391.75	407.0833	-889.083	-1153.08	-69.5833	-30.1667	-938.636	52.45455	-321.273
12	-958.09091	-235.091	-32.3636	-504.182	423.7273	-784.727	2	998	-73.5	-633.909	-337.909	-746.727	-959.545	-202	-153.455	-263.727	431	-915.455	-896	-58.8182	-709.727	-939.364	18.18182	-294.455
13	-960.18182	-228.455	-14.6364	-490.727	409.2727	-838.091	1.083333	997.6667	-77.0833	-631.364	-343.636	-740.273	-967.182	-197.545	-126.818	-229.636	453.6364	-917.091	-747.545	61.54545	-230.364	-944.667	93.25	-303.417
14	-1125.25	-246.417	2.5	-545.333	480.0833	-679.3	-0.25	996.5833	-75.75	-631.273	-349.273	-739.364	-974.091	-160.727	-114.364	-230.909	440.6364	-914.091	-1219.73	-267.091	-69.5455	-954.455	37.45455	-280.182
15	-957.45455	-194.818	15.09091	-497.636	446.4545	-810.909	2.833333	998.25	-76	-633.909	-348.091	-739.727	-977.273	-140.818	-119.727	-215.5	447.5	-926.667	-720.583	123.3333	-655.5	-985.364	37.45455	-304.273
16	-966.58333	-195.667	14.5	-435.818	425	-859.909	-1.61538	999	-73.8462	-634.909	-341.545	-738.273	-978.273	-124.273	-114.455	-382.8	445.3	-893.8	-1446.82	64.45455	-376.364	-937.182	30.63636	-303.364
17	-970.72727	-182.182	-2.81818	-572.909	225	-808.545	0.727273	998.7273	-74.3636	-637.727	-344.727	-737.909	-982.545	-117.818	-103.818	-497.25	472.6667	-762.333	-847.455	-205	-77	-953.25	30.16667	-286.667
18	-1047.25	-175.083	-17.6667	-656.25	183.75	-794.417	-2.5	997.0833	-74.75	-639.636	-348.636	-732.455	-979.167	-117.583	-78.75	-545	428.8182	-770.455	-1297.64	-25.3636	-207.364	-958.182	103.5455	-307.182
19	-535.27273	-139.909	-4.72727	-589.909	129.2727	-831.455	0.166667	996.3333	-74.5	-643.833	-348.417	-728.333	-979.455	-117.545	-69.1818	-470.917	266.1667	-885.417	-741.364	46.27273	-458.273	-908.364	41.81818	-286.455
20	-963.25	-197.083	26.5	-686.636	-10.3636	-816	1.666667	996.9167	-76.9167	-643.273	-358.636	-720.364	-977.667	-111.417	-72.0833	-500.818	311.6364	-861.636	-957.636	-164.636	-234.182	-974.455	88.45455	-297.455
21	-958.09091	-215.545	38.54545	-657.727	-251.818	-742.727	0.25	1006.25	-77.8333	-638.909	-356.818	-730.364	-982.455	-121	-89	-490.909	448.8182	-789.091	-1175.55	-88.8182	83.27273	-967.455	-22.2727	-310.273
22	-950.58333	-244.25	63.83333	-595.909	-385.727	-725.091	-3.83333	995.4167	-75.5833	-642.364	-353.727	-719.182	-974.364	-100	-100.727	-512.182	439.9091	-749.091	-716.25	-128.5	121	-965	-5.25	-307.833
23	-932.45455	-153.727	30.18182	-569.545	-442.273	-766.091	-0.58333	996.75	-78.5	-653	-349.545	-714.909	-986.333	-41.25	-55.3333	-453.273	569.0909	-728.364	-1399.36	15.72727	-510.364	-924.636	-14.8182	-301.727
24	-947	-209.917	48.08333	-531.833	-486.75	-730.167	1.615385	997.4615	-77.3846	-652.909	-358.455	-716	-948.455	219.8182	106.4545	-460.167	454.6667	-815.083	-772.4	-91.8	281.7	-937.545	-27.9091	-264.182
25	-977	-157.273	25.36364	-630.455	-501.364	-600.727	1.583333	997.5	-78.4167	-643.273	-362.364	-716.909	-1028.45	129.3636	-105.182	-434.455	202.3636	-915.091	-880.667	-276.333	-200.667	-952.455	-75.8182	-337.727
26	-349.25	-78.5	-14.1667	-607.364	-487.091	-676.727	2.166667	998.75	-67.8333	-643.909	-367.636	-714.818	-947.727	-261.182	-199.545	-469.091	34.90909	-926.636	-881.182	-43.0909	291.8182	-965.25	4.666667	-311.333
27	-939.36364	-58.2727	-16.5455	-607.455	-495.636	-663.909	-0.45455	994.8182	-66.5455	-640.091	-352.636	-733	-961.273	-214.818	-150.909	-513.167	6.666667	-909.833	-1342.36	39	241.7273	-946.636	-34.4545	-331.091
28	-961.66667	-155	-35.9683	-586.833	-468.818	-766.727	-0.416667	996.5693	-75.0526	-630.364	-360.545	-736.091	-975.182	-163.091	-113.727	-510	16.09091	-900.091	-781.333	81.66667	-333.636	-851.333	15.45455	-388.273

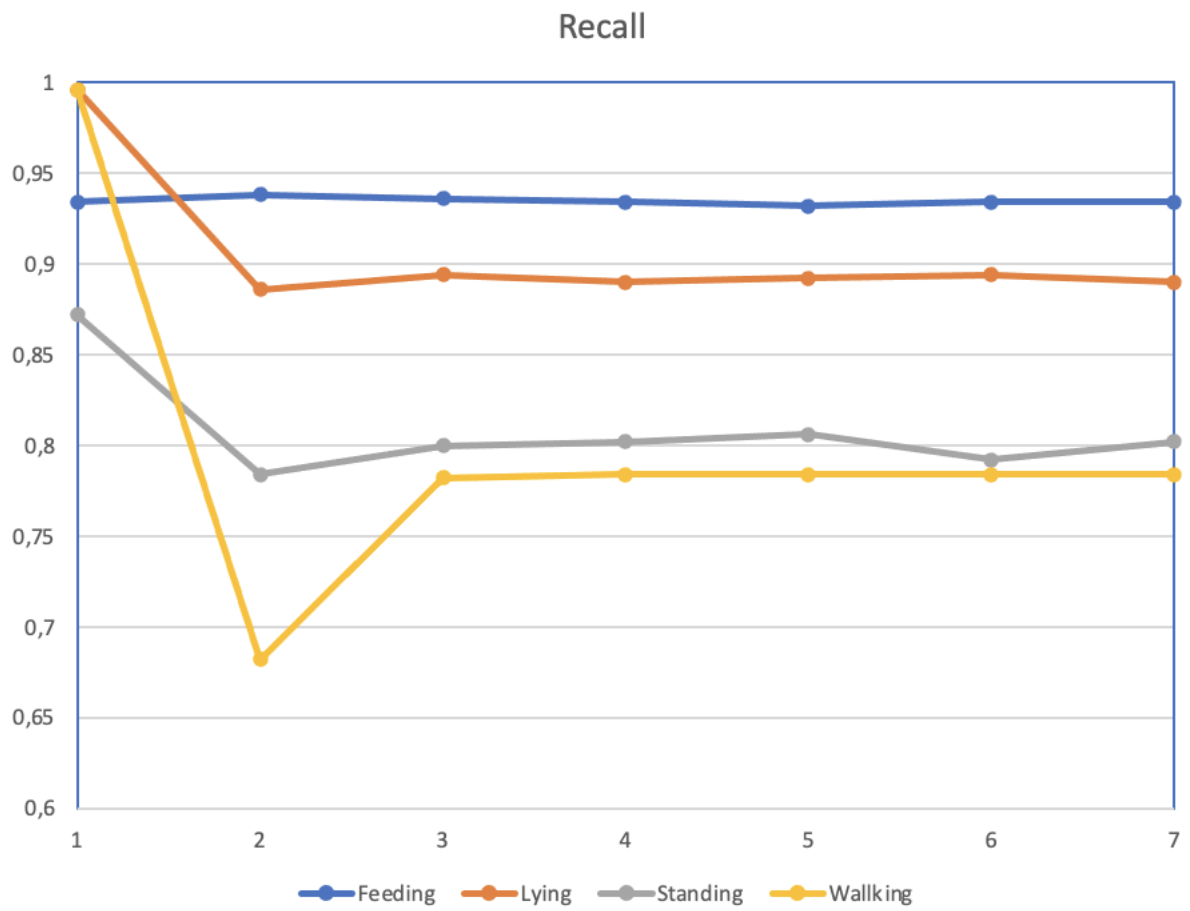
Hình 12: Bảng dữ liệu thu thập 1CowData_2sensor phục vụ đánh giá



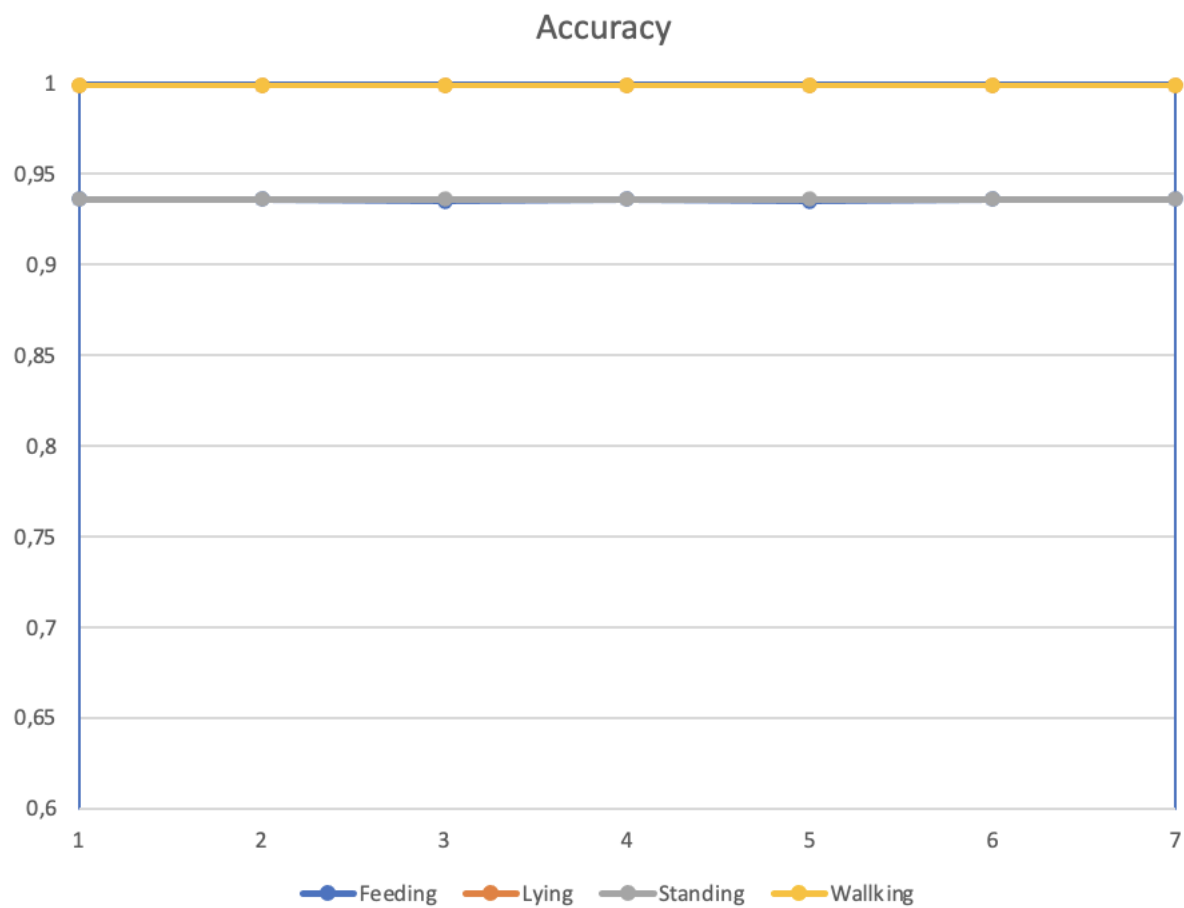
Hình 13a: Accuracy 4 hành vi của 7 dataset chỉ dùng cõ



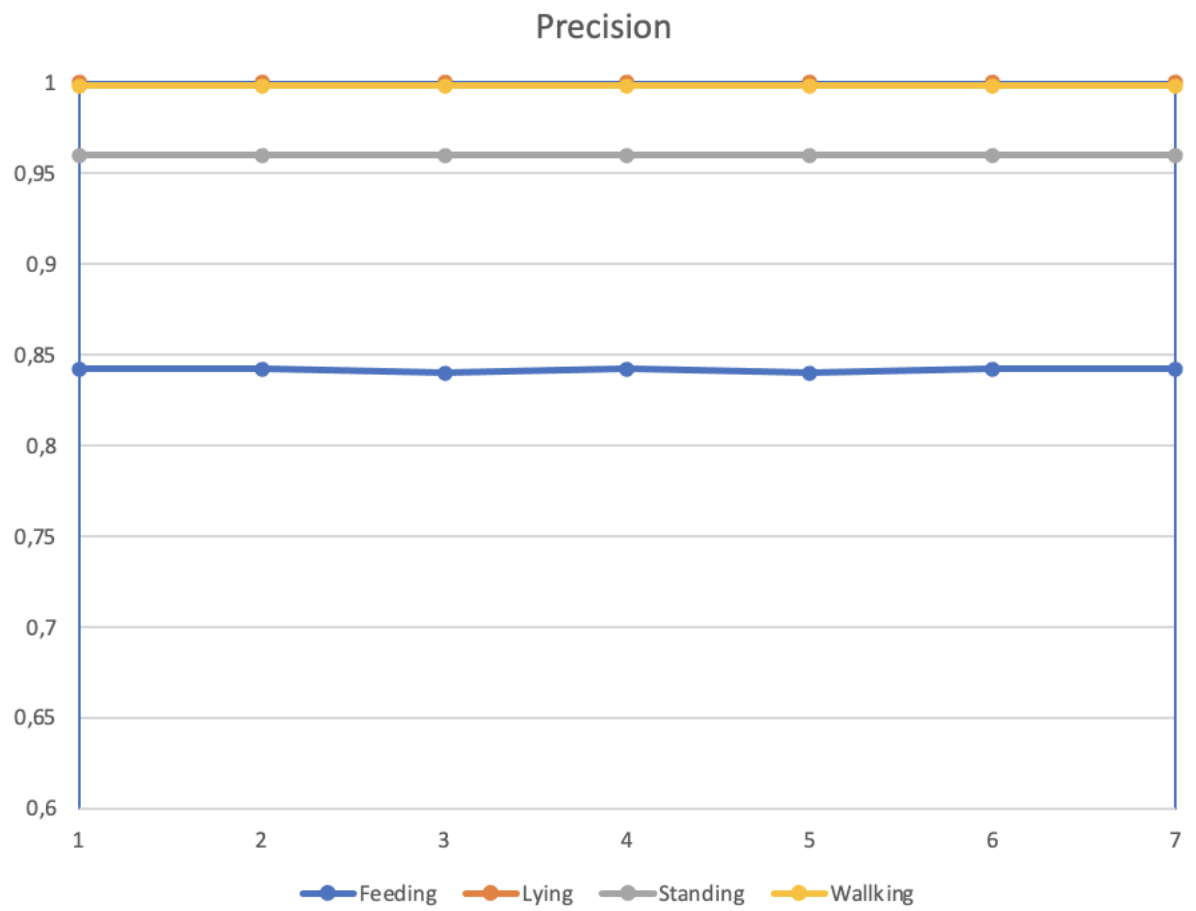
Hình 13b: Precision 4 hành vi của 7 dataset chỉ dùng cổ.



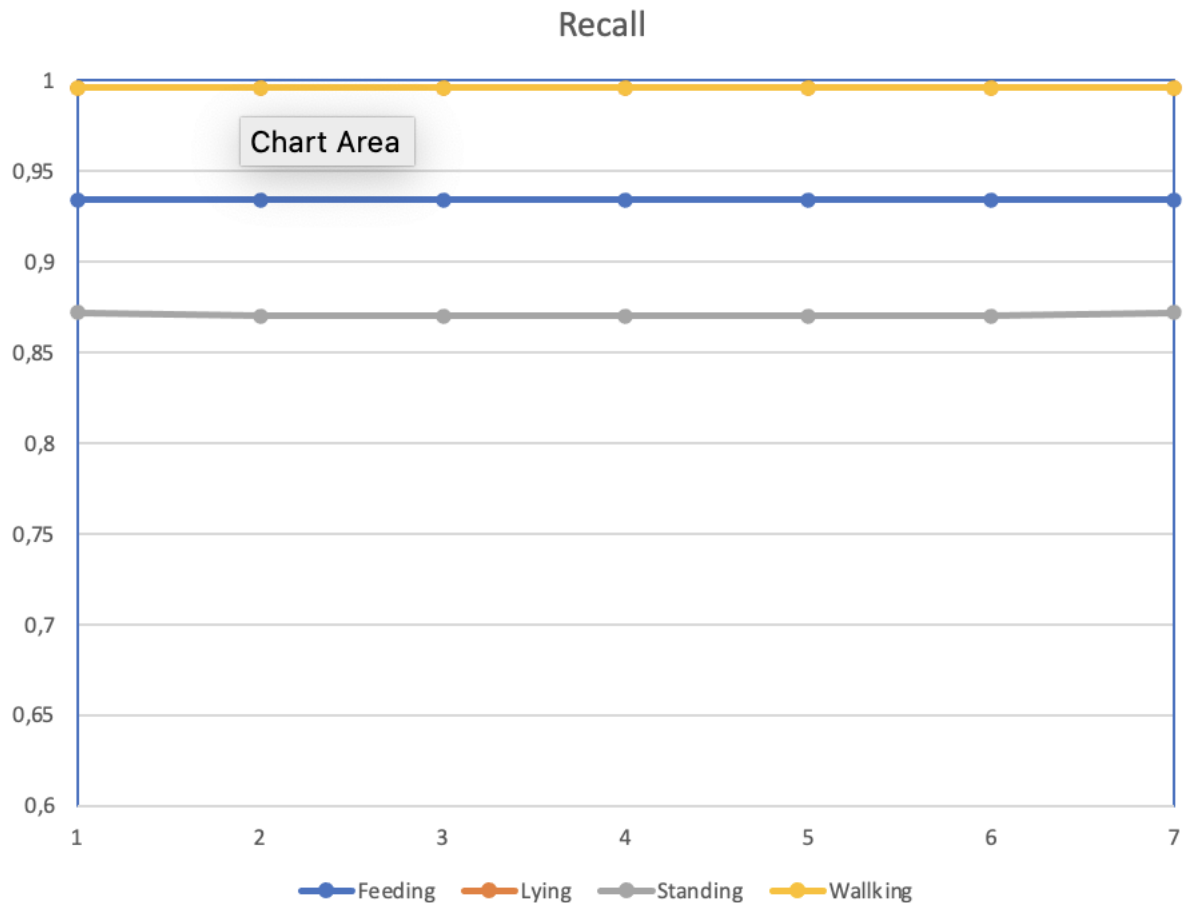
Hình 13c: Recall 4 hành vi của 7 dataset chỉ dùng cổ



Hình 14a: Accuracy 4 hành vi của 7 dataset dùng chân + cổ



Hình 14b: Precision 4 hành vi của 7 dataset dùng chân + cổ



Hình 14c: Recall 4 hành vi của 7 dataset dùng chân + cổ

Hình 13(a), Hình 13(b) và Hình 13(c) biểu thị sự so sánh về sự chính xác, độ chính xác và độ nhạy (trung bình của tất cả các con bò) đối với bốn hành vi tương ứng trong bảy trường hợp khi hệ thống sử dụng và cảm biến gắn trên cổ. Trong Hình 14(a), Hình 14(b) và Hình 14(c), cho thấy sự chính xác, độ chính xác và độ nhạy trung bình trong bảy trường hợp thử nghiệm là tương tự nhau. Điều đó có nghĩa là tác động của độ lệch cảm biến sẽ không đáng kể nếu hệ thống sử dụng cảm biến gắn trên chân và gắn trên cổ. Những kết quả này có ý nghĩa vì lợi thế khi sử dụng cả cảm biến gắn trên chân và gắn trên cổ. Lưu ý rằng trong Hình 13(a), đường Nằm và Đường Đi chồng lên nhau, Đường Cho ăn và Đường Đứng chồng lên nhau. Những điều tương tự có thể được tìm thấy trong Hình 13(b) và Hình 13(c). Nó không thay đổi quan điểm mà chúng tôi quan tâm.

Bảng 2 và Bảng 3 cho thấy kết quả về hiệu suất trung bình trong hai trường hợp: hệ thống sử dụng cảm biến gắn trên cổ và hệ thống sử dụng cảm biến gắn trên chân và cổ. Bảy bộ dữ liệu trong Bảng 4 có độ chính xác, độ nhạy và độ chính xác tương tự nhau. Trong Bảng 2, DATA01 có độ nhạy và độ chính xác cao nhất. Vì vậy, có thể kết luận tương tự:

- Tác động làm lệch cảm biến là đáng kể nếu hệ thống chỉ sử dụng cảm biến gắn ở cổ.
- Tác động làm lệch cảm biến sẽ không đáng kể nếu hệ thống sử dụng cảm biến gắn trên chân và gắn trên cổ.

Bảng 2: Bình quân các chỉ số của cảm biến cổ

DATASET	Behavior pattern	Accuracy	Recall	Perccision
1	Feeding	0,9361	0,934	0,842
	Lying	0,9992	0,996	1
	Standing	0,9361	0,872	0,96
	Walking	0,9989	0,996	0,998
2	Feeding	0,920	0,938	0,824
	Lying	0,940	0,886	0,844
	Standing	0,841	0,784	0,798
	Walking	0,924	0,682	0,842
3	Feeding	0,922	0,936	0,824
	Lying	0,924	0,894	0,864
	Standing	0,866	0,8	0,838
	Walking	0,925	0,782	0,89
4	Feeding	0,923	0,934	0,822
	Lying	0,947	0,89	0,862
	Standing	0,866	0,802	0,838
	Walking	0,953	0,784	0,892

5	Feeding	0,925	0,932	0,83
	Lying	0,948	0,892	0,864
	Standing	0,868	0,806	0,84
	Walking	0,952	0,784	0,888
6	Feeding	0,920	0,934	0,818
	Lying	0,945	0,894	0,862
	Standing	0,864	0,792	0,838
	Walking	0,953	0,784	0,896
7	Feeding	0,923	0,934	0,822
	Lying	0,947	0,89	0,862
	Standing	0,865	0,802	0,838
	Walking	0,952	0,784	0,892

Bảng 3: Bình quân các chỉ số của cảm biến gắn chân + cổ

DATASET	Behavior pattern	Accuracy	Recall	Perccision
1	Feeding	0,9361	0,934	0,842
	Lying	0,9992	0,996	1
	Standing	0,9361	0,872	0,96
	Walking	0,9989	0,996	0,998
2	Feeding	0,936	0,934	0,842
	Lying	0,999	0,996	1
	Standing	0,936	0,87	0,96
	Walking	0,999	0,996	0,998
3	Feeding	0,935	0,936	0,84
	Lying	0,999	0,996	1
	Standing	0,936	0,87	0,96
	Walking	0,999	0,996	0,998

4	Feeding	0,936	0,934	0,842
	Lying	0,999	0,996	1
	Standing	0,936	0,872	0,96
	Walking	0,999	0,996	0,998
5	Feeding	0,935	0,936	0,84
	Lying	0,999	0,996	1
	Standing	0,936	0,87	0,96
	Walking	0,999	0,996	0,998
6	Feeding	0,936	0,934	0,842
	Lying	0,999	0,996	1
	Standing	0,936	0,87	0,96
	Walking	0,999	0,996	0,998
7	Feeding	0,936	0,934	0,842
	Lying	0,999	0,996	1
	Standing	0,936	0,872	0,96
	Walking	0,999	0,996	0,998

KẾT LUẬN

Trong luận văn này, tôi đã nghiên cứu để xác định mức độ ảnh hưởng của độ lệch cảm biến gia tốc đến kết quả phân loại hành vi bò. Từ một dataset luận văn mô phỏng thực nghiệm cho các dataset ứng với mức độ đo lệch khác nhau của thiết bị. Từ đó, luận văn đưa giá kết quả đánh giá cho các tham số Accuracy, Precision và Recall cho bài toán phân loại 4 hành vi: đi, nằm, ăn và đứng của bò. Kết quả nghiên cứu có thể góp phần giúp các nhà nghiên cứu dự đoán được độ chính xác của các hệ thống phân loại hành vi trong bài toán quản lý nông trại thông minh. Mức độ ảnh hưởng của các độ lệch thiết bị có khả năng gây ảnh hưởng tới bài toán phân loại từ đó có giải pháp phù hợp cho việc thiết kế cách đo thiết bị và gia cố phù hợp hơn trong chăn nuôi bò.

Những đóng góp của luận văn

- ✓ Xây dựng tập dữ liệu từ 1 dataset dữ liệu không đo lệch thành 7 dataset khác nhau ứng với các độ lệch: 0, 4, 8, 12, 16, 20, 24 và 28 phục vụ cho việc đánh giá phân loại hành vi
- ✓ Đưa ra các biểu đồ đánh giá các tiêu chí Accuracy, Precision và Recall cho bài toán phân loại
- ✓ Ứng dụng bài toán phân loại bằng Random Forest để phân loại hành vi bò

Định hướng nghiên cứu trong tương lai

Trong các nghiên cứu tiếp theo, học viên sẽ mở rộng đánh giá các yếu tố khác có thể ảnh hưởng đến bài toán phân loại. Xây dựng một hệ thống giúp giám sát hoạt động quản lý chăn nuôi bò để phục vụ nghiên cứu và ứng dụng vào thực tiễn trong chăn nuôi.

Tài liệu tham khảo

1. F. W. Oudshoorn, C. Cornou, A. L. F. Hellwing, H. H. Hansen, L. Munksgaard, P. Lund, T. Kristensen, "*Estimation of grass intake on pasture for dairy cows using tightly and loosely mounted di- and tri-axial accelerometers combined with bite count*", *Comput. Electron. Agric.*, vol. 99, pp. 227–235, Sep. 2013.
2. D.S. Conner, V. Campbell-Arvai, M.W. Hamm, "*Value in the values: pasture-raised livestock products offer opportunities for reconnecting producers and consumers*", *Renew. Agric. Food Syst.*, vol. 23, pp. 62–69, 2008.
3. B. Robert, B. J. White, D. G. Renter, R. L. Larson, "*Evaluation of three-dimensional accelerometers to monitor and classify behavior patterns in cattle*", *Comput. Electro. Agric.*, vol. 67, no. 1-2, pp. 80–84, Mar. 2009.
4. C. Arcidiacono, S. M. Porto, M. Mancino, G. Cascone, "*A threshold-based algorithm for the development of inertial sensor-based systems to perform real-time cow step counting in free-stall barns*", *Biosyst. Eng.*, vol. 153, pp. 99–109, Nov. 2016.
5. C. Arcidiacono, S. M. C. Porto, M. Mancino, G. Cascone, "*Development of a threshold-based classifier for real-time recognition of cow feeding and standing behavioral activities from accelerometer data*", *Comput. Electro. Agric.*, vol. 134, pp. 124–134, Jan. 2017.
6. N. B. Cook, T. B. Bennett, K. V. Nordlund, "*Monitoring Indices of Cow Comfort in Free-Stall-Housed Dairy Herds*", *J. Dairy Sci.*, vol. 88, no. 11, pp. 3876–3885, Nov. 2005.
7. S. A. Schoenig, T. S. Hildreth, L. Nagl, H. Erickson, M. F. Spire, D. Andresen, S. Warren, "*Ambulatory Instrumentation Suitable for Long-Term Monitoring of Cattle Health*", in *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, San Francisco, CA, USA, 2004.
8. F. Mahmoud, B. Christopher, A. Maher, H. Jürg, S. Alexander, S. Adrian, H. Gaby, "*Prediction of calving time in dairy cattle*", *Anim. Reprod. Sci.*,

- vol. 187, pp. 37–46, Dec. 2017.
9. H. C. Weigele, L. Gygax, A. Steiner, B. Wechsler, J. B. Burla, "*Moderate lameness leads to marked behavioral changes in dairy cows*", J. Dairy Sci., vol. 3101, pp. 2370–2382, Mar. 2018.
 10. G. M. Pereira, J. H. Bradley, I. E. Marcia, "*Validation of an eartag accelerometer sensor to determine rumination, eating, and activity behaviors of grazing dairy cattle*", J. Dairy Sci., vol. 101, pp. 2492–2495, Jan. 2019.
 11. M. R. Borchers, Y. M. Chang, "*A validation of technologies monitoring dairy cow feeding, ruminating, and lying behaviors*", J. Dairy Sci., vol. 999, pp. 7458–7466, May 2016.
 12. S. M. C. Porto, C. Arcidiacono, "*Localization and identification performances of a real-time 4 system based on ultra wide band technology for monitoring and tracking dairy cow behavior in semi-open free-stall barn*", Comput. Electro. Agric., vol. 108, pp. 221–229, Oct. 2014.
 13. M. J. Gibb, C. A. Huckle, R. Nuthall, A. J. Rook, "*Effect of sward surface height on intake and grazing behaviour by lactating Holstein Friesian cows*", Grass Forage Sci., vol. 52, pp. 309–321, 1997.
 14. E. D. Ungar, N. Ravid, T. Zada, E. Ben-Moshe, R. Yonatan, H. Baram, A. Genizi, "*The implications of compound chew–bite jaw movements for bite rate in grazing cattle*", Appl. Anim. Behav. Sci., vol. 98, pp. 183–195, 2006.
 15. Y. Ueda, F. Akiyama, S. Asakuma, N. Watanabe, "*The use of physical activity monitor to estimate the eating time of cows in pasture*", J. Dairy Sci., vol. 94, pp. 3498–3503, 2011.
 16. C. W. Maina, "*IoT at the Grassroots–Exploring the Use of Sensors for Livestock Monitoring*", in Ist-Africa Week, 2017, pp. 1–8.
 17. J. A. V. Diosdado, Z. E. Barker, "*Classification of behavior in housed dairy cows using an accelerometer-based activity monitoring system*", Anim. Biotelemetry, vol. 3, pp. 1–14, Jun. 2015.

18. E. S. Nadimi, H. T. Sogaard, "*Observer Kalman filter identification and multiple-model adaptive estimation technique for classifying animal behavior using wireless sensor networks*", *Comput. Electro. Agric.*, vol. 68, pp. 9–17, Aug. 2009.
19. K. O'Driscoll, L. Boyle, A. Hanlon, "*A brief note on the validation of a system for recording lying behavior in dairy cows*", *Appl. Anim. Behav. Sci.*, vol. 111, pp. 195–200, May 2008.
20. M. S. Shahriar, D. Smith, "*Detecting heat events in dairy cows using accelerometers and unsupervised learning*", *Comput. Electro. Agric.*, vol. 128, pp. 20–26, Oct. 2016.
21. J. M. Talavera, L. E. Tobón, J. A. Gómez, M. A. Culman, J. M. Aranda, D. T. Parra, et al., "*Review of IoT applications in agro-industrial and environmental fields*", *Comput. Electro. Agric.*, vol. 142, pp. 283–297, Nov. 2017.
22. J. Wang, Z. He, "*Development and validation of an ensemble classifier for real-time recognition of cow behavior patterns from accelerometer data and location data*", *PLoS One*, vol. 13, Aug. 2018.
23. B. D. Robért, B. J. White, D. G. Renter, R. L. Larson, "*Determination of lying behavior patterns in healthy beef cattle by use of wireless accelerometers*", *Am. J. Vet. Res.*; vol. 72, pp. 467–473, Apr. 2011.
24. J. Wang, Z. He, J. Ji, K. Zhao, H. Zhang, "*IoT-based measurement system for classifying cow behavior from tri-axial accelerometer*", *Cienc. Rural*, vol. 49, pp. 1–13, Mar. 2019.
25. C. P. K. Phung, D. T. Tran, V. T. Duong, H. T. Nguyen, D. N. Tran, "*The new design of cows' behavior classifier based on acceleration data and proposed feature set*", *Math. Biosci. Eng.*, vol. 17, no. 4, pp. 2760-2780, March 2020.
26. P. Martiskainen, M. Jarvinen, "*Cow behavior pattern recognition using a three-dimensional accelerometer and support vector machines*", *Appl. Anim. Behav. Sci.*, vol. 119, pp. 32–38, Jun. 2009.

27. J. A. Vázquez Diosdado, Z. E. Barker, H. R. Hodges et al., "*Classification of behavior in housed dairy cows using an accelerometer-based activity monitoring system*", *Anim. Biotelemetry*, vol. 3, no. 15, Jun. 2015.
28. Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2008). *The Elements of Statistical Learning* (2nd ed.). Springer. ISBN 0-387-95284-5.
- 29.[29] Liaw A (16 October 2012). "*Documentation for R package randomForest*" (PDF). Retrieved 15 March 2013.