

**ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ**



NGUYỄN ĐÌNH CHINH

**NGHIÊN CỨU THUẬT TOÁN HỌC MÁY ÁP DỤNG
CHO HỆ THỐNG GIÁM SÁT VÀ NHẬN DẠNG HÀNH VI TRÊN BỜ**

**LUẬN VĂN THẠC SĨ: NGÀNH CÔNG NGHỆ KỸ THUẬT ĐIỆN TỬ,
TRUYỀN THÔNG**

Hà Nội – 2017

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

NGUYỄN ĐÌNH CHINH

NGHIÊN CỨU THUẬT TOÁN HỌC MÁY ÁP DỤNG
CHO HỆ THỐNG GIÁM SÁT VÀ NHẬN DẠNG HÀNH VI TRÊN BÒ

Ngành: Công nghệ Kỹ thuật Điện tử, Truyền thông
Chuyên Ngành: Kỹ thuật Điện tử
Mã Số: 60 52 02 03

LUẬN VĂN THẠC SĨ: NGÀNH CÔNG NGHỆ KỸ THUẬT ĐIỆN TỬ,
TRUYỀN THÔNG

NGƯỜI HƯỚNG DẪN KHOA HỌC: TS. LÊ VŨ HÀ

PGS. TS. TRẦN ĐỨC TÂN

Hà Nội - 2017
LỜI CAM ĐOAN

Tôi xin cam đoan luận văn tốt nghiệp: “**Nghiên cứu thuật toán học máy áp dụng cho hệ thống giám sát và nhận dạng hành vi trên bờ**” là công trình nghiên cứu của riêng tác giả. Các số liệu, kết quả trình bày trong luận văn là hoàn toàn trung thực, chưa từng được công bố trong các bất kỳ công trình nào khác.

Trong luận văn có dùng một số tài liệu tham khảo như đã nêu trong phần tài liệu tham khảo.

Tác giả luận văn

Nguyễn Đình Chính

LỜI CẢM ƠN

Để hoàn thành luận văn tốt nghiệp này, trước tiên, tôi xin gửi lời cảm ơn chân thành và sâu sắc nhất và tình cảm đặc biệt tới người Thầy của tôi PGS. TS. Trần Đức Tân. Thầy là người đã luôn theo sát tôi, tận tình chỉ bảo, góp ý và hướng dẫn, định hướng cho tôi trong suốt quá trình làm luận văn này tại Khoa Điện tử Viễn thông, Trường Đại học Công nghệ. Tôi không chỉ được học ở Thầy phương pháp luận nghiên cứu khoa học, tôi còn tích lũy được rất nhiều bài học quý báu về cách làm việc chuyên nghiệp, lối tư duy đánh giá sự việc, những kinh nghiệm làm việc rất quan trọng cho tôi trong công việc sau này. Em cảm ơn Thầy rất nhiều!

Tôi xin cảm ơn đến thầy TS. Lê Vũ Hà đã cung cấp cho tôi các kiến thức nền tảng về lĩnh vực học máy, người đồng hướng dẫn tôi trong luận văn này.

Tôi xin cảm ơn các thầy, các anh là đồng nghiệp của tôi tại bộ môn Vi cơ điện tử và vi hệ thống – trường Đại học Công nghệ, họ luôn là tấm gương trong nghiên cứu khoa học và người luôn sát cánh, động viên tinh thần cũng như giúp đỡ tôi trong nghiên cứu. Cảm ơn anh Phùng Công Phi Khanh – Nghiên cứu sinh tại bộ môn, người đi cùng tôi trong nghiên cứu này, và cảm ơn các bạn sinh viên trong nhóm nghiên cứu, các em đã hỗ trợ để tôi có thể hoàn thành nghiên cứu.

Tiếp theo, tôi cũng xin gửi lời cảm ơn sâu sắc tới các Thầy, các Cô và các anh chị em trong Khoa đã luôn sẵn sàng giúp đỡ tạo điều kiện tốt nhất cho tôi trong quá trình làm luận văn.

Cuối cùng, tôi xin gửi những lời cảm ơn chân thành và yêu thương nhất tới bố mẹ của tôi, những người luôn luôn ủng hộ, động viên tôi cả về vật chất lẫn tinh thần để tôi có thể hoàn thành luận văn tốt nhất. Con cảm ơn bố mẹ thật nhiều!

Mặc dù có nhiều cố gắng, song thời gian thực hiện luận văn có hạn, nên luận văn còn nhiều hạn chế. Tôi rất mong nhận được nhiều sự góp ý, chỉ bảo của các thầy, cô để hoàn thiện hơn luận văn của mình.

Tôi xin chân thành cảm ơn!

Hà Nội, ngày ... tháng 04 năm 2017

Học viên

Nguyễn Đình Chinh

TÓM TẮT

Ngành chăn nuôi, sản xuất sữa và thực phẩm là ngành kinh tế quan trọng trong phát triển kinh tế xã hội và an ninh lương thực thế giới. Để bảo đảm sự phát triển bền vững của các ngành này việc giám sát và chăm sóc sức khỏe của gia súc có vai trò rất quan trọng và nó là nhu cầu thiết yếu đối với ngành chăn nuôi. Tại Việt Nam, có một số công ty sữa lớn như TH Truemilk, VINAMILK, để nâng cao năng suất và chất lượng sản phẩm, họ rất quan tâm đến vấn đề sức khỏe của bò. Vì vậy, họ có nhu cầu giám sát về thể chất và sinh lý của đàn bò càng thường xuyên càng tốt. Phát sinh từ bản chất của những khó khăn liên quan đến việc quản lý trang trại với các khu chăn thả lớn, các nhà chăn nuôi luôn luôn có nhu cầu "giám sát" động vật của họ một cách tự động và tiết kiệm chi phí nhất. Công nghệ mạng cảm biến không dây là một giải pháp khả thi cho vấn đề này.

Trong các thông tin cần cho việc chăm sóc sức khỏe gia súc thì hành vi là một trong những cơ sở quan trọng và nhạy cảm nhất. Việc giám sát hành vi của bò trên thực tế thường được tiến hành theo hướng quan sát chuyên động trên cổ bò hoặc chuyên động trên chân bò. Dữ liệu từ các thiết bị quan sát đó có thể được lưu lại tại thiết bị để xử lý sau hoặc được truyền thông không dây về một thiết bị trung tâm để xử lý. Tuy nhiên, việc thực thi các kỹ thuật trên còn rất cơ bản chưa đáp ứng được việc giám sát thời gian thực hay quy mô chăn thả lớn, nhiều hệ thống chỉ xác định một hoặc hai hành vi hoặc trạng thái động vật tại một thời điểm hoặc nhiều hành vi nhưng độ chính xác còn thấp. Trong các hệ thống này, cảm biến gia tốc được sử dụng khá phổ biến để theo dõi hành vi và tình trạng sức khỏe của động vật. Luận văn này đề xuất giám sát tình trạng hành vi của gia súc thông qua mạng cảm biến không dây. Trong đó, mỗi nút mạng sẽ là sự kết hợp của hai thiết bị giám sát hành vi trên cổ bò và chân của bò, chúng được kết nối không dây với nhau và được phát triển dựa trên cảm biến gia tốc 3 trục (cảm biến MPU6050) giúp xác định tình trạng chính xác hơn. Báo cáo này đang hoàn thiện mô hình hệ thống giám sát và sẽ được thực thi trong thực tế.

Một số phương pháp gần đây được đề xuất để phân loại hành vi tự động ở động vật chủ yếu dựa trên các thuật toán học máy khác nhau như cây quyết định, k-means, SVM và HMM. Luận văn sử dụng 02 thuật toán học máy: thuật toán cây quyết định và thuật toán SVM để phân loại hành vi của bò dựa trên dữ liệu cảm biến gia tốc ba trục từ cổ và chân bò được truyền đến một máy chủ thông qua mô hình mạng cảm biến không dây Zigbee. Tại máy chủ, các thuật toán giúp xử lý và phân loại hành vi được áp dụng nhằm đưa ra trạng thái chính xác. Kết quả nghiên cứu chỉ ra khả năng phân biệt các trạng thái hành vi của bò và sự vượt trội về kết quả khi sử dụng thuật toán SVM so với cây quyết định và kết hợp thông số gia tốc trên chân và cổ bò để đưa ra đánh giá chính xác về hành vi so với việc chỉ dùng thông số gia tốc trên cổ. Nghiên cứu đã chỉ ra khả năng phân biệt 05 hành vi: đi, đứng, nằm, ăn và uống nước của bò.

Từ khóa: *Giám sát, Phân loại hành vi, cảm biến gia tốc, mạng cảm biến không dây, Cây quyết định, Máy vector hỗ trợ (SVM).*

MỤC LỤC

DANH MỤC HÌNH VẼ	7
DANH MỤC BẢNG BIỂU.....	9
DANH MỤC VIẾT TẮT	10
DANH MỤC KÝ HIỆU	11
MỞ ĐẦU.....	12
Chương 1.....	14
TỔNG QUAN VỀ CÁC HỆ THỐNG GIÁM SÁT VÀ NHẬN DẠNG HÀNH VI TRÊN BÒ.....	14
1.1. Giới thiệu	14
1.2. Giới thiệu hệ thống giám sát bò	15
1.3. Giới thiệu về thuật toán học máy	16
1.3.1. Các loại học máy.....	17
1.3.2. Học máy theo đầu ra cụ thể.....	17
1.3.3. Các dạng học máy.....	18
1.4. Kết luận chương 1	20
Chương 2.....	21
THUẬT TOÁN HỌC MÁY GIÚP NHẬN DẠNG HÀNH VI TRÊN BÒ	21
2.1. Nhận diện hành vi trên bò.....	21
2.2. Thuật toán học máy cho nhận dạng hành vi trên bò	22
2.2.1. Thuật toán cây quyết định	22
2.2.2. Thuật toán SVM.....	24
2.3. Đánh giá hiệu năng của thuật toán học máy	30
2.4. Kết luận chương 2.....	31
Chương 3.....	32
THIẾT KẾ HỆ THỐNG, THỰC THI VÀ ĐÁNH GIÁ THUẬT TOÁN	32

3.1. Thiết kế hệ thống	32
3.1.1. Thiết kế hệ thống thu thập dữ liệu huấn luyện	32
3.1.2. Thiết kế hệ thống giám sát.....	34
3.2. Thực thi và đánh giá thuật toán.....	42
3.2.1. Thuật toán cây quyết định	42
3.2.2. Thuật toán SVM.....	47
3.2.3. Đánh giá hiệu năng	50
3.3. Kết luận chương 3	51
KẾT LUẬN.....	53
DANH SÁCH CÁC CÔNG BỐ	53
TÀI LIỆU THAM KHẢO	54

DANH MỤC HÌNH VẼ

Hình 2.1: Xác định ngưỡng theo theo đặc trưng	23
Hình 2.2: Mô hình cây quyết định	23
Hình 2.3: Đường phân tách mềm của thuật toán SVM	26
Hình 2.4: Thuật toán SVM một đối một.....	30
Hình 3.1: Sơ đồ nguyên lý của hệ thống thu dữ liệu mẫu	32
Hình 3.2: Cảm biến gia tốc 3 trục MPU6050	33
Hình 3.3: Module Bluetooth HC05	33
Hình 3.4: Bên trong thiết bị đo dữ liệu mẫu	34
Hình 3.5: Giao diện phần mềm thu dữ liệu mẫu	34
Hình 3.6: Mô hình hệ thống giám sát	35
Hình 3.7: Mô hình mạng cảm biến.....	36
Hình 3.8: Sơ đồ nguyên lý thiết bị đo gia tốc trên chân	37
Hình 3.9: Thiết bị đo gia tốc trên chân.....	37
Hình 3.10: Sơ đồ nguyên lý thiết bị đo gia tốc trên cổ	39
Hình 3.11: Module Zigbee.....	39
Hình 3.12: Thiết bị đo gia tốc trên cổ	39
Hình 3.13: Sơ đồ truyền nhận dữ liệu giữa 2 thiết bị đo gia tốc trên chân và cổ bò	40
Hình 3.14: Thiết bị được đeo trên bò.....	40
Hình 3.15: Raspberry Pi 3	41
Hình 3.16: Trung tâm điều phối.....	41
Hình 3.17: Sơ đồ thuật toán cây quyết định.....	43
Hình 3.18: Dữ liệu gia tốc 3 trục tại cổ bò.....	44
Hình 3.19: Dữ liệu gia tốc 3 trục trên chân bò.....	44
Hình 3.20: Biểu diễn giá trị VeDBA và SCAY của dữ liệu huấn luyện.....	45

Hình 3.21: Biểu diễn phân bố dữ liệu theo mỗi quan hệ gửi VeDBA và SCAY của dữ liệu huấn luyện.....	45
Hình 3.22: Đường cong ROC để xác định Threshold A.....	46
Hình 3.23: Đường cong ROC để xác định Threshold B.....	46
Hình 3.24: Đường cong ROC để xác định Threshold C1.....	47
Hình 3.25: Đường cong ROC để xác định Threshold C2.....	47
Hình 3.26: Sơ đồ thực thi thuật toán SVM.....	48
Hình 3.27: Đồ thị xác định CV của thuật toán SVM với dữ liệu trên cổ.....	49
Hình 3.28: Đồ thị xác định CV của thuật toán SVM với dữ liệu trên cổ và chân	50

DANH MỤC BẢNG BIỂU

Bảng 2.1: Ma trận chông chéo.....	31
Bảng 3.1: Định dạng chuỗi dữ liệu gửi từ thiết bị	33
Bảng 3.2: Thông số Vi điều khiển Dspic33EP64M và module nRF24L01	36
Bảng 3.3: Số SVs phân bổ tại các lớp với dữ liệu trên cổ.....	48
Bảng 3.4: Giá trị b cho hàm quyết định với dữ liệu cổ	49
Bảng 3.5: Số SVs phân bổ tại các lớp với dữ liệu trên cổ và chân	49
Bảng 3.6: Giá trị b cho hàm quyết định với dữ liệu cổ và chân	50
Bảng 3.7: Đánh giá hiệu năng của các thuật toán	50

DANH MỤC VIẾT TẮT

SVM	Support Vector Machine	Máy vector hỗ trợ
SVMs	Support Vector Machines	Các máy vector hỗ trợ
SVs	Support Vectors	Các Vector hỗ trợ
CV	Cross-Validation	Đánh giá chéo
HMM	Hidden Markov Models	Mô hình Markov ẩn
WSN	Wireless Sensor Network	Mạng cảm biến không dây
ANN	Artificial Neural Network	Mạng nơ ron nhân tạo
NN	Neural Network	Mạng nơ ron
K-NN	K-nearest Neighbor	Lân cận gần nhất
TPR	True Positive Rate	Tỉ lệ dương tính đúng
FPR	False Positive Rate	Tỉ lệ dương tính sai
TP	True Positive	Dương tính đúng
TN	True Negative	Âm tính đúng
FP	False Positive	Dương tính sai
FN	False Negative	Âm tính sai
RFID	Radio Frequency Identification	Nhận biết tần số vô tuyến điện
DAG	Directed acyclic graph	Đồ thị không tuần hoàn có hướng
DBA	Dynamic Body Acceleration	Gia tốc cơ thể động
DBA_x	Dynamic Body Acceleration of x-axis	Gia tốc cơ thể động trục x
DBA_y	Dynamic Body Acceleration of y-axis	Gia tốc cơ thể động trục y
DBA_z	Dynamic Body Acceleration of z-axis	Gia tốc cơ thể động trục z
VeDBA	Vector of Dynamic Body Acceleration	Vector gia tốc cơ thể động
SCAY	Static Component of the Acceleration in the Y-axis	Thành phần gia tốc tĩnh của trục y
ROC	Receiver Operating Characteristic	
OCR	Optical Character Recognition	Nhận dạng ký tự quang học
ADC	Analog to Digital Converter	Bộ chuyển đổi tương tự sang số
UART	Universal asynchronous receiver/transmitter	Truyền nhận nối tiếp không đồng bộ
I2C	Inter-Integrated Circuit	Mạch tích hợp

DANH MỤC KÝ HIỆU

Ký hiệu	Đơn vị
VeDBA	m/s^2
SCAY	m/s^2
DBA_x	m/s^2
DBA_y	m/s^2
DBA_z	m/s^2

MỞ ĐẦU

Việc phân tích hành vi là một trong những cách thức thường được sử dụng trong chăm sóc động vật. Ngày nay, công nghiệp thực phẩm vẫn là một lĩnh vực được xã hội rất quan tâm như công nghiệp sữa, thực phẩm và chăn nuôi gia súc. Nó không chỉ được quan tâm về mặt số lượng mà còn là chất lượng. Tại Việt Nam, có một số công ty sữa lớn như: TH Truemilk, VINAMILK, ... Để nâng cao năng suất và chất lượng sản phẩm, họ quan tâm đến vấn đề sức khỏe của bò. Vì vậy, nhu cầu giám sát cá nhân và chăn nuôi gia súc từ quan điểm thể chất và sinh lý phát sinh từ bản chất của những khó khăn liên quan đến việc quản lý trang trại với các khu chăn thả lớn. Các nhà chăn nuôi luôn luôn có nhu cầu "giám sát" động vật của họ càng thường xuyên càng tốt. Chú ý đến chăm sóc động vật, cho dù đó là vấn đề sức khỏe hay cách chăm sóc có thể làm giảm năng suất và các tổn thương trên vật nuôi.

Chăn nuôi theo hình thức nhỏ lẻ thì vấn đề chăm sóc và giám sát vật nuôi thường khá đơn giản. Tuy nhiên, với những trang trại có những khu chăn thả lớn và bò là di động, họ không thể quan sát đúng thông qua việc nhìn. Vì vậy, một số hệ thống giám sát và phân loại được thiết kế của để theo dõi hành vi của bò. Thiết bị được đeo trên cổ bò, và sử dụng dữ liệu cảm biến gia tốc để có được từ hệ thống để phân loại hành vi của bò. Một số phương pháp gần đây đề xuất để phân loại hành vi tự động ở động vật chủ yếu dựa trên các thuật toán học máy khác nhau như cây quyết định, k-means, SVM và HMM.

Trong luận văn này, hệ thống giám sát bò được thiết kế dựa trên kỹ thuật mạng cảm biến không dây Zigbee. Hệ giám sát được xây dựng gồm hệ cơ sở dữ liệu giúp lưu trữ trạng thái trực tiếp của bò, một máy tính cỡ nhỏ giúp xử lý và đưa ra trạng thái của bò, đây cũng đóng vai trò là nút điều phối cho mạng cảm biến, và các nút mạng được gắn trên bò gồm 02 thiết bị thu dữ liệu chuyển động gồm một thiết bị gắn trên chân và một thiết bị được gắn trên cổ bò. Thuật toán nhận dạng và phân loại hành vi của bò được sử dụng trong hệ thống là hai thuật toán học máy: thuật toán cây quyết định và thuật toán SVM. Luận văn đưa ra đánh giá cụ thể về hiệu năng của hai thuật toán này.

Luận văn được chia thành 4 phần như sau:

Chương 1: Tổng quan về các hệ thống giám sát và nhận dạng hành vi trên bò

Chương này sẽ tóm lược về các hệ thống giám sát bò và kỹ thuật được sử dụng trong việc phân loại hành vi trên bò.

Chương 2: Thuật toán học máy giúp nhận dạng hành vi trên bò

Thuật toán học máy sử dụng cho nhận dạng và phân loại hành vi trên bò sẽ được thảo luận và phân tích tại chương 2.

Chương 3: Thiết kế hệ thống và đánh giá hiệu năng thuật toán

Trình bày quy trình thiết kế chi tiết về hệ thống giám sát hành vi trên bờ được đề xuất trong luận văn. Thực thi thuật toán cây quyết định và SVM cũng như đánh giá về hiệu năng của các thuật toán nhận dạng sẽ được trình bày ở nội dung chương này.

Kết luận

Tóm tắt những kết quả mà đã đạt được trong luận văn, đồng thời chỉ ra những mặt hạn chế cần phải cải tiến các sản phẩm trong công trình.

TỔNG QUAN VỀ CÁC HỆ THỐNG GIÁM SÁT VÀ NHẬN DẠNG HÀNH VI TRÊN BÒ

1.1. Giới thiệu

Ngày nay, nhu cầu về thực phẩm ngày càng cao, không những đòi hỏi về số lượng mà còn về chất lượng. Trước đó, việc chăn nuôi gia súc theo hộ gia đình là rất phổ biến với các hộ gia đình và các nông trại trên toàn thế giới, với quy mô chăn nuôi nhỏ thì việc giám sát tình trạng sức khỏe, sinh sản vật nuôi sẽ rất đơn giản và theo kinh nghiệm của người chăn nuôi. Tuy nhiên, với những trang trại hay doanh nghiệp cần chăn nuôi với số lượng lớn gia súc phục vụ cho mục đích lấy thịt và sữa hay lấy con giống...thì việc giám sát trở nên rất khó khăn nếu áp dụng giám sát theo hình thức thủ công. Hình thức giám sát phổ biến nhất hay được dùng là sử dụng mã vạch được gắn trên vật nuôi và ghi chép lại thông tin của từng con theo ngày, tuần, tháng... Cách thức này gây nhiều bất tiện cho người chăn sóc. Một trong những hệ được phát triển theo hình thức hiện đại hơn đó là hệ giám sát tự động tình trạng sức khỏe và các thông số đặc thù theo mong muốn từ người chăn nuôi. Các hình thức giám sát hiện đại giúp người chăn nuôi gia súc tiết kiệm được đáng kể thời gian giám sát và nâng cao tính hiệu quả trong việc chăn sóc và phúc lợi cho vật nuôi.

Sức khỏe và phúc lợi chung của gia súc thường có thể được kiểm chứng và xác định theo mô hình hành vi của bò. Hành vi vật lý của bò được báo cáo như là một cách phát hiện sớm các bệnh như bệnh tim và chỉ thị về đau đớn, stress nhiệt và tương tác xã hội trong một đàn. Hành vi thay đổi khi động vật bị ốm có thể bao gồm giảm hoạt động thường ngày, hoạt động sinh sản, ăn uống, chăm sóc và các hành vi xã hội khác [12].

Việc giám sát bò được tập chung theo ba hướng phổ biến nhất đó là giám sát phân loại hành vi của bò; phát hiện chấn thương; xác định thời điểm sinh sản của bò.

1.2. Giới thiệu hệ thống giám sát bò

Việc phát triển các hệ thống giám sát về sức khỏe và phúc lợi của bò để được tập trung theo ba hướng chính là nhận dạng và phân loại hành vi, phát hiện chấn thương, và xác định thời điểm sinh sản của bò. Có rất nhiều hệ giám sát bò được đề xuất và đã có hệ được thương mại hoá phục vụ cho các mục đích trên. Các hệ giám sát được phát triển theo nhiều hướng khác nhau như: giám sát theo thiết bị độc lập được áp dụng cho từng cá thể, dữ liệu được gửi về một thiết bị cầm tay khi có yêu cầu hoặc được lưu vào bộ nhớ của thiết bị, dữ liệu được tổng hợp theo ngày hoặc theo tuần... Hình thức khác là giám sát thời gian thực theo mạng, mỗi thiết bị giám sát trạng thái của bò sẽ đóng vai trò như là một nút mạng, dữ liệu sẽ được truyền từ nút mạng về trung tâm tức thời, việc giám sát này mang lại nhiều ưu việt và hiệu quả hơn trong việc giám sát so với các hệ trước đó.

Với những phương thức giám sát trên, có nhiều báo cáo được công bố cũng như sản phẩm được thương mại hoá. Trong bài báo [1] nhóm tác giả đã xây dựng hệ giám sát và phân loại hành vi trên động vật nhằm nâng cao khả năng chăm sóc trong hệ thống sản xuất thực phẩm hiện đại. Báo cáo cũng chỉ ra, hành vi của động vật cung cấp thông tin thực tế về sức khỏe và phúc lợi của động vật. Trong báo cáo [3], nhóm tác giả đã xây dựng hệ giám sát bò sử dụng công nghệ RFID giúp định danh bò và giám sát ở các nông trại chăn nuôi gia súc nhỏ, công nghệ này thay thế cho kỹ thuật giám sát thủ công trước đó là sử dụng công nghệ mã vạch. Bên cạnh đó, có rất nhiều các thiết bị giám sát đã được thương mại hoá ở các tính năng sử dụng cơ bản như thiết bị giám sát của công ty Afimilk. Công ty cung cấp ra thị trường 2 thiết bị giám sát là Afimilk Silent Herdman và AfiAct II, sản phẩm cung cấp khả năng tốt hơn về các thời điểm thay đổi nhiệt độ của bò giúp tăng tỉ lệ thụ tinh, và có khả năng giám sát sức khỏe như phân tích về thời gian ăn, nghỉ và sự kích thích động dục giúp giám sát sức khỏe bò hiệu quả. Công ty Cattle Watch cung cấp hệ thống giám sát gồm các tính năng: giám sát vật nuôi từ xa, xác định vị trí và khu vực di chuyển của vật nuôi, cảnh báo sớm về những rủi ro.

Có rất nhiều kỹ thuật trong việc xây dựng hệ giám sát vật nuôi như kỹ thuật dựa trên mạng cảm biến không dây (WSN), các loại cảm biến xác định nhiệt, cảm biến chuyển động, xác định tọa độ và kỹ thuật xử lý ảnh... Đã có rất nhiều báo cáo đề cập đến các kỹ thuật mạng cảm biến không dây trong giám sát như trong các báo cáo [1, 4, 5, 11, 15, 16], các nhóm tác giả đã xây dựng hệ giám sát dựa trên kỹ thuật mạng cảm biến không dây, một trong những mạng tiêu biểu là mạng Zigbee. Mạng cảm biến bao gồm

các nút mạng có vai trò là thiết bị được gắn trên vật nuôi, được kết nối với nút mạng điều phối, tất cả dữ liệu từ nút mạng sẽ được gửi về nút điều phối và được quản lý tại đây. Một trong những đặc điểm của mạng cảm biến không dây là: tần số hoạt động 2,4Ghz, tốc độ bit 256kbps, tiêu thụ công suất thấp [6, 20], và mạng cảm biến không dây mang lại một mức độ mới về khả năng ứng dụng trong lĩnh vực giám sát vật nuôi với khả năng tăng độ phân giải không gian và thời gian của dữ liệu đo [5]. Công nghệ WSN này góp phần nâng cao chất lượng cuộc sống, năng suất, an toàn và chất lượng cho cuộc sống thông qua các ứng dụng của nó trong các ngành công nghiệp khác nhau như phân phối, hậu cần, xây dựng, giao thông, quân sự và dịch vụ y tế [11].

Việc nhận dạng và phân loại hành vi của động vật dựa trên chuyển động được áp dụng khá phổ biến, như trong báo cáo [7, 10] các tác giả thực hiện phân loại hành vi của động thông qua việc phân tích các chuyển động trên cỏ bò hay theo dõi bản đồ di chuyển của nó. Một trong những cảm biến chuyển động được dùng phổ biến nhất trong các nghiên cứu và ứng dụng là cảm biến gia tốc. Trong các báo cáo [4, 6, 8, 9, 13, 14, 16] các nhóm tác giả đã sử dụng cảm biến gia tốc hai, ba trục để nhận dạng các chuyển động của đối tượng vật nuôi được quan sát. Trong báo cáo [4, 6, 14], tác giả đã sử dụng cảm biến gia tốc ba trục được gắn trên cổ hoặc chân bò để nhận diện hành vi của bò. Báo cáo [2] sử dụng cảm biến gia tốc được đeo trên cổ bò, đã chỉ ra khả năng nhận dạng 8 hành vi như đứng, nằm, ăn, đi, đứng lên, nằm xuống, chấn thương chân lúc đi và trạng thái kích thích. Tuy nhiên, về mức độ chính xác và độ tin cậy còn thấp. Báo cáo [6] sử dụng cảm biến gia tốc ba trục được đeo ở chân bò, báo cáo chỉ đưa ra khả năng nhận dạng được hai hành vi là đi và đứng của bò. Trong báo cáo [14], tác giả xây dựng thuật toán nhận dạng hai hành vi của bò theo thời gian thực là hành vi đứng và ăn dựa trên ngưỡng của cảm biến gia tốc được gắn trên cổ bò. Bên cạnh đó, các báo cáo [8, 9], nhóm tác giả cũng xây dựng hệ giám sát và phân loại hành vi của lợn dựa trên cảm biến gia tốc. Như vậy, cảm biến gia tốc được sử dụng rất phổ biến trong việc xây dựng hệ giám sát và phân loại hành vi trên động vật.

1.3. Giới thiệu về thuật toán học máy

Học máy là một lĩnh vực của khoa học máy tính, theo Arthur Samuel vào năm 1959, "máy tính có khả năng học hỏi mà không cần được lập trình một cách rõ ràng." [17] Phát triển từ nghiên cứu về nhận dạng mẫu và lý thuyết học tính toán trong trí tuệ nhân tạo. Học máy xây dựng các thuật toán có thể học hỏi và thực hiện các dự đoán về dữ liệu [18,19], các thuật toán như vậy vượt qua các hướng dẫn chương trình nghiêm ngặt bằng cách dự đoán dữ liệu hoặc quyết định thông qua xây dựng một mô hình từ mẫu đầu vào [20]. Học máy được sử dụng trong một loạt các tác vụ điện toán khi thiết kế và lập trình các thuật toán rõ ràng với hiệu năng tốt là khó hoặc không khả thi; Các ví dụ ứng dụng bao gồm lọc email, phát hiện các kẻ xâm nhập mạng hoặc những người

trong nội bộ đang làm việc để phá vỡ dữ liệu, [21] nhận dạng ký tự quang học (OCR), [22] học để xếp hạng và tầm nhìn máy tính.

Học máy liên quan chặt chẽ và thường chồng lấp đến thống kê tính toán, tập trung vào việc dự đoán bằng cách sử dụng máy tính. Nó có quan hệ chặt chẽ với việc tối ưu hóa toán học, cung cấp các phương pháp, lý thuyết và các lĩnh vực ứng dụng cho lĩnh vực này. Việc học máy đôi khi được kết hợp với việc khai thác dữ liệu [23], trong đó lĩnh vực nhỏ thứ hai tập trung nhiều hơn vào phân tích dữ liệu thăm dò và được biết đến là học không giám sát. [24] Học máy cũng có thể là giám sát [25] và được sử dụng để tìm hiểu và thiết lập hồ sơ hành vi cơ bản cho các thực thể khác nhau [26] và sau đó được sử dụng để tìm các dị thường có ý nghĩa.

Trong lĩnh vực phân tích dữ liệu, học máy là một phương pháp được sử dụng để đưa ra các mô hình phức tạp và các thuật toán cho phép dự đoán; Trong sử dụng thương mại, điều này được gọi là phân tích tiên đoán. Các mô hình phân tích này cho phép các nhà nghiên cứu, các nhà khoa học dữ liệu, các kỹ sư và các nhà phân tích "đưa ra các quyết định, kết quả đáng tin cậy, lặp lại" và khám phá những "cái nhìn sâu sắc ẩn giấu" thông qua việc học hỏi từ các mối quan hệ và xu hướng lịch sử trong dữ liệu [27].

1.3.1. Các loại học máy

Học máy thường được phân thành ba loại, tùy thuộc vào tính chất của "tín hiệu" hoặc "phản hồi" của hệ thống học. Gồm có:

- Học có giám sát (Supervised learning): Máy tính được thể hiện với các đầu vào ví dụ và các đầu ra mong muốn của nó, được cho bởi "người dạy" và mục tiêu là tìm hiểu một quy tắc chung cho phép ánh xạ các đầu vào cho đầu ra. [28]
- Học không giám sát (Unsupervised learning): Không có nhãn nào được đưa ra cho thuật toán học, mục đích là để nó tự tìm ra cấu trúc trong đầu vào của nó. Học tập không giám sát có thể là một mục tiêu tự nó (khám phá các mẫu ẩn trong dữ liệu). [28]
- Học tăng cường (Reinforcement learning): Một chương trình máy tính tương tác với một môi trường động, trong đó nó phải thực hiện một mục đích nhất định. Chương trình được cung cấp phản hồi về đầu ra khi nó điều hướng các vấn đề của nó. [28]

1.3.2. Học máy theo đầu ra cụ thể

Nếu xem xét đặc tính đầu ra của thuật toán học máy, học máy có thể được phân loại thành các loại:

- Trong phân loại, đầu vào được chia thành hai hoặc nhiều lớp, và người học phải tạo ra một mô hình gán các đầu vào không nhìn thấy cho một hoặc nhiều (phân loại nhiều loại) của các lớp này. Điều này thường được giải quyết theo cách giám sát. [28]
- Trong hồi quy, đầu ra là liên tục chứ không phải rời rạc. [28]
- Trong phân cụm, tập dữ liệu đầu vào được chia thành các nhóm. Không giống như trong phân loại, các nhóm không được biết trước, nó thường là một nhiệm vụ học không giám sát. [28]
- Ước lượng mật độ tìm thấy sự phân bố đầu vào trong một không gian. [28]
- Giảm số chiều làm đơn giản hoá đầu vào bằng cách ánh xạ chúng vào không gian thấp hơn. Mô hình hóa chủ đề là một vấn đề liên quan, trong đó một chương trình được cung cấp một danh sách các tài liệu ngôn ngữ của con người và có nhiệm vụ tìm ra tài liệu nào bao gồm các chủ đề tương tự. [28]

1.3.3. Các dạng học máy

Có rất nhiều dạng học máy được nghiên cứu và ứng dụng, có thể kể đến các thuật toán học máy dưới đây:

- Thuật toán cây quyết định (Decision tree learning): là một phương pháp không tham biến (không cần đưa vào dạng hàm phụ thuộc vector tham biến của hàm mục tiêu). Phương pháp này sẽ chia không gian vector đặc trưng thành các miền, rồi xấp xỉ hàm mục tiêu trên mỗi miền bởi một hàm đơn giản nhất, gọi là hàm hằng. Khi đó, toàn bộ hàm mục tiêu sẽ được xấp xỉ bởi một hàm có thể mô tả dưới dạng cấu trúc cây [36].
- Học theo quy tắc kết hợp (Association rule learning): là một phương pháp học dựa trên quy tắc để học tập các mối quan hệ thú vị giữa các biến trong các cơ sở dữ liệu lớn. Nó nhằm xác định các quy tắc mạnh mẽ được phát hiện trong cơ sở dữ liệu sử dụng một số biện pháp thú vị. [29]
- Thuật toán mạng neural nhân tạo (ANN): thường được gọi là "mạng nơ-ron" (NN), là một thuật toán học được lấy cảm hứng từ cấu trúc và các khía cạnh chức năng của mạng nơ-ron sinh học. Tính toán được cấu trúc theo một nhóm các neuron nhân tạo kết nối, xử lý thông tin bằng cách sử dụng một cách tiếp cận kết nối để tính toán. Mạng nơ-ron hiện đại là các công cụ mô hình hoá dữ liệu không tuyến tính. Chúng thường được sử dụng để mô hình các mối quan hệ phức tạp giữa đầu vào và đầu ra, để tìm các mẫu trong dữ liệu hoặc để nắm bắt cấu trúc thống kê trong phân bố xác suất chung không xác định giữa các biến quan sát được. [28]

- Học sâu (Deep learning): bao gồm nhiều lớp ẩn trong một mạng nơ-ron nhân tạo. Cách tiếp cận này cố gắng mô hình hóa cách não người xử lý hình ảnh và âm thanh bởi thị giác và thính giác. [28]
- SVMs là một tập hợp các phương pháp học tập có giám sát liên quan được sử dụng để phân loại và hồi quy. Với một tập hợp các ví dụ huấn luyện, được đánh dấu là thuộc một trong hai loại, thuật toán huấn luyện SVM xây dựng mô hình dự đoán xem một ví dụ mới có thuộc một loại hay không. [28]
- Phân cụm (Clustering): Phân cụm là sự phân chia của một tập các quan sát thành các tập con (gọi là các cụm) để các quan sát trong cùng một cụm tương tự theo một số tiêu chí hoặc tiêu chí được chuẩn bị trước, trong khi các quan sát được rút ra từ các cụm khác nhau thì không giống nhau. Các kỹ thuật phân cụm khác nhau đưa ra các giả định khác nhau về cấu trúc dữ liệu, thường được xác định bởi một số chỉ số tương tự và được đánh giá bởi sự tương tự giữa các thành viên của cùng một cụm và sự tách biệt giữa các cụm khác nhau. Các phương pháp khác dựa trên mật độ và kết nối đồ thị. Phân cụm là một phương pháp học không được giám sát và là một kỹ thuật phổ biến để phân tích dữ liệu thống kê. [28]
- Mạng Bayesian là một mô hình đồ thị xác suất (một loại mô hình thống kê) đại diện cho một tập các biến ngẫu nhiên và các phụ thuộc có điều kiện của chúng qua một đồ thị trực không hướng (DAG) [28]
- Học máy dựa trên nguyên tắc là một thuật ngữ chung cho bất kỳ phương pháp học máy nào xác định, học hỏi, hoặc phát triển “các quy tắc” để lưu trữ, vận dụng hoặc áp dụng kiến thức. Đặc điểm xác định của học máy dựa trên quy tắc xác định và sử dụng một tập hợp các quy tắc quan hệ đại diện cho kiến thức thu được bởi hệ thống. Điều này trái ngược với những học máy khác thường xác định một mô hình độc nhất có thể được áp dụng phổ quát cho bất kỳ trường hợp nào để dự đoán. Các phương pháp học tập dựa trên nguyên tắc máy học bao gồm các hệ thống phân loại học tập, học về quy tắc kết hợp và các hệ thống miễn dịch nhân tạo. [28].
- Bên cạnh đó còn một số thuật toán học máy khác như: thuật toán di truyền (Genetic algorithm), Học máy dựa trên nguyên tắc (Rule-based machine learning), hệ thống phân loại học (Learning classifier systems).

1.4. Kết luận chương 1

Chương này đã giới thiệu tổng quan về các hệ giám sát bò và vai trò của các hệ giám sát hiện đại trong việc giám sát sức khỏe và phúc lợi của gia súc, giúp nâng cao năng suất và khi quy mô chăn nuôi lớn. Tùy theo mục đích giám sát mà các hệ cũng được xây dựng theo mục đích phù hợp. Việc giám sát và phân loại hành vi gia súc có thể quy về việc giám sát và phân loại hành vi trên bò. Các kỹ thuật học máy được áp dụng rất phổ biến trong việc nhận dạng và phân loại đối tượng, các thuật toán được áp dụng tùy theo đặc trưng của việc phân loại.

THUẬT TOÁN HỌC MÁY GIÚP NHẬN DẠNG HÀNH VI TRÊN BÒ

2.1. Nhận diện hành vi trên bò

Một trong những vấn đề được quan tâm nhất trong những hệ giám sát và nhận dạng hành vi trên bò nói riêng và trên động vật nói chung là khả năng nhận dạng và phân loại, mức độ tin cậy và độ chính xác trong nhận dạng mỗi hành vi. Có rất nhiều thuật toán học máy được áp dụng trong việc nhận dạng và phân loại hành vi động vật như: thuật toán mạng neural nhân tạo, thuật toán SVM, thuật toán cây quyết định hay K-NN... Các thuật toán này được quy về hai phương pháp: phương pháp tham biến (hàm mục tiêu phụ thuộc vào các tham biến và việc học được quy về việc đánh giá các tham biến từ tập ví dụ huấn luyện được cung cấp [36]) và không tham biến. Trong thực tế, việc áp dụng thuật toán để nhận dạng và phân loại phụ thuộc vào đặc trưng đầu vào của đối tượng cần phân loại.

Đã có nhiều báo cáo đề cập đến việc sử dụng các thuật toán học máy trong việc nhận dạng và phân loại hành vi trên bò. Trong báo cáo [1], tác giả đã sử dụng thuật toán mạng neural nhân tạo đa lớp để phân loại 5 hành vi trên bò với độ chính xác trung bình là 76,2%. Báo cáo [2] đã sử dụng thuật toán SVM để phân loại 8 hành vi trên bò với độ chính xác trung bình là 78%. Trong báo cáo [12], tác giả đã sử dụng nhiều thuật toán phân loại gồm loại học có giám sát và học không có giám sát, báo cáo đã so sánh và chỉ ra hiệu năng của từng thuật toán. Trong bài toán phân loại này, thuật toán bagging ensembles with Tree learner cho hiệu năng tốt nhất. Báo cáo [16] đã sử dụng K-NN để phân loại tự động hành vi của bò từ dữ liệu gia tốc.

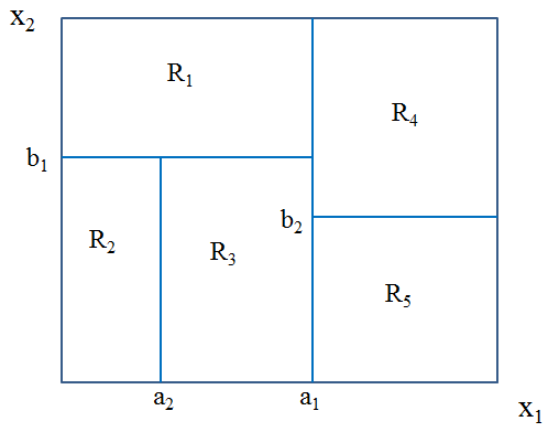
2.2. Thuật toán học máy cho nhận dạng hành vi trên bò

Trong luận văn này, tôi đề xuất dùng 2 thuật toán nhận dạng và phân loại hành vi trên bò là thuật toán cây quyết định và thuật toán SVM. Phần dưới đây sẽ trình bày cụ thể về cơ sở lý thuyết toán học học máy của hai thuật toán này.

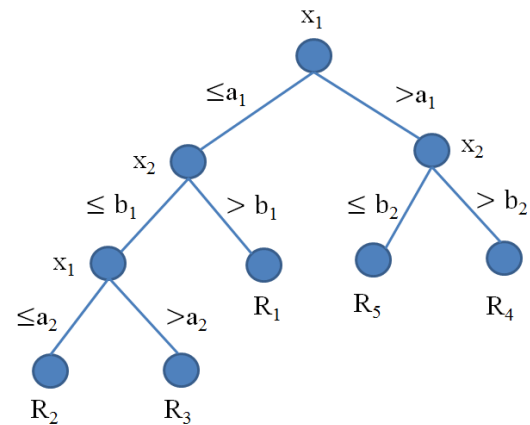
2.2.1. Thuật toán cây quyết định

Phương pháp học máy cây quyết định là chia không gian vector đặc trưng thành các miền, sau đó xấp xỉ hàm mục tiêu trên mỗi miền bởi một hàm hằng, chính là ngưỡng chia. Khi đó toàn bộ mục tiêu sẽ được xấp xỉ bởi một ngưỡng có thể mô tả dưới dạng cấu trúc cây. Cây quyết định thường được thực thi dưới dạng cây nhị phân.

Cây nhị phân có thể thực thi dưới dạng như sau, giả sử ta cần phân các đối tượng X được mô tả bởi vector đặc trưng trong không gian 2 chiều có dạng $X = (x_1, x_2)$. Ta chia không gian đặc trưng thành các hình chữ nhật bởi các đường thẳng song song với các trục tọa độ [28]. Đầu tiên, ta chia không gian thành 2 miền bởi đường thẳng song song với trục tung tại a_1 , tại đây ta có 2 miền, một miền các điểm X mà $x_1 > a_1$ và một miền có $x_1 \leq a_1$. Tiếp theo ta chia miền thứ nhất bởi đường thẳng song song với trục hoành tại điểm b_2 , ta nhận được 2 vùng chia là R_4 , các điểm X mà có $x_2 > b_2$, và vùng R_5 chứa các điểm X có $x_2 \leq b_2$. Tương tự cho các vùng còn lại. Kết thúc quá trình chia ta nhận được các vùng từ R_1 đến R_5 như hình 2.1. Quá trình phân chia không gian đặc trưng và xây dựng hàm phân lớp xấp xỉ như trên có thể biểu diễn như quá trình xây dựng cây nhị phân. Mỗi đỉnh biểu diễn một miền, đầu tiên gốc cây nhị phân biểu diễn toàn bộ không gian đặc trưng, vì chọn x_1 là biến chia, nên gốc cây được gán nhãn x_1 . Từ gốc có 2 nhánh đi ra, nhánh trái được gán nhãn $\leq a_1$ và nhánh phải có nhãn $> a_1$. Mỗi nhánh dẫn đến một đỉnh biểu diễn miền con gồm các điểm X mà $x_1 \leq a_1$. Miền này lại được chia đôi bởi biến x_2 tại điểm chia b_1 , do đó đỉnh con trái của gốc có nhãn là x_2 và nhánh trái của nó có nhãn là $\leq b_1$, nhánh phải có nhãn $> b_1$. Quá trình phát triển cây vẫn tiếp tục. Phát triển một đỉnh tương ứng ta chọn cho nó biến chia và điểm chia cho miền được biểu diễn bởi đỉnh đó. Quá trình xây dựng cây nhị phân của hình 2.1 được biểu diễn như hình 2.2.



Hình 2.1: Xác định ngưỡng theo theo đặc trưng



Hình 2.2: Mô hình cây quyết định

Sau khi xây dựng được mô hình cây, việc sử dụng cây cho tiên đoán các giá trị đến là rất đơn giản. Ta chỉ cần đi từ gốc xuống lá với đường đi là câu trả lời cho những phép so sánh tại mỗi đỉnh. Ta gọi cây như trên là cây quyết định. Trong thuật toán cây quyết định, mỗi đỉnh của cây biểu diễn một miền con của không gian đặc trưng, và phát triển một đỉnh là chọn một biến thuộc tính để chia miền con đó thành các miền con nhỏ hơn, gắn nhãn cho đỉnh đó bởi biến chia đã lựa chọn và xác định các nhánh đi ra từ đỉnh đó. Đó đó, để xây dựng cây quyết định ta cần:

- Chọn thuộc tính làm nhãn cho đỉnh cần phát triển.
- Thời điểm một đỉnh trở thành đỉnh lá của cây.
- Nếu một đỉnh là đỉnh lá của cây, thì chọn giá trị làm cho nhãn cho đỉnh lá đó.
- Xác định thời điểm dừng phát triển cây.

Trong quá trình phát triển cây quyết định, ta giải quyết vấn đề then chốt nhất là lựa chọn biến chia và điểm chia để chia một miền thành hai miền con. Trước tiên, ta có một đỉnh cây biểu diễn một miền con R – đỉnh lá của cây được xây dựng tại thời điểm hiện tại, ta cần phát triển đỉnh đó, ta cần chọn một thuộc tính x_k và một điểm chia a để chia miền R thành hai miền con R_l và R_r , khi đó đỉnh phát triển sẽ có đỉnh con trái biểu diễn miền con R_l và đỉnh con phải biểu diễn miền con R_r . Ta đánh giá hiệu quả của sự phân chia miền con R bởi cặp (x_k, a) thành hai miền con R_l, R_r bởi độ đo ký hiệu là $G(R, x_k, a)$, có dạng phương trình sau:

$$G(R, x_k, a) = \text{Impurity}(R) - (P_l \text{Impurity}(R_l) + P_r \text{Impurity}(R_r)) \quad (1)$$

Ở đây, P_l là tỷ số giữa ví dụ huấn luyện thuộc R_l trên số ví dụ huấn luyện thuộc R , tương tự ta cũng có với R_r , do đó:

$$P_l \text{Im purity}(R_l) + P_r \text{Im purity}(R_r) \quad (2)$$

là giá trị trung bình của các độ đo hỗn tạp của R_l và R_r .

Độ hỗn tạp (*Impurity*) là thông số đo độ thuần nhất của một miền con R . Một miền con R có độ hỗn tạp càng nhỏ thì nó càng thuần nhất (miền con R hoàn toàn thuần nhất khi tất cả các mẫu huấn luyện trong miền con này có cùng một nhãn lớp). Độ hỗn tạp được xác định bởi công thức:

$$\text{Im purity}(R) = -p_1 \log p_1 - p_2 \log p_2 \quad (3)$$

trong đó

$$p_1 = \frac{N_1}{N} \quad p_2 = \frac{N_2}{N}$$

N là số đối tượng cần phân lớp, N_1 là số đối tượng thuộc cần phân lớp sang phân lớp số 1, N_2 số đối tượng cần phân lớp sang phân lớp số 2.

Ta mong muốn sự phân chia rút bớt độ hỗn tạp nhất có thể, do đó ta lựa chọn biến chia x_k và điểm chia a sao cho $G(R, x_k, a)$ là lớn nhất.

Ta có thể dừng phát triển cây bởi việc ta sẽ dừng phát triển đỉnh biểu diễn miền con R – cho nó thành đỉnh lá cây, khi độ đo hỗn tạp của miền R nhỏ hơn một ngưỡng đã đưa ra do ta xác định trước.

2.2.2. Thuật toán SVM

Ý tưởng của thuật toán SVM là ánh xạ tập dữ liệu nguyên thủy vào không gian đặc trưng có số chiều lớn hơn số chiều ban đầu của dữ liệu nguyên thủy và tìm siêu phẳng tách tối ưu trong không gian mới này. Cuối cùng, ta tiến hành phân tách tối ưu trong không gian mới này. Ưu điểm của phương pháp này là nhận hàm quyết định trong không gian dữ liệu ban đầu (tương đương với siêu phẳng tối ưu trong không gian đặc trưng) không phải là hàm tuyến tính mà là hàm tổng quát hơn nào đó, có thể là đa thức bậc $k > 2$. Khi tập dữ liệu không phân tách được tuyến tính, nếu ánh xạ dữ liệu vào không gian đặc trưng mới, thì khả năng phân tách tuyến tính dữ liệu trong không gian mới là rất lớn [28]. Để làm rõ vấn đề này, trước hết ta xét bài toán phân tách dữ liệu thành 2 lớp.

2.2.2.1. Thuật toán SVM cho bài toán phân tách 2 lớp

Giả sử ta có M tập dữ liệu huấn luyện đầu vào trong không gian m chiều là X_i ($i = 1, \dots, M$) thuộc phân lớp 1 hoặc phân lớp 2. Nếu đầu ra y_i tương ứng với X_i là 1

thì X_i thuộc lớp 1, -1 thì X_i thuộc lớp 2. Nếu dữ liệu có thể phân tách tuyến tính, thì ta có thể xác định hàm quyết định như sau:

$$D(X) = W^T X + b \quad (4)$$

ở đây, W là vector m chiều, b là độ lệch cho tất cả $i = 1, \dots, M$. Với mỗi điểm X_i ta có:

$$W^T X_i + b \begin{cases} > 0, y_i = 1 \\ < 0, y_i = -1 \end{cases} \quad (5)$$

Vì theo giả thiết dữ liệu huấn luyện là phân tách tuyến tính nên không có dữ liệu nào thoả mãn $W^T X + b = 0$. Do đó, bất phương trình (5) có thể thay thế bởi:

$$W^T X_i + b \begin{cases} \geq 1, y_i = 1 \\ \leq -1, y_i = -1 \end{cases} \quad (6)$$

Ta có thể rút gọn thành:

$$y_i(W^T X_i + b) \geq 1, (i = 1, \dots, M) \quad (7)$$

Siêu phẳng phân tách các điểm X_i ($i = 1, \dots, M$) có dạng tổng quát:

$$D(X) = W^T X_i + b = c, (-1 < c < 1) \quad (8)$$

Khi $c = 0$ mặt phẳng phân tách nằm giữa 2 mặt phẳng khi $c = 1$ và $c = -1$. Khoảng cách giữa mặt phân tách và mẫu dữ liệu huấn luyện gần nhất với mặt phân cách gọi là

lề. Khoảng cách đó có độ lớn là: $\frac{1}{|W|}$

Mục đích của thuật toán học máy này là tìm tìm siêu phẳng tách có lề lớn nhất, do đó ta cần giải vấn đề tối ưu với các ràng buộc sau:

$$\min f(W) = \frac{1}{2}(W, W) \quad (9)$$

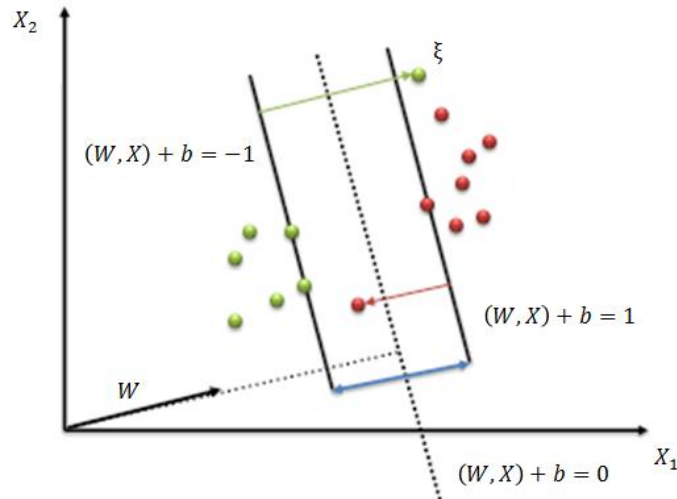
Với ràng buộc:

$$y_i(W^T X_i + b) \geq 1, (i = 1, \dots, M) \quad (10)$$

2.2.2.2. Thuật toán SVM với phân lề mềm

Tuy nhiên, thực tế tập dữ liệu huấn luyện không tách được tuyến tính trong không gian đặc trưng. Bên cạnh đó, tập huấn luyện có thể chứa các điểm ngoại lai, có thể có các mẫu bị phân lớp sai. Để xử lý vấn đề này, ta có thể đưa vào biến lỏng, biến siêu phẳng thành siêu phẳng mềm, mỗi biến tương ứng với một điểm huấn luyện:

$$\xi_i \geq 0, (i = 1, \dots, M) \quad (11)$$



Hình 2.3: Đường phân tách mềm của thuật toán SVM

Lúc này ta có ràng buộc mới

$$y_i [(W, X_i) + b] \geq 1 - \xi_i, (i = 1, \dots, M) \quad (12)$$

Do đó, nếu điểm dữ liệu huấn luyện (X_i, y_i) thỏa mãn bất phương trình trên với $0 \leq \xi_i \leq 1$ thì X_i được phân lớp chính xác bởi siêu phẳng, còn nếu ξ_i đủ lớn thì ràng buộc trên sẽ mặc nhiên được thỏa mãn. Lúc này ta cần giải quyết vấn đề tối ưu ràng buộc sau:

$$\min f(W, x) = \frac{1}{2}(W, W) + C \sum_{i=1}^N \xi_i \quad (13)$$

Với ràng buộc:

$$y_i [(W, X_i) + b] \geq 1 - \xi_i, (i = 1, \dots, M) \quad (14)$$

$$\xi_i \geq 0, (i = 1, \dots, M) \quad (15)$$

Trong đó, C là hằng số > 0 đã chọn trước, đóng vai trò như hệ số điều chỉnh trong hàm lỗi điều chỉnh.

Như vậy, siêu phẳng mềm là siêu phẳng $(W, X) + b = 0$, mà vector trọng số W được xác định là nghiệm của vấn đề tối ưu ràng buộc trên. Để tìm nghiệm của vấn đề tối ưu, ta xét hàm Lagrange tương ứng

$$L(W, x, b, a, \beta) = \frac{1}{2}(W, W) + C \sum_{i=1}^N \alpha_i - \sum_{i=1}^N \alpha_i \{y_i [(W, X_i) + b] - 1 + x_i\} - \sum_{i=1}^N \beta_i x_i \quad (16)$$

trong đó, $\alpha_i \geq 0$ và $\beta_i \geq 0$ ($i = 1, \dots, M$) là nhân tử Lagrange. Các điều kiện tương ứng là:

$$\begin{aligned} \alpha_i \{y_i [(W, X_i) + b] - 1 + x_i\} &= 0, (i = 1, \dots, M) \\ \beta_i x_i &= 0 \end{aligned} \quad (17)$$

Nghiệm của vấn đề tối ưu là điểm yên ngựa của hàm Lagrange, tại đó hàm cực tiểu theo W , ξ và b , cực đại theo α và β . Thông qua các đạo hàm riêng tương ứng ta có:

$$W = \sum_{i=1}^M \alpha_i y_i X_i \quad (18)$$

$$\sum_{i=1}^M \alpha_i y_i = 0 \quad (19)$$

$$\alpha_i = C - \beta_i, (i = 1, \dots, M) \quad (20)$$

Sau khi khử được W , ξ và b ta nhận được hàm Lagrange có dạng

$$L(a) = \sum_{i=1}^M \alpha_i - \frac{1}{2} \sum_{i=1}^M \sum_{j=1}^M \alpha_i \alpha_j y_i y_j (X_i, X_j) \quad (21)$$

Từ đó, vấn đề tối ưu được chuyển về tìm max của hàm (21) với ràng buộc

$$0 \leq \alpha_i \leq C \quad (22)$$

$$\sum_{i=1}^M \alpha_i y_i = 0 \quad (23)$$

Cuối cùng ta nhận được hàm quyết định có dạng

$$D(X) = \text{sign} \left(\sum_{i=1}^M a_i y_i (X_i, X) + b \right) \quad (24)$$

Ta thấy rằng các vector X_i ứng với các $\alpha_i = 0$ không đóng góp gì vào việc ra quyết định, nên để tiên đoán cho vector mới X , ta chỉ sử dụng các vector hỗ trợ với $\alpha_i > 0$. Do đó, ta có thể viết lại hàm quyết định như sau:

$$D(X) = \text{sign} \left(\sum_{i \in S} a_i y_i (X_i, X) + b \right), \quad (25)$$

với S là tập các chỉ số của các vector hỗ trợ. Ký hiệu R là tập các chỉ số của vector hỗ trợ tương ứng với các $0 < \alpha_i < C$, ta xác định b là giá trị trung bình lấy trên tập R ,

$$b = \frac{1}{N_R} \sum_{i \in R} \left\{ y_i - \sum_{j \in S} a_j y_j (X_j, X_i) \right\} \quad (26)$$

với N_R là số vector trong R .

2.2.2.3. Sử dụng hàm nhân trong thuật toán SVM

Khi tập dữ liệu không phân tách được tuyến tính, nếu ánh xạ dữ liệu vào không gian đặc trưng mới, thì khả năng phân tách tuyến tính dữ liệu trong không gian mới là rất lớn. Gọi hàm $\varphi(X)$ là hàm ánh xạ X sang không gian đặc trưng mới. Trong trường hợp này, ta sử dụng hàm nhân $k(\dots)$ thay thế cho việc tính toán nhân ánh xạ của X trong không gian đặc trưng mới. Khi đó tích $(\varphi(X_i), \varphi(X_j))$ trong không gian mới được thay thế bằng hàm $k(X_i, X_j)$. Do đó, hàm quyết định $D(X)$ và giá trị b trên được thay thế bằng:

$$D(X) = \text{sign} \left(\sum_{i=1}^M a_i y_i k(X_i, X) + b \right) \quad (27)$$

$$b = \frac{1}{N_R} \sum_{i \in R} \left\{ y_i - \sum_{j \in S} a_j y_j k(X_j, X_i) \right\} \quad (28)$$

Trên thực tế có 7 loại hàm nhân được sử dụng trong thuật toán học máy. Ví dụ, hàm uniform, triangle, Epanechnikov, quartic (biweight), tricube, triweight, Gaussian, quadratic and cosine. Trong luận văn này tôi sử dụng hàm nhân Gaussian (Radial Basis Function) có dạng như sau:

$$k(X_i, X_j) = e^{-g \|X_i - X_j\|^2}, g > 0 \quad (29)$$

2.2.2.4. Kiểm chứng chéo trong quá trình huấn luyện

Trong phần này ta đề cập đến kiểm chứng chéo (Cross-Validation - CV) trong việc huấn luyện mô hình phân lớp tối ưu. Kiểm chứng chéo được sử dụng để đo lỗi tổng quát của các phân loại cho một số lượng hạn chế các dữ liệu thu thập được. Trong kiểm chứng chéo, M dữ liệu đầu vào sẽ được chia thành 2 tập dữ liệu: tập dữ liệu huấn luyện S_i^{tr} gồm l dữ liệu huấn luyện và tập dữ liệu kiểm tra S_i^{ts} gồm $M - l$ dữ liệu còn lại. Sau đó việc huấn luyện trên tập dữ liệu huấn luyện sẽ được kiểm tra bởi tập dữ liệu kiểm tra. Việc này được lặp lại trên tất cả tập dữ liệu, và tổng tất cả tỉ số nhận dạng trên tập dữ liệu kiểm tra được tính toán như là ước lượng hiệu năng phân loại. Do sự tiêu tốn về thời gian trong công việc này, nên việc kiểm chứng chéo k-fold thường được sử dụng. [17]

Trong kiểm chứng chéo $k - fold$, dữ liệu huấn luyện thường được chia ngẫu nhiên thành k tập dữ liệu với số lượng xấp xỉ nhau, và phân loại sẽ sử dụng $k - 1$ tập và việc kiểm tra sẽ dùng tập còn lại. Việc huấn luyện sẽ được thực thi k lần, và tổng tỉ số nhận dạng cho tất cả k tập sẽ được tính toán. [17]

Việc kiểm chứng chéo sẽ giúp tìm ra cặp hệ số (C, γ) tốt nhất cho mô hình huấn luyện tương ứng với giá trị CV tốt nhất.

2.2.2.5. Thuật toán SVM một đối một cho phân tách đa lớp

Trong luận văn này, việc nhận dạng và phân loại hành vi của bò được xử lý bởi việc phân tách đa lớp. Thuật toán SVM có một số kỹ thuật cho phân tách đa lớp như: một đối tất cả, một đối một, tất cả đối một. Trong khuôn khổ luận văn này, tôi sử dụng kỹ thuật một đối một để xử lý vấn đề phân tách đa lớp trong phân loại hành vi của bò.

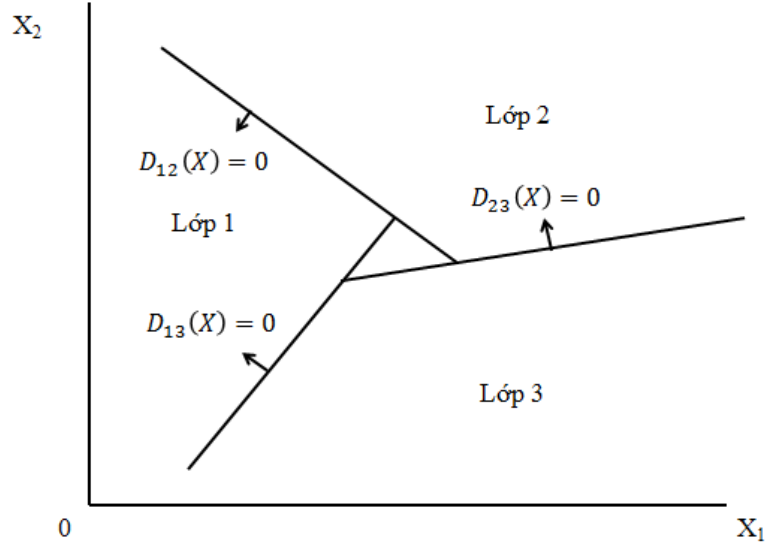
Trong kỹ thuật này ta xây dựng hàm quyết định theo từng cặp. Khi đó, với lớp thứ i và j ta xác định một hàm quyết định $D_{ij}(X)$ ($i \neq j, i, j = 1, \dots, M$), sao cho:

$$D_{ij}(X) > 0 \quad (30)$$

Khi X thuộc phân lớp i

$$D_{ij}(X) < 0 \quad (31)$$

Khi X thuộc phân lớp j



Hình 2.4: Thuật toán SVM một đối một

Với bài toán cần phân lớp cho M lớp ta cần $M(M - 1)/2$ hàm quyết định. Hàm quyết định (27) có thể viết lại thành hàm (32).

$$D_{ij}(X) = \text{sign} \left(\sum_{t=1, j \neq t}^M a_{ij} y_{ij} k(X_{ij}, X) + b_{ij} \right) \quad (32)$$

Ta có vùng không chồng chập

$$R_i = \left\{ X \mid D_{ij}(X) > 0, j = 1, \dots, M, j \neq i \right\} \quad (33)$$

Nếu X nằm trong vùng R_i , X được xem là thuộc lớp i . Có một vấn đề là X có thể không thuộc bất cứ vùng R_i nào. Khi đó, ta tiến hành phân lớp X bởi việc bầu chọn, với mỗi đầu vào X ta tính toán

$$D_i(X) = \underset{j=1, j \neq i}{\overset{M}{\mathop{\text{arg}}{\max}}} \text{sign}(D_{ij}(X)) \quad (34)$$

Và chúng ta sẽ phân lớp X thuộc lớp mà

$$\arg \max_{i=1, \dots, M} D_i(X) \quad (35)$$

2.3. Đánh giá hiệu năng của thuật toán học máy

Để đánh giá hiệu năng của thuật toán học máy, ta thường đánh giá ba thông số sau: độ nhạy, độ chính xác và độ chỉ rõ [47]. Các giá trị trên được xác định dựa trên ma trận chồng chéo sau:

Bảng 2.1: Ma trận chông chéo

	Chỉ định dương	Chỉ định âm
Dương tính thực tế	TP	FN
Âm tính thực tế	FP	TN

- **TP** (true positive): Số hành vi được phân lớp dương tính đúng so với quan sát hành vi dương tính thực tế
- **FN** (false negative): Số hành vi phân lớp dương tính nhưng nhầm thành phân lớp âm tính
- **FP** (false positive): Số hành vi phân lớp âm tính nhưng nhầm thành phân lớp dương tính
- **TN** (true negative): Số hành vi được phân lớp âm tính đúng so với quan sát hành vi âm tính thực tế

Khi đó, các giá trị độ nhạy (*sen*), độ chính xác (*pre*), và độ chỉ rõ (*spe*) được tính toán theo các công thức sau:

$$sen = \frac{TP}{TP + FN} \quad (36)$$

$$pre = \frac{TP}{TP + FP} \quad (37)$$

$$spe = \frac{TN}{TN + FP} \quad (38)$$

2.4. Kết luận chương 2

Như vậy, chương 2 đã trình bày chi tiết về cơ sở toán học cho 2 thuật toán là cây quyết định và thuật toán SVM với kỹ thuật phân tách đa lớp “một đối một” cho bài toán nhận dạng và phân loại hành vi trên bò. Chương này cũng đưa ra cách thức đánh giá hiệu năng của thuật toán học máy dựa trên 3 thông số: độ nhạy, độ chính xác và độ chỉ rõ.

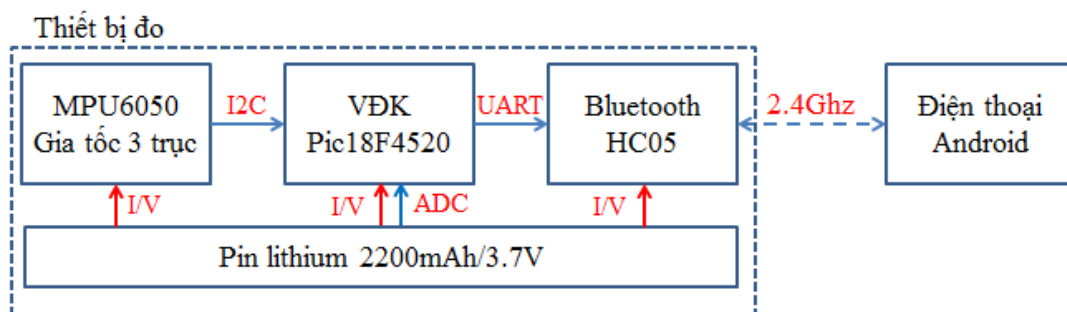
THIẾT KẾ HỆ THỐNG, THỰC THI VÀ ĐÁNH GIÁ THUẬT TOÁN

3.1. Thiết kế hệ thống

3.1.1. Thiết kế hệ thống thu thập dữ liệu huấn luyện

Trong vấn đề học máy có giám sát, việc huấn luyện mô hình dự đoán cần phải có tập dữ liệu huấn luyện thật chính xác và bao quát các trường hợp. Phần này trình bày cách thức xây dựng hệ thống thu nhận và gán nhãn một cách chính xác cho tập dữ liệu huấn luyện.

Hệ thống gồm hai cặp thiết bị giống nhau nhằm mục đích thu nhận dữ liệu gia tốc trên chân bò và trên cổ bò. Mỗi cặp thiết bị gồm một thiết bị đo gia tốc được kết nối không dây theo phương thức truyền Bluetooth và một điện thoại thông minh chạy trên hệ điều hành android. Hình 3.1 mô tả sơ đồ nguyên lý của hệ thống.



Hình 3.1: Sơ đồ nguyên lý của hệ thống thu dữ liệu mẫu

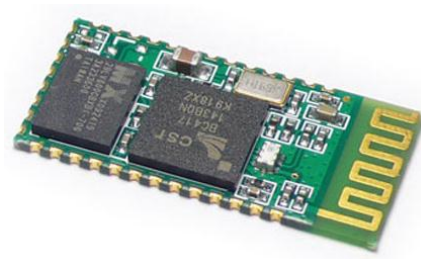
Thiết bị đo gồm 4 phần chính: Vi điều khiển, cảm biến gia tốc, bộ phận truyền thông không dây, và pin. Vi điều khiển trung tâm là pic18f4520 chức năng thu nhận, xử lý

tín hiệu, cảm biến gia tốc 3 trục MPU6050 thu nhận trạng thái chuyển động trên các vị trí tương ứng của bàn và trả về giá trị số, Bluetooth HC05 chịu trách nhiệm truyền dữ liệu thu được về điện thoại, và cuối cùng là pin lithium có dung lượng 2200mAh.

- Cảm biến gia tốc MPU6050 có thông số cơ bản như sau: Dải điện áp hoạt động từ 3-5V. Sử dụng bộ chuyển đổi ADC 16-bit. Gia tốc có thể cấu hình theo các giải đo: $\pm 2g$, $\pm 4g$, $\pm 8g$, $\pm 16g$. Cảm biến MPU6050 được giao tiếp với vi điều khiển thông qua chuẩn truyền thông I2C. Tốc độ lấy mẫu dữ liệu gia tốc là 1kHz. Hình 3.2 là hình ảnh về cảm biến gia tốc MPU6050.
- Bluetooth HC05 có thông số cơ bản như sau: Dải điện áp hoạt động từ 3-5V. Dòng điện khi HC05 hoạt động ổn định là 8mA, dải tần sóng hoạt động là 2.4GHz, HC05 có chuẩn Bluetooth là v2.0. Module giao tiếp với vi điều khiển thông qua chuẩn giao tiếp UART. Hình 3.3 là hình ảnh về Bluetooth HC05.



Hình 3.2: Cảm biến gia tốc 3 trục MPU6050



Hình 3.3: Module Bluetooth HC05

Điện thoại android sử dụng trong công việc này là điện thoại Sony Z3 được cài đặt sẵn ứng dụng để ghi nhận dữ liệu truyền về từ thiết bị đo. Ứng dụng gồm 2 chức năng chính là chức năng ghi nhận dữ liệu và chức năng ghi video thời gian thực. Hai tính năng này được đồng bộ thời gian với nhau giúp gán nhãn dữ liệu của sự kiện gửi về một cách chính xác. Việc gán nhãn cho dữ liệu sẽ giúp đánh giá hiệu năng của các thuật toán học máy sau này. Dữ liệu được ghi trên điện thoại gồm các trường:

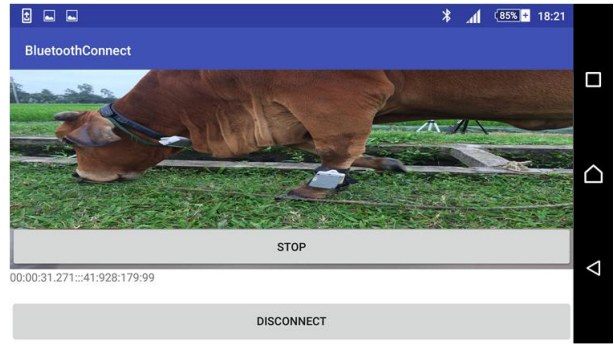
Bảng 3.1: Định dạng chuỗi dữ liệu gửi từ thiết bị

giờ: phút: giây	Gia tốc trục x	Gia tốc trục y	Gia tốc trục z	Giá trị pin (%)
-----------------	----------------	----------------	----------------	-----------------

Hình 3.4 và 3.5 là hình ảnh thiết bị đo và giao diện của ứng dụng trên điện thoại.



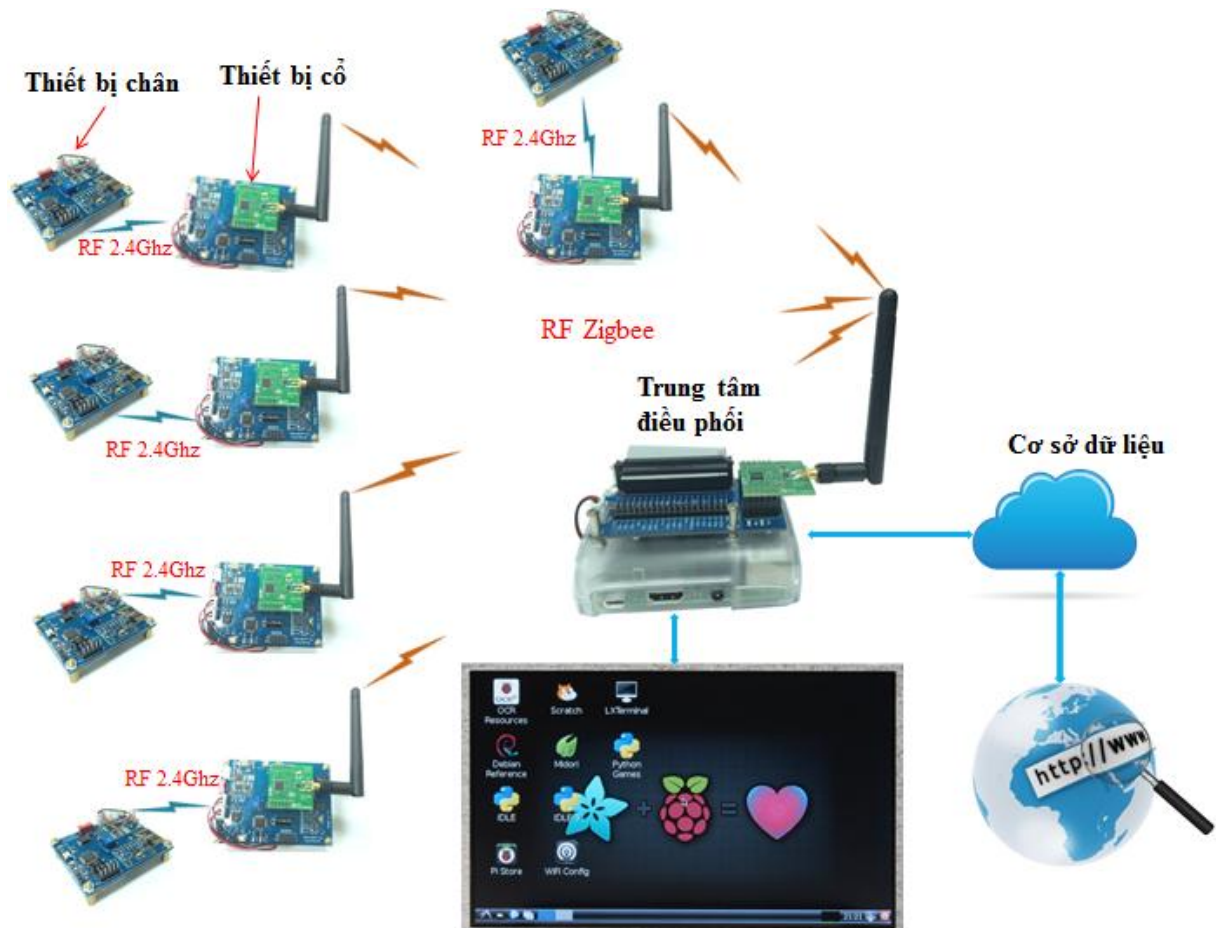
Hình 3.4: Bên trong thiết bị đo dữ liệu mẫu



Hình 3.5: Giao diện phần mềm thu dữ liệu mẫu

3.1.2. Thiết kế hệ thống giám sát

Luận văn này đề xuất việc xây dựng hệ thống giám sát và phân loại hành vi của bò. Hệ thống giám sát được thiết kế dựa trên kỹ thuật mạng cảm biến không dây Zigbee. Mỗi nút mạng gồm 2 thiết bị thu nhận dữ liệu chuyển động, một thiết bị thu dữ liệu chuyển động tại chân bò và một thiết bị thu dữ liệu chuyển động tại cổ bò. Trung tâm điều phối mạng là một mini máy tính Raspberry Pi 3, chịu trách nhiệm điều phối toàn mạng và thực thi thuật toán nhận dạng và phân loại hành vi của bò dựa trên dữ liệu chuyển động từ nút mạng gửi về. Tại đây, trạng thái của bò cũng được đẩy lên cơ sở dữ liệu online thông qua kết nối mạng tới trung tâm Raspberry. Người chăn nuôi có thể theo dõi bất cứ đâu thông qua giao diện website. Hình 3.6 là mô tả về hệ thống.

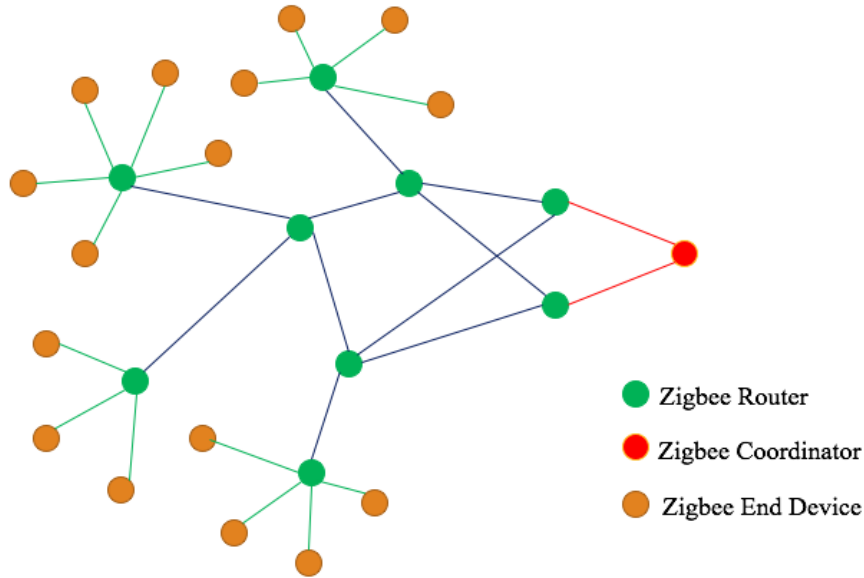


Hình 3.6: Mô hình hệ thống giám sát

3.1.2.1. Mô hình mạng truyền thông

Mô hình truyền thông của hệ giám sát được cấu hình là mạng truyền thông Zigbee dạng mạng Mesh. Mạng gồm nút điều phối (Zigbee Coordinator), các nút truyền vận (Zigbee Router) và các nút thiết bị cuối (Zigbee End Device). Hình 3.7 là cấu hình mạng giám sát được đề xuất trong luận văn. Thông số mạng như sau:

- Tần số sóng 2.4Ghz, có thể thay đổi 2405Mhz-2480Mhz, bước nhảy 5Mhz.
- Khoảng cách truyền nhận lý tưởng nút tới nút: 1.6Km
- Độ nhạy thu -110 dBm



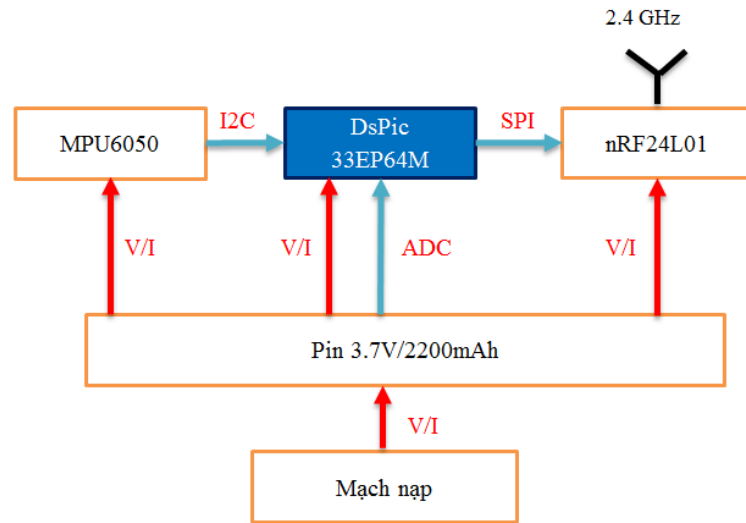
Hình 3.7: Mô hình mạng cảm biến

3.1.2.2. Thiết bị đo chuyển động trên chân

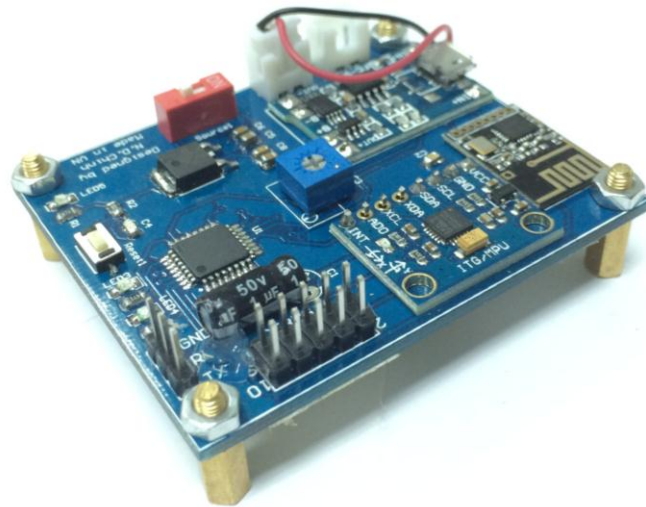
Thiết bị đo chuyển động trên chân bò sẽ đo thông số về gia tốc 3 trục tại chân, sau quá trình tiền xử lý, giá trị gia tốc sẽ được chuyển về thiết bị đo chuyển động trên cổ để hợp nhất dữ liệu trước khi gửi về nút điều phối. Hình 3.8 và hình 3.9 là sơ đồ nguyên lý của thiết bị và mạch thực tế. Thiết bị đo chuyển động trên chân gồm: 01 cảm biến gia tốc MPU6050, 01 Vi điều khiển Dspic33EP64M, 01 module truyền thông nRF24L01, 01 pin lithium 3.7V/2200mAh và 01 mạch sạc pin. Cảm biến gia tốc MPU6050 được kết nối với vi điều khiển Dspic33EP64M để đo và tiền xử lý dữ liệu gia tốc 3 trục tại chân. Dữ liệu sau đó được gửi về thiết bị trên cổ thông qua truyền thông RF trên module nRF24L01. Một vài thông số về vi điều khiển Dspic33EP64M và module nRF24L01 được chỉ ra trong bảng 3.2:

Bảng 3.2: Thông số Vi điều khiển Dspic33EP64M và module nRF24L01

Dspic33EP64M	Module nRF24L01
<ul style="list-style-type: none"> - Điện áp cấp 3-3.6V - Cấu trúc thanh ghi 16 bits - Tốc độ CPU: 70MIPS - Bộ nhớ chương trình 64 KB - Bộ nhớ RAM: 8192 Bytes - Dải nhiệt độ làm việc: -40 đến 125°C - Hỗ trợ các chuẩn giao tiếp: UART, 	<ul style="list-style-type: none"> - Tần số sóng: 2.4Ghz - Tiêu thụ năng lượng thấp - Điện áp cung cấp: từ 1.9 đến 3.6V - Dòng điện tiêu thụ: 11.3mA TX tại 0dBm và 13.5mA RX tại 2Mbps. - Tốc độ dữ liệu: 250kbps, 1Mbps và 2Mbps - Khoảng cách truyền khoảng 100m



Hình 3.8: Sơ đồ nguyên lý thiết bị đo gia tốc trên chân



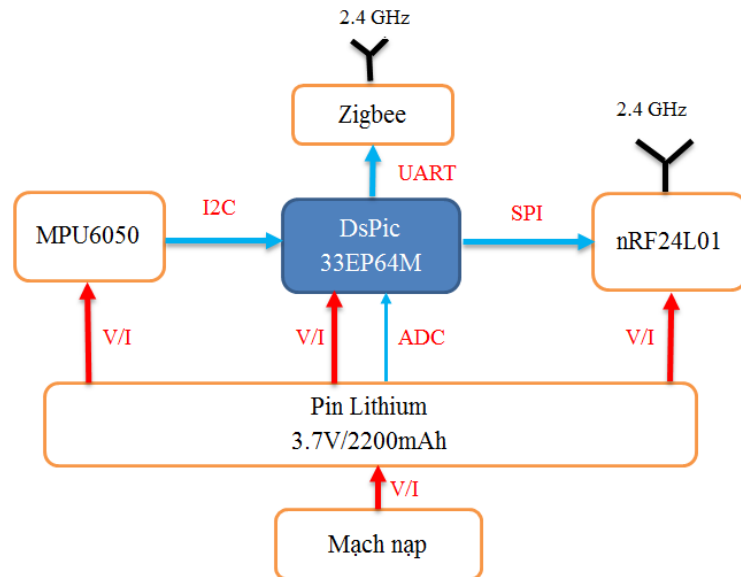
Hình 3.9: Thiết bị đo gia tốc trên chân

3.1.2.3. Thiết bị đo chuyển động trên cổ

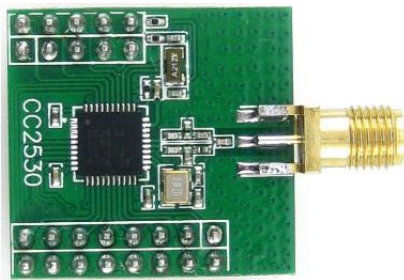
Thiết bị đo chuyển động trên cổ sẽ đo thông số về gia tốc 3 trục tại cổ bò, sau quá trình tiền xử lý, giá trị gia tốc tại cổ sẽ được đồng nhất với dữ liệu gia tốc tại chân để gửi về trung tâm điều phối. Thiết bị trên cổ gồm: 01 cảm biến gia tốc 3 trục MPU6050, 01 vi điều khiển Dspic33EP64M, 01 module truyền thông nRF24L01, 01 module Zigbee, 01 pin lithium 3.7V/2200mAh và 01 mạch nạp. Module Zigbee tại đây đóng vai trò là một thiết bị cuối (End Device) hoặc thiết bị chuyển vận (Router). Hình 3.10 là sơ đồ

nguyên lý của thiết bị trên cơ và hình 3.11 là hình ảnh thực về thiết bị. Hình 3.12 là hình ảnh về module Zigbee. Dưới đây là thông số của Module Zigbee:

- Tên: mạch thu phát RF Zigbee DRF1605H
- Dòng điện khi truyền lớn nhất là 110mA và khi nhận là 35mA.
- Chạy trên chip CC2530F256, 256Kflash
- Chuẩn truyền sóng Zigbee 2.4Ghz, có thể thay đổi 2405Mhz-2480Mhz, bước nhảy 5Mhz.
- Khoảng cách truyền nhận lý tưởng 1.6Km
- Độ nhạy thu -110 dBm
- Chuẩn kết nối với Vi điều khiển: UART



Hình 3.10: Sơ đồ nguyên lý thiết bị đo gia tốc trên cổ



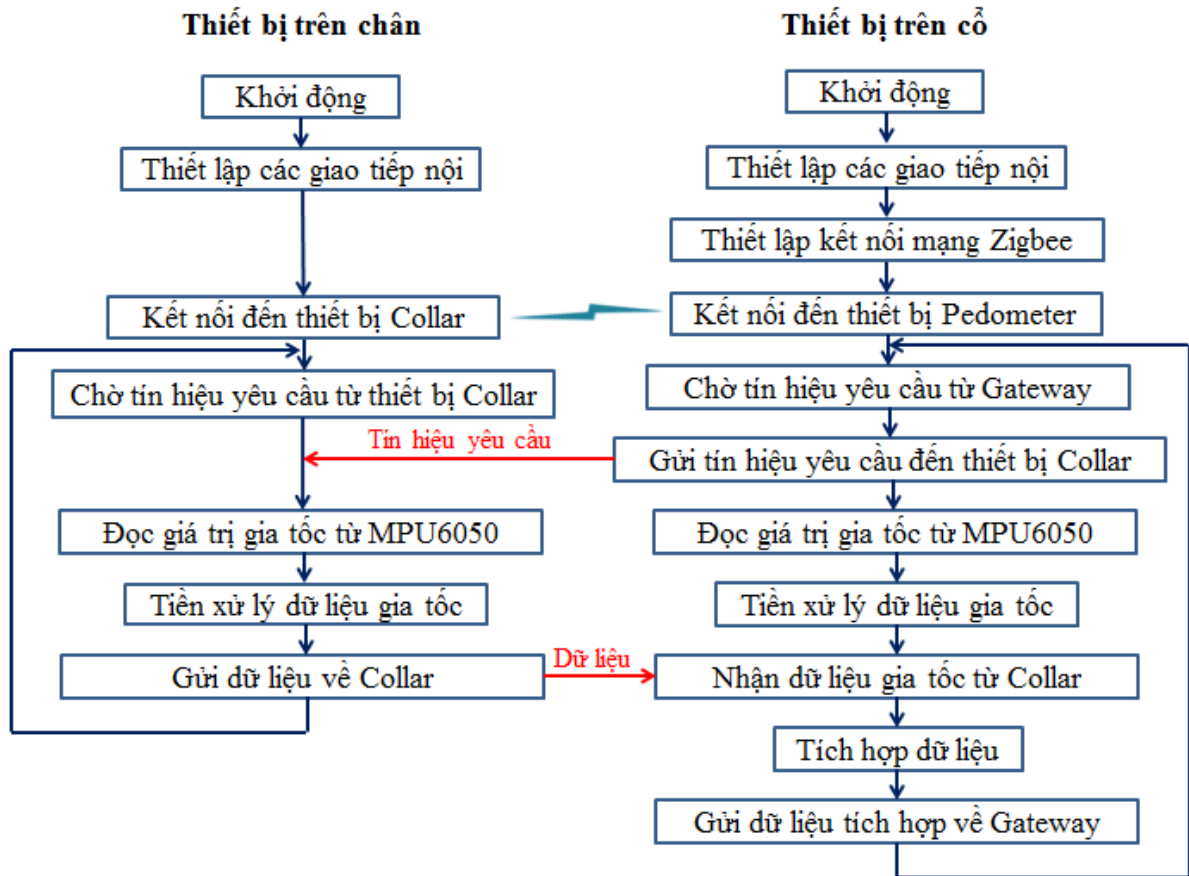
Hình 3.11: Module Zigbee



Hình 3.12: Thiết bị đo gia tốc trên cổ

3.1.2.4. Kết nối hai thiết bị chân và cổ

Dữ liệu gia tốc trên chân và cổ bò được xử lý truyền nhận một cách nhịp nhàng để gửi về trung tâm. Hình 3.13 là sơ đồ nguyên tắc kết nối và truyền nhận dữ liệu giữa 2 thiết bị đo chuyển động tại chân và cổ bò.



Hình 3.13: Sơ đồ truyền nhận dữ liệu giữa 2 thiết bị đo gia tốc trên chân và cổ bò

Hình 3.14 là hình ảnh sau khi 2 thiết bị được gắn trên bò:



Hình 3.14: Thiết bị được đeo trên bò

3.1.2.5. Trung tâm điều phối

Trung tâm điều phối là một máy tính nhỏ Raspberry Pi 3 được kết nối với một module truyền thông Zigbee với vai trò là nút điều phối. Dưới đây là thông số chi tiết của Raspberry Pi 3:

- Chạy trên hệ điều hành Raspbian OS
- Chíp Broadcom BCM2837 chạy ở xung nhịp 1.2 GHz
- 64-bit quad-core ARM Cortex-A53
- 802.11 b/g/n Wireless LAN
- Bluetooth 4.1 (Classic & Low Energy)
- Hai nhân Videocore IV® Multimedia co-processor
- Bộ nhớ RAM 1 GB LPDDR2
- Hỗ trợ tất cả bản phân phối cuối cùng ARM GNU/Linux và Windows 10 IoT
- Điện áp cấp 5V/ 2.5 A
- 1 x 10/100 Ethernet port
- 1 x HDMI video/audio connector
- 1 x RCA video/audio connector
- 4 x USB 2.0 ports
- 40 GPIO pins
- Chip antenna
- Kết nối màn hình DSI
- Khe thẻ nhớ MicroSD [30]

Dưới đây là hình ảnh của Raspberry Pi 3 và trung tâm điều phối sau khi kết nối Raspberry với module Zigbee.



Hình 3.15: Raspberry Pi 3



Hình 3.16: Trung tâm điều phối

3.2. Thực thi và đánh giá thuật toán

Trong luận văn này tôi sử dụng 2 thuật toán học máy để nhận dạng và phân loại hành vi trên bò là thuật toán cây quyết định và thuật toán SVM. Thuật toán cây quyết định được thực thi trên tập dữ liệu gia tốc 3 trục của thiết bị đo gia tốc trên cổ. Thuật toán SVM được chỉ làm 2 trường hợp thực thi: trường hợp 1 thực thi trên dữ liệu gia tốc 3 trục của thiết bị đo gia tốc trên cổ và trường hợp thứ 2 thực thi trên dữ liệu gia tốc 3 trục của cả 2 thiết bị đo gia tốc trên chân và trên cổ. Cuối cùng luận văn sẽ đánh giá hiệu năng của hệ thống với các trường hợp trên.

3.2.1. Thuật toán cây quyết định

Trong luận văn này, các hoạt động mức cao của bò là hoạt động ăn và đi, các hoạt động mức thấp của bò là hoạt động uống, đứng và nằm. Các hoạt động này được phân loại dựa trên 2 thông số VeDBA (gia tốc trung bình) và SCAY (thành phần gia tốc tĩnh theo trục y) bởi thuật toán cây quyết định. Dữ liệu gia tốc được dùng là gia tốc của thiết bị đeo trên cổ bò.

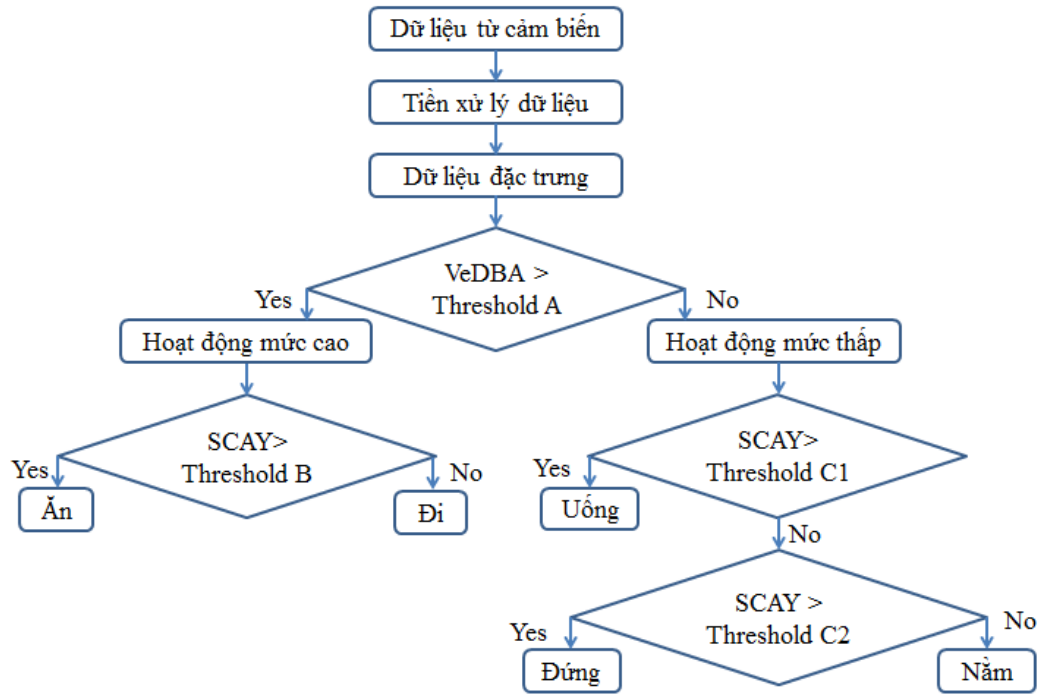
Giá trị VeDBA được xác định dựa trên công thức:

$$\text{VeDBA} = \sqrt{\text{DBA}_x^2 + \text{DBA}_y^2 + \text{DBA}_z^2} \quad (39)$$

Trong đó, giá trị DBA (gia tốc động) được tính bởi việc trừ gia tốc cho một lượng gia tốc tĩnh μ_x, μ_y, μ_z như dưới đây:

$$\begin{aligned} \text{DBA}_x &= A_x - \mu_x \\ \text{DBA}_y &= A_y - \mu_y \\ \text{DBA}_z &= A_z - \mu_z \end{aligned} \quad (40)$$

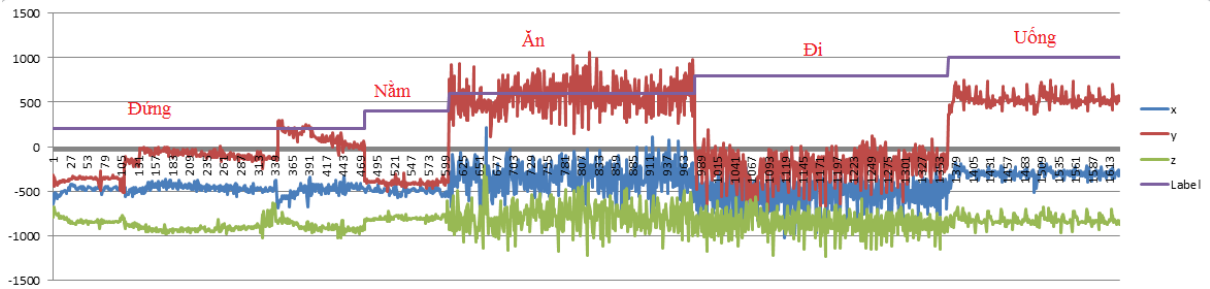
Hình 3.17 là sơ đồ thuật toán cây quyết định được dùng trong luận văn này. Thuật toán cây quyết định được thực thi với 4 ngưỡng: Threshold A, Threshold B, Threshold C1 và Threshold C2 để phân tách 5 hành vi của bò là đi, đứng, nằm, ăn và uống.



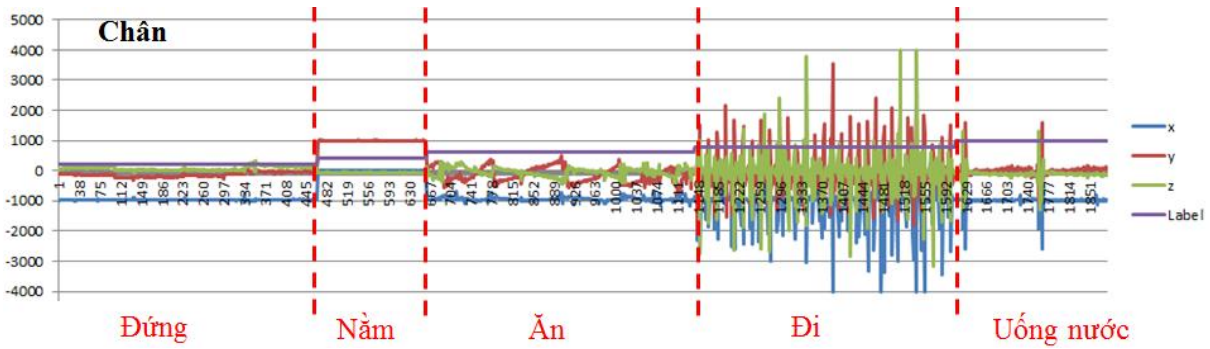
Hình 3.17: Sơ đồ thuật toán cây quyết định

- Tiền xử lý dữ liệu

Việc tiền xử lý dữ liệu được thực thi với bộ lọc trung bình, đây là bộ lọc được sử dụng khá phổ biến để làm mịn dữ liệu đầu vào. Cửa sổ kích thước M của bộ lọc được sử dụng để giảm nhiễu và giữ được hình dạng của tín hiệu. Hình 3.18 và hình 3.19 biểu diễn dữ liệu gia tốc thô từ cổ bò và chân bò cho 5 trạng thái: đi, đứng, nằm, ăn và uống nước trong thang đo giá trị gia tốc là mg ($1g = 1000mg$).

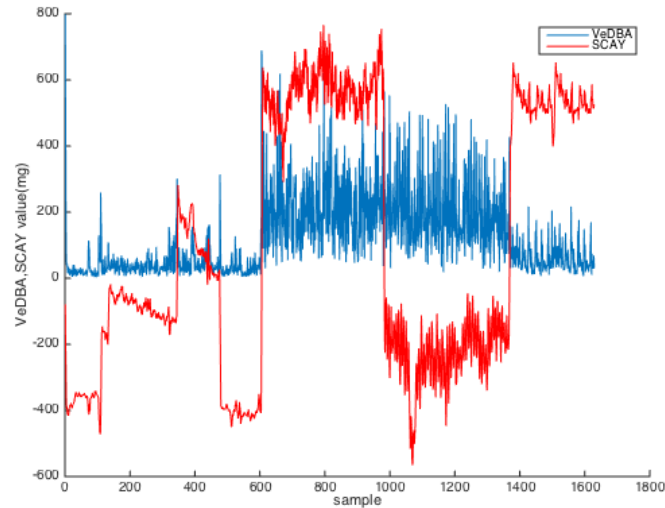


Hình 3.18: Dữ liệu gia tốc 3 trục tại cổ bò

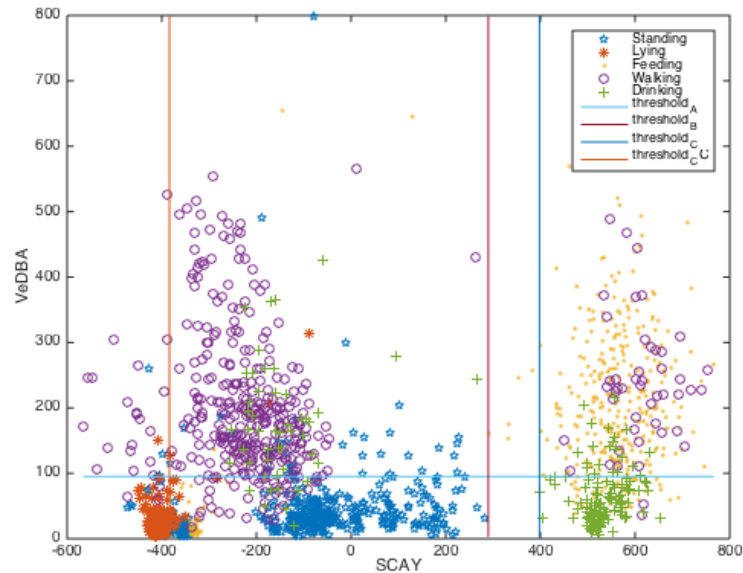


Hình 3.19: Dữ liệu gia tốc 3 trục trên chân bò

Hình 3.20 biểu diễn giá trị của VeDBA và SCAY của dữ liệu huấn luyện. Phân bố dữ liệu theo mối quan hệ giữa giá trị VeDBA và SCAY tại mỗi mẫu được biểu diễn trong hình 3.21. Hình 3.21 cũng biểu diễn các ngưỡng để phân tách các hành vi của bò.



Hình 3.20: Biểu diễn giá trị VeDBA và SCAY của dữ liệu huấn luyện



Hình 3.21: Biểu diễn phân bố dữ liệu theo mỗi quan hệ giữa VeDBA và SCAY của dữ liệu huấn luyện

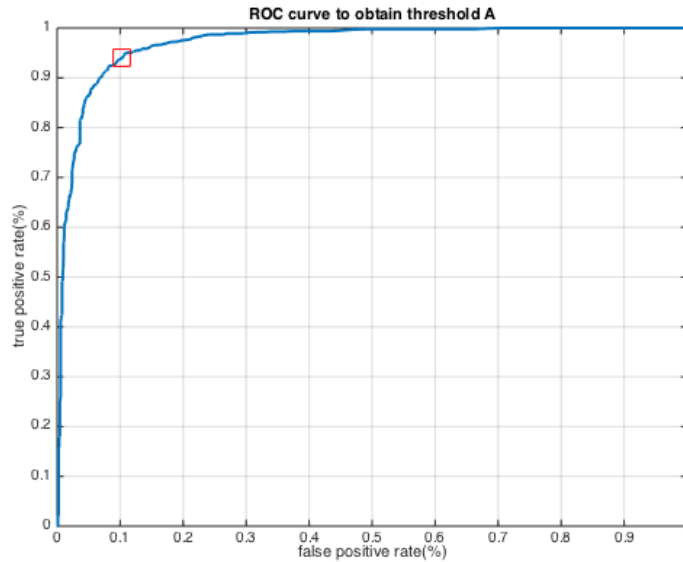
- Xác định các ngưỡng

Hình 3.17, giá trị VeDBA được so sánh với Threshold A để phân loại hành vi thành 2 loại là hoạt động mức cao và hoạt động mức thấp. Trong khi đó, giá trị SCAY được sử dụng để phân tách chi tiết các hành vi khi được so sánh với các Threshold B, Threshold C1 và Threshold C2. Các ngưỡng được chọn dựa trên đường cong ROC. Đường cong ROC được xác định dựa trên tập TPR và FPR [34].

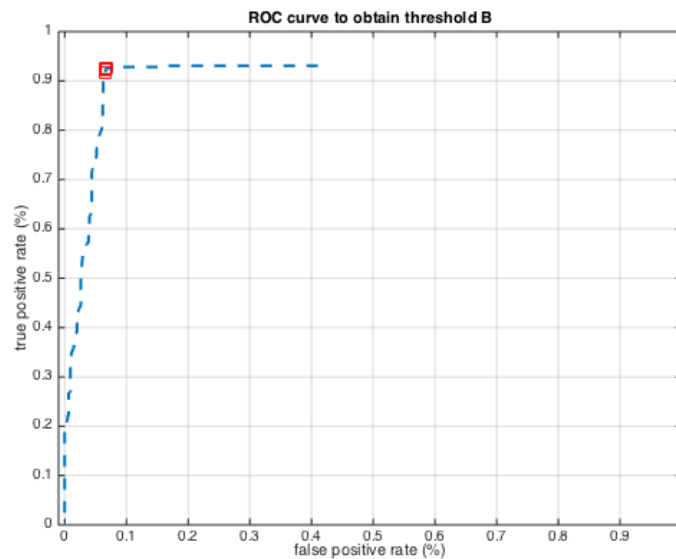
$$TPR = \frac{TP}{TP + FN} \quad (41)$$

$$FPR = \frac{FP}{FP + TN} \quad (42)$$

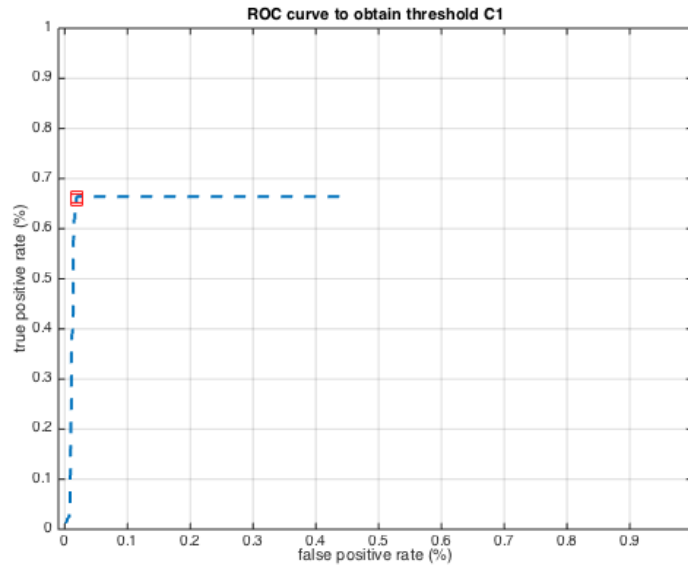
Trong luận văn này, Threshold A, Threshold B, Threshold C1 và Threshold C2 được chọn lần lượt là 75mg, 354mg, 404mg và -389mg. Việc chọn các ngưỡng chính là chọn điểm trên cùng phía bên trái của các đồ thị dưới đây, hình 3.22, 3.23, 3.24 và 3.25. Mục đích việc chọn là để có TPR là lớn nhất và FPR là nhỏ nhất. Điểm được chọn để xác định ngưỡng là vị trí ô vòng đỏ trên đồ thị.



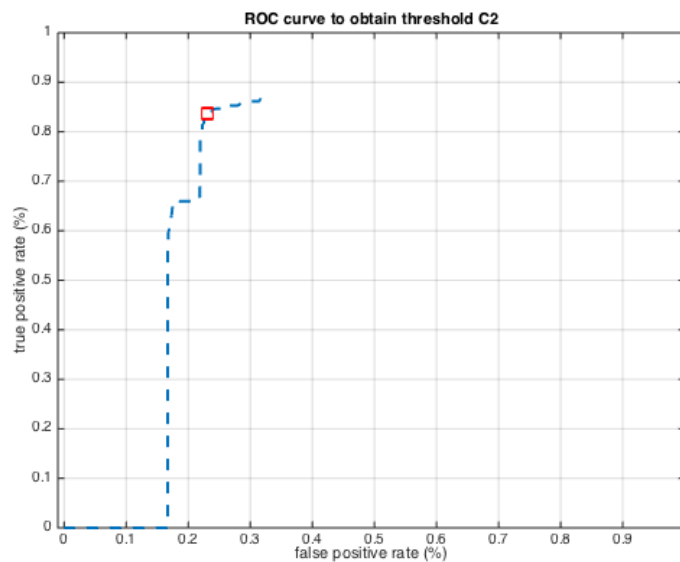
Hình 3.22: Đường cong ROC để xác định Threshold A



Hình 3.23: Đường cong ROC để xác định Threshold B



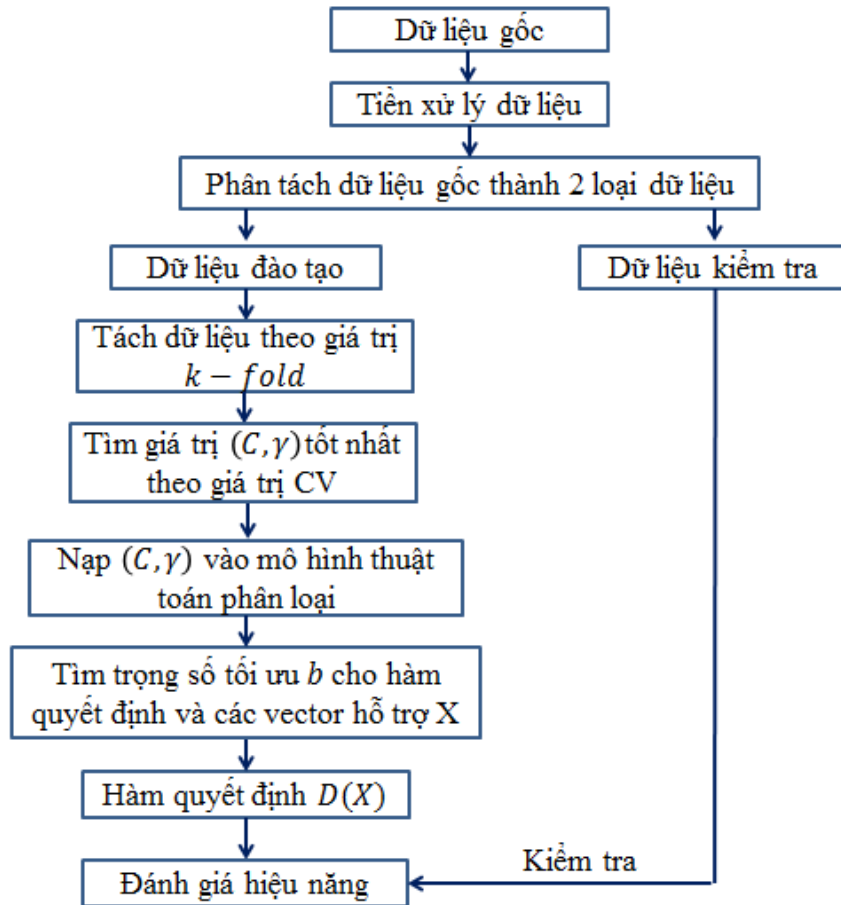
Hình 3.24: Đường cong ROC để xác định Threshold C1



Hình 3.25: Đường cong ROC để xác định Threshold C2

3.2.2. Thuật toán SVM

Trong luận văn này, thư viện SVM [33] trên Matlab được sử dụng để huấn luyện ra mô hình dự đoán. Tiến trình huấn luyện được mô tả trong sơ đồ hình 3.26. Trong thuật toán này, dữ liệu sẽ được tiền xử lý để ra được giá trị VeDBA và SCAY cho dữ liệu gia tốc từ thiết bị trên cở bò. Việc xử lý này hoàn toàn giống với việc tiền xử lý dữ liệu của thuật toán cây quyết định. Các dữ liệu từ cảm biến đều được tiền xử lý thông qua bộ lọc trung bình của số M.



Hình 3.26: Sơ đồ thực thi thuật toán SVM

3.2.2.1. Trường hợp phân loại hành vi từ dữ liệu gia tốc trên cổ

Trường hợp này, dữ liệu được lấy sau khi tiền xử lý và thu được giá trị VeDBA và SCAY như hình 3.20 giống như trường hợp cây quyết định phía trên và giá trị k-fold được lấy là 3.

Đồ thị hình 3.27 là mối quan hệ giữa giá trị CV với cặp giá trị (C, γ) tương ứng trong định dạng $\log_2 C$ và $\log_2 \gamma$. Đồ thị cho thấy giá trị CV lớn nhất là 87,94%. Khi đó cặp giá trị tốt nhất của (C, γ) là $(8, 2^{-11})$.

Số vector hỗ trợ được xác định là 853. Số SVs được phân bố theo các lớp trạng thái được thể hiện trong bảng 3.3. Các giá trị b tìm được cho hàm quyết định của mỗi cặp một đối một được thể hiện trong bảng 3.4.

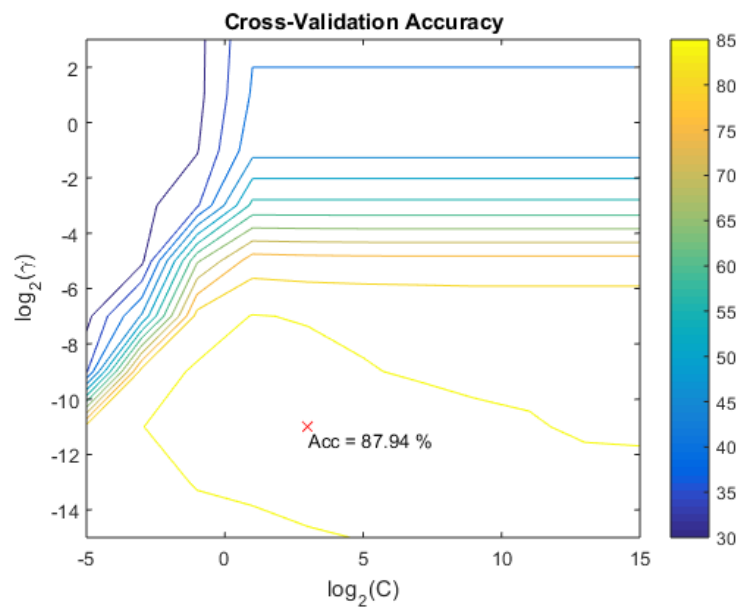
Bảng 3.3: Số SVs phân bố tại các lớp với dữ liệu trên cổ

Lớp trạng thái	Số vector hỗ trợ
Đứng	136
Nằm	73
Ăn	275
Đi	247

Uống	122
Tổng SVs	853

Bảng 3.4: Giá trị b cho hàm quyết định với dữ liệu cổ

Cặp một đối một	Giá trị b
Đứng & Nằm	0,5018
Đứng & Ăn	-0,1112
Đứng & Đi	-0,2834
Đứng & Uống	0,3619
Nằm & Ăn	-0,5726
Nằm & Đi	-0,8527
Nằm & Uống	-0,1606
Ăn & Đi	-0,0382
Ăn & Uống	0,6240
Đi & Uống	0,4822



Hình 3.27: Đồ thị xác định CV của thuật toán SVM với dữ liệu trên cổ

3.2.2.2. Trường hợp phân loại hành vi từ dữ liệu gia tốc trên cả cổ và chân

Trong trường hợp này, giá trị k – fold được lấy là 3. Giá trị CV lớn nhất được xác định là 96,76% như trong đồ thị hình 3.28. Khi đó cặp giá trị tốt nhất của (C, γ) tương ứng là $(32, 2^{-15})$.

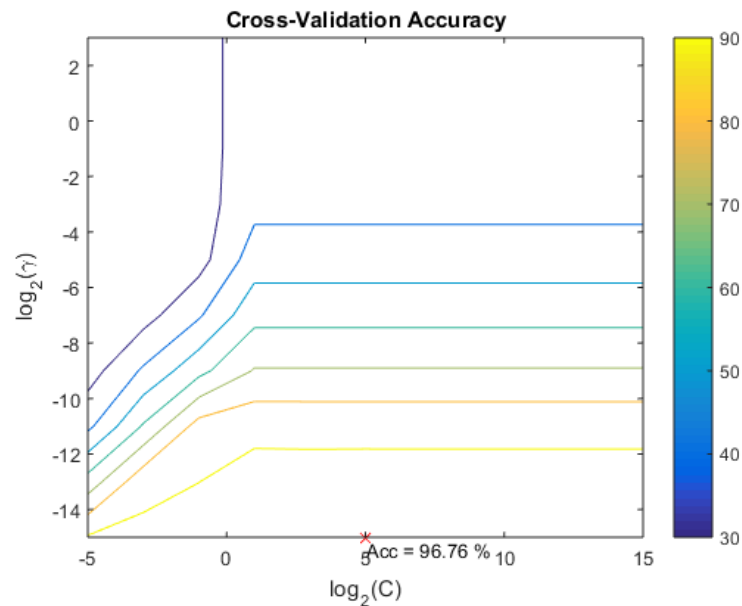
Số vector hỗ trợ được xác định là 590 và được phân bổ cho các lớp trạng thái như bảng 3.5. Các giá trị b tìm được cho hàm quyết định của mỗi cặp một đối một được thể hiện trong bảng 3.6.

Bảng 3.5: Số SVs phân bổ tại các lớp với dữ liệu trên cổ và chân

Lớp trạng thái	Số vector hỗ trợ
Đứng	37
Nằm	8
Ăn	156
Đi	314
Uống	75
Tổng SVs	590

Bảng 3.6: Giá trị b cho hàm quyết định với dữ liệu cổ và chân

Cặp một đối một	Giá trị b
Đứng & Nằm	0,5439
Đứng & Ăn	-0,5770
Đứng & Đi	-0,9590
Đứng & Uống	-0,2476
Nằm & Ăn	-0,8494
Nằm & Đi	-0,9706
Nằm & Uống	-0,6883
Ăn & Đi	-0,6947
Ăn & Uống	0,8316
Đi & Uống	0,8565



Hình 3.28: Đồ thị xác định CV của thuật toán SVM với dữ liệu trên cổ và chân

3.2.3. Đánh giá hiệu năng

Hiệu năng của thuật toán được đánh giá bởi việc đánh giá mô hình phân tách với dữ liệu kiểm tra, dữ liệu này hoàn toàn độc lập với dữ liệu huấn luyện cho việc xây dựng mô hình thuật toán. Luận văn so sánh hiệu năng của 2 thuật toán: cây quyết định và SVM trong các trường hợp khác nhau.

Bảng 3.7: Đánh giá hiệu năng của các thuật toán

		Độ nhạy (%)	Độ chính xác (%)
Thuật toán cây quyết định			
Gia tốc 3 trục trên cổ	Ăn	88,5	83,3
	Nằm	9,7	34,2
	Đứng	84,5	71,9
	Đi	90,5	78
	Uống	70,6	92,2
	Tổng thể	68,8	71,9
Thuật toán SVM			
Gia tốc 3 trục trên cổ	Ăn	84,5	89,6
	Nằm	31,3	43,8
	Đứng	84,5	68,7
	Đi	70,7	81,4
	Uống	84,4	85,2
	Tổng thể	71,1	73,7
Gia tốc 6 trục trên cổ & trên chân	Ăn	79,2	80,7
	Nằm	100	100
	Đứng	94,7	95
	Đi	95,2	85,7
	Uống	72,9	84,3
	Tổng thể	88,4	89,1

Kết quả cho thấy, thuật toán SVM cho kết quả phân loại tốt hơn so với cây quyết định với dữ liệu gia tốc 3 trục trên cổ bò. Sự cải thiện thấy rõ nhất là hành vi nằm của bò, Sensitivity được cải thiện từ 9,7% lên 31,34%. Kết quả phân loại thấp khi chỉ dùng giá trị gia tốc trên cổ bò, đây chính là lý do mà luận văn đề xuất việc sử dụng dữ liệu kết hợp giữa giá trị gia tốc ở chân và cổ.

Kết quả cũng cho thấy, việc kết hợp gia tốc trên chân và trên cổ giúp cho hiệu suất nhận diện và phân loại hành vi trên bò được cải thiện một cách đáng kể. Độ nhạy từ 71,1% với dữ liệu trên cổ lên đến 88,4% khi kết hợp dữ liệu ở cả cổ và chân trong thuật toán SVM. Trong khi đó trạng thái nằm được cải thiện một cách vượt trội từ 9,7% ở cây quyết định lên đến 100% trong thuật toán SVM.

3.3. Kết luận chương 3

Như vậy chương 3 đã trình bày việc xây dựng hệ thống giám sát và phân loại hành vi trên bò. Luận văn đề xuất việc sử dụng 2 thiết bị đo chuyển động tại chân và cổ bò.

Chương 3 cũng đã chỉ rõ việc thực thi các thuật toán học máy và tìm các giá trị tối ưu cho việc phân loại các hành vi đồng thời đánh giá hiệu năng của mỗi thuật toán và cho mỗi trường hợp khác nhau của việc chỉ sử dụng dữ liệu gia tốc tại chỗ và việc kết hợp giá trị gia tốc của cả chân và cổ.

Trong suốt thời gian nghiên cứu thực hiện luận văn, với sự hướng dẫn tận tình của PGS. TS. Trần Đức Tân và TS. Lê Vũ Hà, cùng với những cố gắng và nỗ lực của bản thân, toàn bộ nội dung của luận văn đã hoàn thiện và đáp ứng được các yêu cầu đặt ra.

Luận văn đã đề xuất, thiết kế hệ thống giám sát và phân loại hành vi trên gia súc – bò. Hệ thống được thiết kế dựa trên kỹ thuật mạng cảm biến không dây, mà mỗi vật nuôi sẽ đóng vai trò là 1 nút mạng. Với hệ thống này, người chăn nuôi có thể giám sát vật nuôi tại bất cứ đâu. Điểm mới của luận văn là việc đo chuyển động và giám sát hành vi của bò được thực hiện trên cả chân và cổ so với các nghiên cứu trước đó là chỉ có một thiết bị duy nhất. Luận văn thực hiện phân loại hành vi được thực thi trên hai thuật toán: thuật toán cây quyết định và thuật toán SVM.

Các thuật toán nhận diện và phân loại hành vi được huấn luyện dựa trên dữ liệu thực mà luận văn đã thu thập. Kết quả huấn luyện được đánh giá hiệu năng thông qua dữ liệu kiểm tra. Luận văn đã chỉ ra việc sử dụng thuật toán SVM cho kết quả thực thi tốt hơn cây quyết định và việc kết hợp dữ liệu gia tốc ở cả chân và cổ cho hiệu năng tốt hơn rất nhiều.

Trên cơ sở các kết quả đã thu được, luận văn có thể được phát triển theo các hướng tiếp theo như sau:

- Đánh giá về thời gian sống của mỗi nút mạng và áp dụng kỹ thuật giúp nâng cao thời gian sống của mỗi nút mạng.
- Cải thiện hiệu năng trong phân loại hành vi và mở rộng mô hình dự đoán hành vi cho nhiều đối tượng gia súc khác.
- Chế tạo và triển khai thực tế.

DANH SÁCH CÁC CÔNG BỐ

Nguyễn Đình Chinh, Phùng Công Phi Khanh, Trần Đức Tân, Lê Vũ Hà, “Nghiên cứu và thiết kế hệ thống giám sát hành vi trên bò”, Hội nghị quốc gia về Điện tử, Truyền thông và Công nghệ thông tin 2016 (REV-2016), pp. 6-19 – 6-22, Vol.2016.

TÀI LIỆU THAM KHẢO**Tiếng Anh**

- [1]. Nadimi, Esmail S., et al. "Monitoring and classifying animal behavior using ZigBee-based mobile ad hoc wireless sensor networks and artificial neural networks." *Computers and Electronics in Agriculture* 82 (2012): 44-54.
- [2]. Martiskainen, Paula, et al. "Cow behaviour pattern recognition using a three-dimensional accelerometer and support vector machines." *Applied animal behaviour science* 119.1 (2009): 32-38.
- [3]. Stankovski, S., Ostojic, G., Senk, I., Rakic-Skokovic, M., Trivunovic, S., & Kucevic, D., "Dairy cow monitoring by RFID". *Scientia Agricola*, 69(1), (2012): 75-80.
- [4]. Nadimi, Esmail Shahrak, Henning Tangen Sjøgaard, and Thomas Bak. "ZigBee-based wireless sensor networks for classifying the behaviour of a herd of animals using classification trees." *Biosystems engineering* 100.2 (2008): 167-176.
- [5]. Guo, Y., Corke, P., Poulton, G., Wark, T., Bishop-Hurley, G., & Swain, D. "Animal behaviour understanding using wireless sensor networks". *In Local Computer Networks, Proceedings 2006 31st IEEE Conference on*, (2006, November), IEEE: 607-614..
- [6]. Nielsen, Lars Relund, et al. "Quantifying walking and standing behaviour of dairy cows using a moving average based on output from an accelerometer." *Applied Animal Behaviour Science* 127.1 (2010): 12-19.
- [7]. Schwager, Mac, et al. "Robust classification of animal tracking data." *Computers and Electronics in Agriculture* 56.1 (2007): 46-59.
- [8]. Marchioro, Gilberto Fernandes, et al. "Sows' activity classification device using acceleration data—a resource constrained approach." *Computers and electronics in agriculture* 77.1 (2011): 110-117.
- [9]. Cornou, Cécile, and Søren Lundbye-Christensen. "Classifying sows' activity types from acceleration patterns: an application of the multi-process Kalman filter." *Applied Animal Behaviour Science* 111.3 (2008): 262-273.
- [10]. González, L. A., et al. "Behavioral classification of data from collars containing motion sensors in grazing cattle." *Computers and Electronics in Agriculture* 110 (2015): 91-102.
- [11]. Jegadeesan, S., and GKD Prasanna Venkatesan. "Smart cow health monitoring, farm environmental monitoring and control system using wireless

- sensor networks." *Int. J. Adv. Eng. Tech./Vol VII/Issue I/Jan–March* 334 (2016): 339.
- [12]. Dutta, Ritaban, et al. "Dynamic cattle behavioural classification using supervised ensemble classifiers." *Computers and Electronics in Agriculture* 111 (2015): 18-28.
- [13]. Nathan, Ran, et al. "Using tri-axial acceleration data to identify behavioral modes of free-ranging animals: general concepts and tools illustrated for griffon vultures." *Journal of Experimental Biology* 215.6 (2012): 986-996.
- [14]. Arcidiacono, C., et al. "Development of a threshold-based classifier for real-time recognition of cow feeding and standing behavioural activities from accelerometer data." *Computers and Electronics in Agriculture* 134 (2017): 124-134.
- [15]. Guo, Ying, et al. "Animal behaviour understanding using wireless sensor networks." *Local Computer Networks, Proceedings 2006 31st IEEE Conference on*. IEEE, 2006.
- [16]. Bidder, Owen R., et al. "Love thy neighbour: automatic animal behavioural classification of acceleration data using the k-nearest neighbour algorithm." *PloS one* 9.2 (2014): e88609.
- [17]. Internet: Machine Learning and Optimization,
https://www.cims.nyu.edu/~munoz/files/ml_optimization.pdf
- [18]. Internet: machine learning, <https://www.britannica.com/technology/machine-learning>
- [19]. Ron Kohavi; Foster Provost (1998). "Glossary of terms". *Machine Learning*. 30: 271–274.
- [20]. Internet: Machine learning and pattern recognition "can be viewed as two facets of the same field.", <http://www.iascgroup.it/en/machine-learning-en.html>
- [21]. Internet: exploiting machine learning in cybersecurity,
<https://techcrunch.com/2016/07/01/exploiting-machine-learning-in-cybersecurity/>
- [22]. Wernick, Yang, Brankov, Yourganov and Strother, *Machine Learning in Medical Imaging*, IEEE Signal Processing Magazine, vol. 27, no. 4, July 2010, pp. 25–38
- [23]. Mannila, Heikki (1996). *Data mining: machine learning, statistics, and databases*. Int'l Conf. Scientific and Statistical Database Management. IEEE Computer Society.

- [24]. Friedman, Jerome H. (1998). "Data Mining and Statistics: What's the connection?". *Computing Science and Statistics*. 29 (1): 3–9.
- [25]. Internet: 3 flavors of machine learning who what and where, <http://www.darkreading.com/threat-intelligence/3-flavors-of-machine-learning--who-what-and-where/a/d-id/1324278>
- [26]. Internet: lightcybers jason matlof explains how magna detects hackers before they attack, <http://aibusiness.org/lightcybers-jason-matlof-explains-how-magna-detects-hackers-before-they-attack/>
- [27]. Internet: Machine Learning: What it is & why it matters, https://www.sas.com/it_it/insights/analytics/machine-learning.html
- [28]. Internet: Machine learning, https://en.wikipedia.org/wiki/Machine_learning
- [29]. Piatetsky-Shapiro, Gregory (1991), Discovery, analysis, and presentation of strong rules, in Piatetsky-Shapiro, Gregory; and Frawley, William J.; eds., *Knowledge Discovery in Databases*, AAAI/MIT Press, Cambridge, MA.
- [30]. Internet: Mạch Raspberry pi, <https://raspberrypi.vn/shop/mach-raspberry-pi/>
- [31]. Internet: Vi điều khiển dsPIC33EP64MC502, <http://www.microchip.com/wwwproducts/en/dsPIC33EP64MC502>
- [32]. Internet: Module nRF24L01, https://www.sparkfun.com/datasheets/Components/SMD/nRF24L01Pluss_Preliminary_Product_Specification_v1_0.pdf
- [33]. Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1--27:27, 2011. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [34]. Tom Fawcett. "ROC Graphs: Notes and Practical Considerations for Data Mining Researchers," Technical report, HP labs, 2003.
- [35]. Resheff, Yehezkel S., et al. "AcceleRater: a web application for supervised learning of behavioral modes from acceleration measurements." in *Movement ecology*, Vol. 2, No.1, (2014): 1-7.

Tiếng Việt

- [36]. Đinh Mạnh Tường, "Học máy các kỹ thuật cơ bản và hiện đại", nhà xuất bản Đại học Quốc gia Hà Nội, 2015.