CHẨN ĐOÁN TRẠNG THÁI KĨ THUẬT ĐỘNG CƠ Ô TÔ BẰNG DỮ LIỆU ĐÁP ỨNG VỀ NHIÊN LIỆU CỦA ĐỘNG CƠ VÀ THUẬT TOÁN K-NEAREST NEIGHBOR

AUTOMOTIVE ENGINE DIAGNOSTICS USING FUEL TRIM DATA AND K-NEAREST NEIGHBOR ALGORITHM

TRẦN XUÂN THẾ

Viện Cơ khí, Trường Đại học Hàng hải Việt Nam

Email liên hệ: thetx.vck@vimaru.edu.vn

Tóm tắt

Với sự phát triển của khoa học công nghệ, các hệ thống trên ô tô hiện nay đang được nâng cấp và ngày một trở lên phức tạp. Cùng với đó, việc chẩn đoán các sai lệch và hư hỏng của động cơ cũng như các hệ thống trên ô tô đòi hỏi các phương pháp chẩn đoán mới thay vì dựa vào kinh nghiệm của các kỹ thuật viên. Bài báo đi xây dựng mô hình chẩn đoán trạng thái kĩ thuật của động cơ ô tô bằng dữ liệu Fuel Trims của 300 mẫu dữ liệu xe thu thập được, dựa trên thuật toán K-nearest Neighbor (KNN). Bài báo đã xây dựng được mô hình và tiến hành kiểm nghiệm trên tệp dữ liệu kiểm tra và đạt độ chính xác cao nhất là 87%. Căn cứ vào kết quả mô hình đã thể hiện được mối quan hệ giữa các thông số đầu vào bao gồm độ tuổi, giới tính người điều khiển chính, khu vực xe hoạt động chính, quãng đường xe chạy với chỉ số LTFT, giá trị để đánh giá trạng thái kỹ thuật của động cơ.

***Từ khóa**: Chẩn đoán ô tô, fuel trim, học có giám sát, K-nearest neighbor.*

Abstract

Technology and science have been revolutionized recently. As a result, the systems on the car are increasingly complicated than it were the past. Thus, it requires new methods to diagnostic the engine and automotive systems' technical status, rather than depending on the experience of technicans. In this article, I aim to build a model to diagnostic engine status based on Fuel Trims data collected from 300 car samples and using K-nearest Neighbor (KNN) to train this data. The model was successfully built and got the highest accuracy is 87%. The model illustrated the relationship between input data that include age, gender of drivers, the using location of the cars, the milleague of the cars and LTFT index - the index to evaluate the technical status of car engines.

***Keywords**: Vehicle diagnostic, Fuel trim, Supervised Learning, K-nearest neighbor*.

1. Giới thiệu vấn đề nghiên cứu

Hiện nay, các hãng ô tô đang đầu tư phát triển rất nhiều phần mềm, thiết bị chẩn đoán chuyên hãng với độ chính xác và hiệu quả tương đối cao. Ví dụ như phần mềm Techstream của Toyota, Dịch vụ dữ liệu toàn cầu (GDS) của Huyndai, hay Dịch vụ dữ liệu của Honda (HDS),… cũng như các ứng dụng chẩn đoán cá nhân, có thể cài đặt trên điện thoại. Những phần mềm này, cung cấp cho người sử dụng rất nhiều dữ liệu về các hệ thống trên ô tô, qua đó hỗ trợ kỹ thuật viên rất nhiều trong quá trình chẩn đoán và sửa chữa các hư hỏng của ô tô.

Tuy nhiên, các dữ liệu trên chỉ có thể phục vụ cho những kỹ thuật viên, người có kiến thức chuyên ngành về ô tô do tính phức tạp của chúng. Với việc dữ liệu về trạng thái kỹ thuật của ô tô có thể liên tục được cập nhật hàng giờ bởi hàng triệu người dùng ô tô trên thế giới. Nguồn dữ liệu này sẽ là rất lớn và cần được khai phá để phục vụ rộng hơn cho không chỉ các kỹ thuật viên ô tô mà còn người sử dụng xe trên toàn cầu.

Một số nghiên cứu đã áp dụng Machine Learning để phân tích dữ liệu thu thập được trong quá trình vận hành của ô tô. Xác định đặc điểm điều khiển ô tô của người điều khiển qua đó phân cụm đặc tính người điều khiển và đưa ra các hệ thống tối ưu phù hợp với điều khiển của người lái [1]. Dựa vào các từ ngữ trong miêu tả lỗi của khách hàng, gợi ý ra mã chẩn đoán sự cố (DTC) tương ứng, dựa trên sử dụng các thuật toán về xử lý ngôn ngữ tự nhiên [2]. Xây dựng server để thu thập dữ liệu của các hệ thống ô tô, sử dụng Machine Learning để đưa ra gợi ý bảo dưỡng theo tình trạng thực tế của xe trước khi hư hỏng xảy ra [3]. Chẩn đoán trạng thái động cơ ô tô bằng phân tích âm thanh sử dụng thuật toán trí tuệ nhân tạo (ANN) và phân loại theo phân phối xác suất Bayes (NBC) của Machine Learning [4].

Trong những dữ liệu quan trọng giúp chẩn đoán trạng thái kỹ thuật động cơ ô tô, Fuel Trims có thể nói là dữ liệu quan trọng nhất. Fuel Trim là sự điều chỉnh nhiên liệu bù thêm hoặc giảm bớt đi

của ECU ô tô nhằm giúp cho tỉ lệ không khí, nhiên liệu (tỉ lệ A/F) nạp vào động cơ luôn ở tỉ lệ lý tưởng là 14,7:1 [5]. Nếu tỷ lệ trên giảm đi hoặc tăng lên, đều gây ra tình trạng hao phí nhiên liệu và các nguy cơ gây các hư hỏng tới các hệ thống liên quan trên ô tô. Qua đó, có thể nói đáp ứng nhiên liệu của động cơ (Fuel trim - FT) là các thông số phản ánh tình trạng hoạt động của động cơ một các định lượng và hiệu quả.

Đáp ứng nhiên liệu của động cơ về bản chất là thông số ghi lại sự thay đổi về tỉ lệ không khí, nhiên liệu trong quá trình động cơ hoạt động, nếu tỉ lệ không khí nhiên liệu luôn duy trì ở mức tốt nhất 14,7:1, giá trị của đáp ứng nhiên liệu của động cơ bằng 0. Giá trị này tăng lên khi hỗn hợp không khí nhiên liệu ở tình trạng nghèo, tức là tỉ lệ không khí, nhiên liệu tăng lên, do có nhiều hơn không khí được nạp vào so với bình thường, hoặc ít hơn nhiên liệu được phun so với trạng thái bình thường của động cơ, nguyên nhân có thể so một số hư hỏng trong hệ thống nhiên liệu, ví dụ như bơm nhiên liệu hay vòi phun,… Ngược lại, giá trị đáp ứng nhiên liệu của động cơ giảm đi khi hỗn hợp không khí nhiên liệu ở tình trạng giàu, có ít không khí được nạp hơn trạng thái bình thường của động cơ, hoặc có nhiều liệu được phun hơn so với trạng thái bình thường của động cơ

Có hai loại dữ liệu đáp ứng nhiên liệu của động cơ bao gồm đáp ứng nhiên liệu của động cơ trong ngắn hạn (Short term fuel trim - STFT) ghi lại các hiệu chỉnh nhiên liệu của động cơ trong ngắn hạn cập nhật liên tục theo các trạng thái của động cơ sau một vài giây một lần. STFT có khả năng thay đổi để đáp ứng với các trạng thái mới của động cơ. Dữ liệu đáp ứng nhiên liệu của động cơ thứ hai là đáp ứng nhiên liệu của động cơ trong dài hạn (Long term fuel trim - LTFT) dữ liệu này theo dõi tình trạng động cơ trong dài hạn, trong khi STFT có thể thay đổi theo các trạng thái mới của động cơ bao gồm cả các sai lệch trong quá trình làm việc của động cơ, LTFT sẽ theo dõi các thay đổi của STFT và đưa ra đánh giá về tình trạng thực tế của động cơ ở thời điểm hiện tại. Do đó, LTFT thường có ý nghĩa hơn trong chẩn đoán động cơ ô tô. Về giá trị, nếu các giá trị đáp ứng nhiên liệu của động cơ của ô tô trong khoảng từ -8% đến 8% là bình thường, từ 8% - 20% là vùng nguy cơ cao động cơ sẽ xuất hiện sai lệch trong hoạt động, từ 20% - 25% là vùng sai lệch, ECU sẽ thiết lập các mã lỗi để cảnh báo sai lệch này trong hoạt động của động cơ.

Hiện nay, giá trị của Fuel Trims được sử dụng rộng rãi trong công tác chẩn đoán của các kỹ thuật viên ô tô, giá trị này thu được bằng việc sử dụng các phần mềm, thiết bị chẩn đoán như đã đề cập ở trên. Do đó, việc xác định nhanh chóng trạng thái kỹ thuật động cơ đối với người sử dụng xe đôi khi rất khó khăn và tốn kém thời gian hoặc người sử dụng ô tô cần đầu tư mua các thiết bị chẩn đoán cần tay có thể kết nối với các thiết bị di động, điều này là tương đối lãng phí và không quá cần thiết đối với người sử dụng cá nhân.

Do đó, nghiên cứu này nhằm xây dựng một phương pháp chẩn đoán mới, sử dụng các thuật toán của Machine Learning để xây dựng các mô hình dự đoán tình trạng kỹ thuật của động cơ ô tô dựa trên các dữ liệu dễ dàng xác định từ người sử dụng ô tô bao gồm thông tin về tuổi tác, giới tính, vị trí địa lý ô tô thường được sử dụng, quãng đường ô tô đã đi. Với sự hỗ trợ của các thiết bị, và phần mềm chẩn đoán ô tô hiện nay, giá trị về Long Term Fuel Trim của các xe được điều tra dữ liệu sẽ được sử dụng để làm nhãn phân loại cho tình trạng kỹ thuật hiện tại của động cơ. Do dữ liệu dùng để huấn luyện mô hình đã lựa chọn được LTFT làm nhãn do đó bài toán xây dựng mô hình sẽ là bài toán học có giám sát (supervised learning) của Machine Learning.

Các phần tiếp theo của bài báo sẽ trình bày các nội dung sau, phần 2 của bài báo sẽ đi trình bày về phương pháp nghiên cứu bao gồm cách thu thập, xử lý và chuẩn hóa dữ liệu, các thuật toán Machine Learning được sử dụng để phân tích trong dữ liệu, phần 3 bài báo sẽ trình bày kết quả kiểm tra các mô hình được xây dựng, phần 4 sẽ đưa ra một số bàn luận xung quanh vấn đề nghiên cứu.

2. Phương pháp xây dựng và phân tích bộ dữ liệu

Nghiên cứu được thiết kế thông qua thu thập thông tin của 300 chủ xe Kia Morning, loại xe phổ biến nhất tại Việt Nam năm sản xuất từ 2014 cho đến 2016. Dữ liệu được thu thập tại cả khu vực thành thị và ngoại thành Hải Phòng, một trong 4 thành phố lớn nhất của Việt Nam, trong điều kiện thời tiết không mưa, trong hai tháng mùa khô tại Việt Nam là tháng 12 và tháng 1. Nhiệt độ trung bình trong hai tháng này là 20^{o}C, nhiệt độ cao nhất 25^{o}C, thấp nhất 16^{o}C.

Đối tượng điều tra được lựa chọn ngẫu nhiên qua chương trình chẩn đoán ô tô miễn phí mà nghiên cứu cung cấp để thu thập dữ liệu nhanh hơn. Thông qua cân đối giữa thời gian huấn luyện dữ liệu và đặc tính của thông số chúng tôi quyết định thu thập 300 mẫu dữ liệu cho nghiên cứu.

Thiết bị được sử dụng là thiết bị chẩn đoán ô tô Gscan, model 2.0, phiên bản phần mềm quốc tế 2018, nơi sản xuất tại Hàn Quốc.

Việc đo lường được tiến hành ở chế độ không tải (sau khởi động) sau khi động cơ đã được làm nóng năm phút. Kết quả lần đo đầu tiên được loại bỏ, kết quả cuối cùng là trung bình cộng giá trị LTFT của xe ở hai lần đo tiếp theo, mỗi lần đo cách nhau ba phút.

Giá trị LTFT là giá trị được hiển thị trên máy chẩn đoán Gscan 2 trong phần kiểm tra thông số của động cơ. Cơ sở xác định giá trị LTFT dựa trên dữ liệu thực tế thu thập được từ cảm biến tỉ lệ không khí nhiên liệu A/F, so với giá trị LTFT được lập trình sẵn trong bản đồ sử dụng nhiên liệu của ô tô tại các vòng quay và tải của động cơ qua đó tính toán ra giá trị LTFT theo đơn vị phần trăm.

Một bộ câu hỏi cũng được thiết kế để thu thập dữ liệu liên quan đến các yếu tố sử dụng xe bao gồm tuổi của người điều khiển xe chính, giới tính của người điều khiển xe chính, khu vực sử dụng xe phổ biến, quãng đường sử dụng xe. Đây là những dữ liệu đơn giản, dễ xác định cho đối tượng sử dụng của mô hình chẩn đoán là người sử dụng xe ô tô, không phải là kỹ thuật viên hay người có kiến thức chuyên ngành về ô tô.

Bảng 1. Dữ liệu của các quan sát trong tệp dữ liệu khảo sát trước chuẩn hóa

Mã số	**Tuổi người sử dụng xe chính** (tuổi)	**Giới tính người sử dụng xe chính**	**Khu vực sử dụng xe**	**Quãng đường xe chạy** (Km)	**Giá trị LTFT khảo sát được** (%)
001	26	Nam	Thành phố	27.701	21
002	40	Nam	Ngoại thành	20.115	1
003	30	Nữ	Thành phố	21.613	1
004	33	Nữ	Ngoại thành	21.288	10

Dữ liệu sau khi được thu thập sẽ được chuẩn hóa về dạng số, đối với thông tin về giới tính, giới tính Nam sẽ tương ứng với 1, nữ tương ứng với 0; đối với thông tin về khu vực sử dụng xe, khu vực thành phố được quy ước là 1, ngoại thành là 0 và xe sử dụng ở cả hai khu vực trên được quy ước ghi 2.

Đối với dữ liệu LTFT được xác định là nhãn của dữ liệu, để phân loại tình trạng động cơ. Đối với các thuật toán Classification (Phân loại) như KNN, giá trị LTFT được phân ra thành 3 classes (nhóm). Nhóm 0 tương ứng với giá trị tuyệt đối của LTFT < 8%, đại diện cho động cơ xe làm việc bình thường. Nhóm 1 tương ứng với giá trị tuyệt đối của LTFT trong khoảng từ 9% đến 19%, đại diện cho động cơ xe đang có sai số và khả năng cao sẽ gặp hư hỏng. Nhóm 2 tương ứng với giá trị tuyệt đối của LTFT trên 20%, đại diện cho tình trạng động cơ đang gặp sự cố, làm việc không tốt hoặc không làm việc [5]. Sau khi được chuẩn hóa, dữ liệu trong Bảng 1 sẽ tương ứng với dữ liệu trong Bảng 2 sau.

Bảng 2. Dữ liệu của các quan sát trong tệp dữ liệu khảo sát sau chuẩn hóa

Mã số	**Tuổi người sử dụng xe chính** (tuổi)	**Giới tính người sử dụng xe chính**	**Khu vực sử dụng xe**	**Quãng đường xe chạy** (Km)	**Giá trị LTFT khảo sát được** (%)
001	26	1	1	27.701	2
002	40	1	0	20.115	0
003	30	0	1	21.613	0
004	33	0	0	21.288	1

Tập dữ liệu bao gồm 300 mẫu thu được, sẽ được chia thành 2 tập, tập thứ nhất là tập huấn luyện (traning set) dùng để xây dựng mô hình bao gồm 70% dữ liệu (210 mẫu), tập thứ 2 là tập kiểm tra dùng để kiểm tra tính chính xác của mô hình xây dựng bao gồm 30% dữ liệu (90 mẫu).

Với việc dữ liệu được gán nhãn bởi thông số LTFT, các thuật toán học có giám sát (Supervised learning) của Machine Learning sẽ được sử dụng. Sau khi cân nhắc đặc điểm dữ liệu, cũng như đặc điểm của các thuật toán học có giám sát, tác giả sử dụng thuật toán K-nearest Neighbor (KNN) để huấn luyện dữ liệu trong tệp huấn luyện. Thông số đầu vào của thuật toán là các thông trong bộ dữ liệu huấn luyện bao gồm độ tuổi, giới tính người điều khiển chính, khu vực xe hoạt động chính, quảng đường xe chạy là những thông số khảo sát, cùng với đó đầu vào của thuật toán còn bao gồm cả dữ liệu về LTFT dùng làm nhãn để thể hiện trạng thái của động cơ. Các đầu vào này sẽ tạo ra đầu ra là một không gian dữ liệu mà ở đó, tất cả các điểm dữ liệu (xác định bởi 5 thông số trên) đều đã thể hiện tình trạng động cơ bình thường, có nguy cơ hư hỏng hay đang gặp hư hỏng. Dựa vào không gian dữ liệu huấn luyện này, khi đưa các dữ liệu cần kiểm tra (chưa có nhãn, chưa thể hiện

thông tin về tình trạng động cơ) vào, mô hình sẽ căn cứ vào k điểm gần nhất với dữ liệu cần kiểm tra để xác định nhãn cho dữ liệu này (số k sẽ được xác định sau khi chạy mô hình với nhiều giá trị k khác nhau để xác định được giá trị k phù hợp nhất cho bộ dữ liệu và được thể hiện trong Hình 1 dưới đây).

Thuật toán sẽ cho ra kết quả dự đoán về nhóm mà dữ liệu đó thuộc vào tương ứng với LTFT nhóm 0, 1 hay 2.

Một số thuật toán học có giám sát khác cũng có thể được sử dụng như trí tuệ nhân tạo (ANN) và phân loại theo phân phối xác suất Bayes (NBC). Tuy nhiên, so với hai thuật toán trên thuật toán KNN có lợi thế về mặt tốc độ huấn luyện dữ liệu nhanh, không tốn tài nguyên khi huấn luyện dữ liệu và phù hợp cho bài toán có số lượng dữ liệu nhỏ. Đối với thuật toán ANN, thuật toán này thường cho độ chính xác cao với các dữ liệu lớn, cấu trúc mô hình ANN là tương đối phức tạp, do đó mất nhiều thời gian và tài nguyên để huấn luyện mô hình. Thuật toán NBC là thuật toán thiên về sử dụng xác suất có tốc độ huấn luyện nhanh nên rất phù hợp với bài toán có dữ liệu lớn, NBC đặc biệt thích hợp với các bài toán về xử lý ngôn ngữ tự nhiên [6].

Thuật toán được huấn luyện và kiểm tra bằng phần mềm Anaconda, dựa trên ngôn ngữ lập trình Python.

Bản chất của thuật toán KNN là tương đối đơn giản, thuật toán xác định k điểm gần nhất với điểm dữ liệu đang xét trong không gian dữ liệu như đã đề cập tới ở trên, dựa theo khoảng cách E-cơ-lít trong không gian hai chiều, hay giá trị về khoảng cách trong không gian véc tơ (được kí hiệu là Norm 2). Công thức cụ thể như sau:

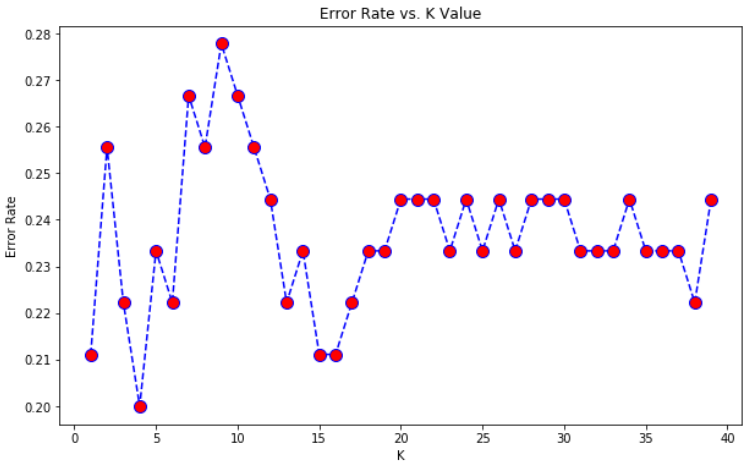
$$d(q,p) = d(p,q) = \sqrt[2]{({q_1}^2 - {p_1}^2) + ({q_2}^2 - {p_2}^2) + \cdots + ({q_n}^2 - {p_n}^2)} = \sqrt[2]{\sum_{i=1}^{n}({q_i}^2 - {p_i}^2)} \quad (1)$$

Trong đó:

• d(q,p) và d(p,q) là khoảng cách giữa điểm dữ liệu đang xét với các điểm dữ liệu gần nó nhất trong không gian huấn luyện.

• q_i, p_i là giá trị tọa độ điểm thứ i của mô hình.

3. Kết quả xây dựng và kiểm tra mô hình xây dựng bằng thuật toán K-nearest Neighbor (KNN)



Hình 1. Mối liên hệ giữa sai số dự đoán với các giá trị của k khi xây dựng mô hình bằng thuật toán KNN

Bảng 3. Kết quả kiểm nghiệm thuật toán KNN cho 90 dữ liệu trong tệp kiểm tra tại k = 4

Tổng số: 90	Dự đoán là 0	Dự đoán là 1	Dự đoán là 2
Nhãn 0	55 (87%)	8 (13%)	0 (0%)
Nhãn 1	5 (31%)	16 (69%)	0 (0%)
Nhãn 2	2 (33%)	4 (66%)	0 (0%)

Trên tập kiểm tra bao gồm 90 mẫu dữ liệu, kết quả dự đoán cho nhóm có nhãn LTFT bằng 0 xe bình thường đạt độ chính xác 87%, 13% kết quả nhóm này bị dự đoán nhầm sang nhãn 1. Độ chính xác dự đoán cho nhóm có nhãn 1 thấp hơn với chỉ 69%, 31% kết quả của nhóm này bị dự đoán nhầm sang nhóm 0. Kết quả dữ đoán cho nhóm nhãn 2 chưa chính xác, do số lượng dữ liệu được dãn nhãn này trong bộ dữ liệu quá nhỏ. Đối với việc thực hiện các vòng lặp của thuật toán,

giá trị k là số điểm dữ liệu gần nhất lấy làm căn cứ để xác định nhóm cho dự liệu cần kiểm tra. Ta nhận thấy, bộ dữ liệu với 300 mẫu dữ liệu là tương đối nhỏ. Do đó thuật toán sẽ cho kết quả chính xác hơn nếu giá trị k nhỏ. Nếu số k lớn sẽ dẫn tới việc đan xen các điểm dữ liệu làm căn cứ phân lớp cho dữ liệu kiểm tra dẫn tới độ chính xác giảm xuống.

4. Kết luận

Nguồn dữ liệu về trạng thái kỹ thuật của ô tô cũng như động cơ ô tô đang được cập nhật hàng ngày. Việc khai thác nguồn dữ liệu này để đưa ra các dự đoán về trạng thái kỹ thuật của ô tô là rất cấp thiết. Bài báo đã sử dụng thuật toán K-nearest Neighbor (KNN) là một thuật toán phân loại đơn giản và hiệu quả cho bài toán phân loại nhiều nhóm dữ liệu. Mô hình của bài báo có thể được áp dụng trong thực tiễn một cách nhanh chóng thông qua tích hợp vào ứng dụng trên điện thoại cá nhân, có thể giúp người điều khiển xe có những gợi ý về tình trạng kĩ thuật của động cơ trên ô tô của họ, qua đó đảm bảo tính an toàn trong quá trình điều khiển và sử dụng ô tô, cũng như giúp người điều khiển xe sớm có kế hoạch bảo dưỡng cho xe của mình.

TÀI LIỆU THAM KHẢO

[1] C. Barreto*, A Machine Learning Approach Based on Automotive Engine Data Clustering for Driver Usage Profiling Classification*, Halmstad University Press, 2014.

[2] M. Yi Lu*, Vehicle Fault Diagnostics Using Text Mining, Vehicle Engineering Structure and Machine Learning,* International Journal of Intelligent Information Systems, 2015.

[3] U. Shafi*, Vehicle Remote Health Monitoring and Prognostic Maintenance System*, Journal of Advanced Transportation, 2018.

[4] C. Barreto*, A Machine Learning Approach Based on Automotive Engine Data Clustering for Driver Usage Profiling Classification*, Australian Information Security Management Conference, 2013.

[5] Internet Resource: http://greencar.vn/dong-co-xang/fuel-trim-la-gi-hieu-ve-su-dieu-chinh-nhien-lieu-cua-dong-co-o-to/.

[6] H. Trevo, *The Elements Of Statictical Learning Second Edition*, Springer, 2018.

Ngày nhận bài: 09/5/2019
Ngày nhận bản sửa: 16/5/2019
Ngày duyệt đăng: 22/5/2019