Cây quyết định trong trích chọn đặc tính gen cho phân loại ung thư sử dụng dữ liệu biểu hiện gen DNA Microarray

Decision Tree Based Feature Selection for Tumor Classification using DNA Microarray Gene Expression Data

Phạm Trung Dũng
Khoa Kỹ thuật Điều khiển
Học viện Kỹ thuật Quân sự
e-mail :
thchdung@yahoo.com

Đặng Thúy Hằng
Khoa Kỹ thuật Điều khiển
Học viện Kỹ thuật Quân sự
e-mail :
hangdtys@gmail.com

Trần Hoài Linh
Viện Điện
Trường ĐHBK Hà Nội
e-mail:
linh.tranhoai@hust.edu.vn

Tóm tắt

Microarray hiện là một trong những công cụ hữu hiệu trong chẩn đoán bệnh do dữ liệu từ các thí nghiệm này chứa các thành phần gen đặc trưng cho một bệnh cụ thể. Tuy nhiên, đặc điểm của loại dữ liệu này là chứa hàng nghìn gen khác nhau trong khi số lượng mẫu lại rất ít nên cần thiết phải có phương pháp lựa chọn các gen phù hợp cho quá trình phân tích và chẩn đoán. Có rất nhiều phương pháp lựa chọn gen đã được nghiên cứu và phát triển. Bài báo này sẽ giới thiệu phương pháp cây quyết định để lựa chọn các gen chứa thông tin đặc trưng. Các giá trị đặc trưng này được tiếp tục xử lý bằng một mô hình phi tuyến để đưa ra được kết quả về phân loại dữ liệu biểu hiện gen để có kết luận về phân loại bệnh ung thư.

Từ khóa: Microarray, ung thư, cây quyết định, mạng MLP

Abstract: Microarray experiments are becoming a powerful tool for clinical diagnosis, as they have the potential to discover gene expression pattern obtaining characteristics of a certain disease. However, the microarray data have thousands of genes within a few samples, it is crucial to develop techniques to effectively gene selection for analysis. In this paper, Decision Tree Algorithm has been proposed to detect information gene for Multi Layer Perceptron for efficient tumor classification.

Keywords: Microarray, cancer, Decision tree, Multi Layer Perceptron

Chữ viết tắt

BL	Burkitt Lymphoma
cDNA	complement Deoxyribo Nucleic Acid
DNA	Deoxyribo Nucleic Acid
EST	Expressed Sequence Tag
EWS	Ewing family of tumors
MPSS	Massively Parallel Signature Sequencing
MLP	Multi Layer Perceptron
NB	Neuroblastoma
RMS	Rhabdomyosarcoma
RT-PCR	Real-time principal component analysis
SRBCT	Small Round Blue Cell Tumor
SAGE	Serial Analysis of Gene Expression

1. Giới thiệu

Ung thư hiện nay là một trong những căn bệnh nguy hiểm với tỷ lệ tử vong rất cao. Kết quả điều trị phụ thuộc rất nhiều vào việc chẩn đoán bệnh. Hiện nay, microarray là một công cụ hữu hiệu giúp cho việc chẩn đoán và điều trị các căn bệnh ung thư hiệu quả hơn. Tuy nhiên, nhược điểm của phương pháp này là lượng dữ liệu đầu vào khá lớn và nhiều chiều khiến cho việc xử lý và phân tích phức tạp hơn. Để giải quyết vấn đề này, nhiều phương pháp giảm chiều dữ liệu và lựa chọn gen đặc trưng đã được đề xuất sử dụng. Cùng chung mục đích đó, bài báo này sẽ đề cập đến thuật toán cây quyết định (*Decision Tree* - DT) để tìm các mẫu gen đặc trưng cho từng nhóm bệnh ung thư. Các mẫu gen được lựa chọn này sau đó sẽ được đưa vào mạng nơ-rôn (cụ thể trong bài báo này là mạng nơ- rôn nhiều lớp MLP) để phân loại và kiểm chứng cho chất lượng của giải pháp được đề xuất. các kết quả tính toán và kiểm nghiệm chứng minh giải pháp đề xuất của bài báo là có triển vọng trong việc phân loại các mẫu bệnh ung thư từ dữ liệu thu nhận được trong các thí nghiệm Micrroarray.

2. Cơ sở sinh học microarray

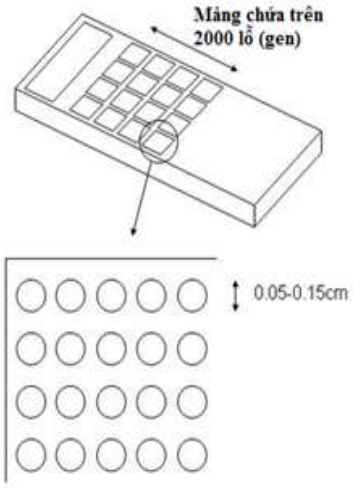
Cơ thể người có rất nhiều loại gen khác nhau. Các gen khác nhau này có thể "bật" hoặc "tắt" khi tồn tại trong các loại tế bào khác nhau. Một cách để phát hiện xem các gen này có hoạt động hay không chính là việc tìm ra các thành phần mang thông tin mRNA, hay ta còn gọi là các gen này có "biểu hiện". Nếu chúng ta có thể đo được số lượng mRNA của tất cả các gen có mặt trong một tế bào hay một mẫu mô sinh học thì ta có thể tạo ra được một bảng các loại gen được "bật" hay các gen được biểu hiện trong những tế bào đó. Nếu so sánh bộ dữ liệu các gen được biểu hiện ra giữa hai loại tế bào khác nhau thì ta sẽ biết được nguyên nhân tạo ra sự khác biệt giữa các tế bào đó [1].

Bảng 1 minh họa một số gen được biểu hiện trong tế bào cơ và tế bào Insulin. Qua bảng 1 có thể nhận thấy gen Myosin được biểu hiện trong tế bào cơ nhưng không được biểu hiện trong tế bào Insulin. Ngược lại, gen Insulin lại biểu hiện trong tế bào Insulin nhưng lại không được biểu hiện trong tế bào cơ. Dựa vào những điểm đặc trưng đó, ta mới có thể phân biệt được hai loại tế bào với nhau.

Bảng 1: *So sánh sự biểu hiện của gen trong tế bào cơ và tế bào insulin [2]*

Tên Gen	*Tế bào cơ*	*Tế bào Insulin*
ACTIN	**ON**	**ON**
MYOSIN	**ON**	OFF
INSULIN	OFF	**ON**
MELANIN	OFF	OFF
CYCLIN 1	**ON**	**ON**
VIBIQUITIN	**ON**	**ON**
HISTONE 2B	**ON**	**ON**

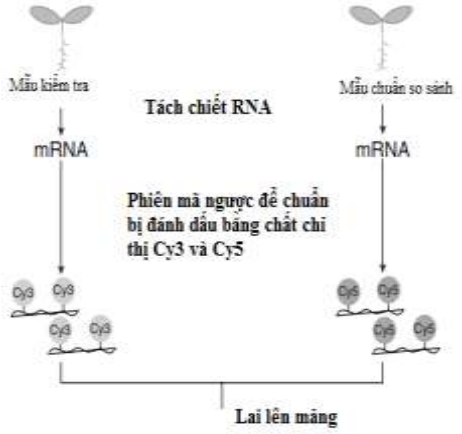
Ngày nay, có rất nhiều phương pháp ứng dụng chương trình máy tính được sử dụng để đo mức biểu hiện gen. Một số phương pháp phổ biếnđược sử dụng có thể kể đến là Northern blots, RT-PCR, Macroarray, Microarrays, phân tích chuỗi gen SAGE, so sánh EST và MPSS [3]. Với đặc tính vượt trội là có thể gắn được hàng nghìn phân tử DNA khác nhau lên một mảng nên có thể đo được biểu hiện của hàng nghìn gen một cách đồng thời trong khi chi phí phù hợp nên Microarray tỏ ra là một công cụ phân tích khá hiệu quả và không thể thiếu trong sinh học hiện nay.



H. 1 *Các lỗ trong một chip microarray*

Microarray gồm một bề mặt rắn, thường là một miếng kính hiển vi, trên đó là các phân tử DNA hoặc các Oligonucleotide được gắn cố định. Mục đích của một microarray là phát hiện sự biểu hiện và số lượng của các DNA được đánh dấu trong một mẫu sinh học. Các DNA trong mẫu sinh học cần kiểm tra được lai với các DNA trên vi mảng microarray thông qua sự ghép cặp theo nguyên lý Watson-Crick và được nhận biết thông qua việc đánh dấu. Điểm mạnh của microarray là có thể gắn được hàng nghìn phân tử DNA khác nhau lên một mảng và do đó ta có thể đo được biểu hiện của hàng nghìn gen một cách đồng thời. Điều này cho phép chúng ta có thể phân tích các thông tin về gen rất nhanh và chính xác [4]. Từ đó tiến tới các nghiên cứu để xác định các gen có biểu hiện khác nhau, phân loại tế bào, xác định các loại bệnh và đưa ra các tương tác điều hòa gen [5].

Quá trình thí nghiệm được minh họa trên H.2. Với một mẫu RNA cần kiểm tra, một chuỗi các phản ứng hóa sinh được thực hiện để tạo ra các đầu dò cRNA hoặc cDNA bổ sung được đánh dấu huỳnh quang. Đầu dò này được lai với microarray và quét bằng chùm tia Laser. Các mức biểu hiện được đo thông qua việc đánh giá cường độ huỳnh quang phát ra từ các lỗ của microarray.



H. 2. *Thí nghiệm Microarray*

Dữ liệu hình ảnh thu được này được gọi là dữ liệu thô. Để đạt được thông tin về mức độ mô tả gene thì dữ liệu hình ảnh này cần phải được phân tích bao gồm: xác định mỗi spot trên mảng sau đó đo và so sánh cường độ của mỗi spot với giá mang. Đây là quá trình lượng hóa hình ảnh. Sau quá trình lượng hóa hình ảnh ta thu được dữ liệu mô tả gen.

3. Thuật toán cây quyết định trong lựa chọn đặc tính

Thuật toán cây quyết định (Decision Tree) là một mô hình phân lớp trong nhận dạng và phân loại dữ liệu [6,7]. Một mô hình cây đơn giản nhất là cây nhị phân, đây là cây chỉ sử dụng điều kiện đơn đơn giản như "if x_i *op* A" tại các nút. Trong đó, *op* là các phép toán so sánh như =, >, <, >=, <=.

H. 3 *Cấu trúc cây quyết định*

Cấu trúc của một cây quyết định được cho trong Hình H.3. Có nhiều thuật toán để huấn luyện một cây, với thuật toán ID3 [6,7] các hàm khuếch đại entropy nút được sử dụng để tối ưu cấu trúc của cây và các điều kiện rẽ nhánh cho từng nút. Do đó, nếu tại nút V ta có N mẫu $x_1, x_2,..., x_N$ thuộc M lớp $C_1, C_2,..., C_M$ thì entropy của lớp đó là

$$E(V) = \sum_{i=1}^{M} - p_i \log_2 \ p_i \qquad (1)$$

trong đó $p_i = \dfrac{\overline{x_j : x_j \in C_i}}{N}$ là xác suất mà một mẫu x_j của nút thuộc lớp C_i. Với một điều kiện S, các mẫu từ nút V được phân thành các nút nhỏ hơn SV_i (với cây nhị phân i=1 hoặc 2) với số lượng các mẫu phù hợp là N_i $\sum_i N_i = N$. Lúc này, hàm entropy cho nút V với điều kiện S cho bởi công thức

$$Gain(V,S) = E(V) - \sum_i \frac{N_i}{N} E(SV_i) \qquad (2)$$

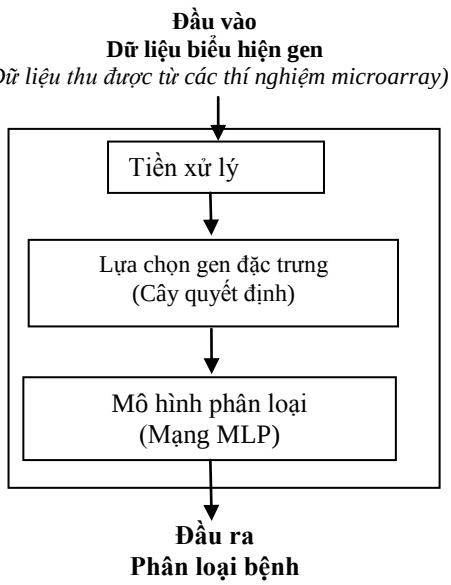
Một điều kiện phân nhánh tốt là điều kiện có giá trị điều chỉnh entropy đạt lớn nhất đối với từng nút.

4. Các kết quả tính toán và mô phỏng

4.1. Cơ sở dữ liệu

Bộ sơ sở dữ liệu sử dụng trong luận án là bộ dữ liệu ung thư tế bào xanh thể cầu [8] lấy từ [9]. Bộ dữ liệu bao gồm 83 mẫu bệnh trong đó có 29 bệnh nhân mắc bệnh ung thư mô xương , 25 bệnh nhân mắc ung thư mô liên kết, 11 bệnh nhân bị u lympho Burkitt và 18 bệnh nhân mắc bệnh u nguyên bào thần kinh. Bộ dữ liệu tổng chứa dữ liệu biểu hiện của 2308 gen.

4.2. Mô hình phân loại

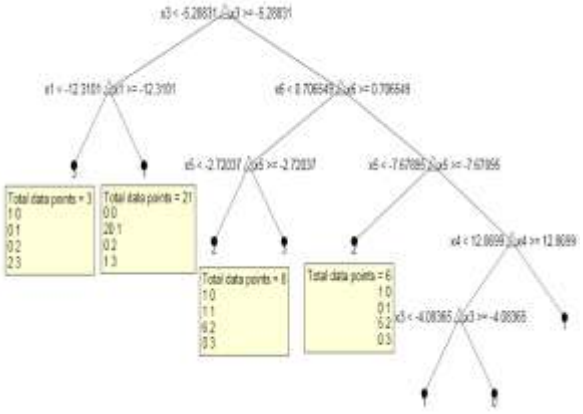


***H.4**. Sơ đồ khối phân loại dữ liệu biểu hiện gen*

Phương pháp đề xuất được thực hiện tuần tự qua ba bước. Trước tiên, ta sẽ thu nhận dữ liệu đo được từ các mảng microarray. Những dữ liệu này được lấy từ các nguồn khác nhau nên cần phải chuẩn hóa và đưa về chung một định dạng dữ liệu của Matlab để thuận tiện cho quá trình tính toán và kiểm nghiệm sau này. Bước thứ hai, sử dụng cây quyết định để lựa chọn các giá trị đặc trưng (hay còn gọi là các đặc tính). Trong bước cuối cùng, bước thứ ba, các giá trị đặc trưng này được xử lý tiếp tục bằng một mô hình phi tuyến để đưa ra được kết quả về phân loại dữ liệu biểu hiện gen. Sơ đồ khối của ý tưởng này được trình bày trên hình H.4.

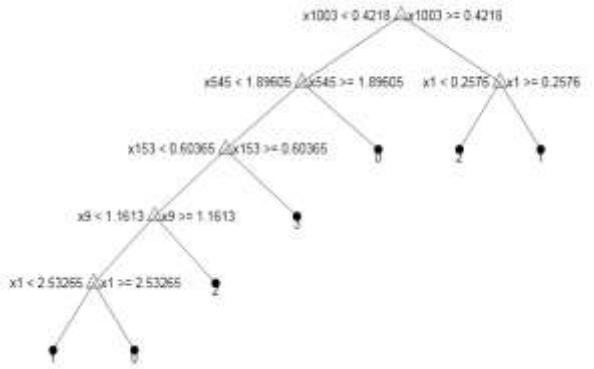
4.3. Kết quả

Bộ cơ sở dữ liệu ung thư tế bào xanh thể cầu có 4 nhóm bệnh khác nhau. Khi sử dụng thuật toán ID3 cho một số thành phần biểu hiện gen đầu tiên thì các mẫu bệnh vẫn còn lẫn vào nhau như minh họa trên hình H.5.



***H. 5** Phân tách nhóm bệnh ung thư tế bào xanh thể cầu nhỏ sử dụng thuật toán ID3 với một số biểu hiện gen đầu tiên*

Ta có thể nhận thấy, với bộ dữ liệu ung thư tế bào xanh thể cầu nếu chỉ sử dụng các biểu diễn gen đầu tiên, các phân bố giá trị của bốn nhóm bệnh RMS, EWS, BL và NB vẫn trùng lên nhau nên khó có thể khoanh vùng để phân tách được các loại bệnh này với nhau. Từ đó dẫn tới việc phân loại các trường hợp bệnh này không hiệu quả. Do đó ta tiến hành tìm kiếm các biểu hiện gen có khả năng phân loại tốt nhất các mẫu bệnh trong tập cơ sở dữ liệu.



***H. 6** Cây quyết định không chứa lỗi phân loại nhóm bệnh ung thư tế bào xanh thể cầu*

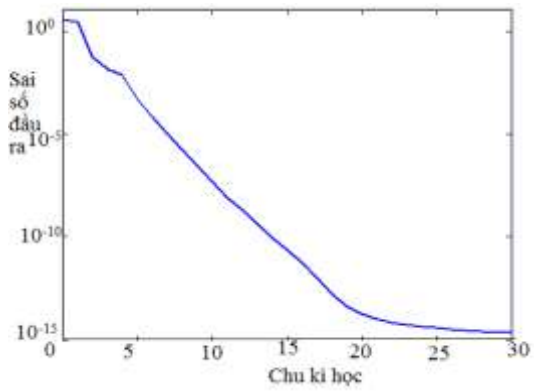
Sử dụng lại thuật toán cây quyết định với một số biểu hiện gen khác trong toàn bộ tập dữ liệu thì kết quả phân tách đã tốt hơn rất nhiều như minh họa trên hình H.6 .

Thuật toán ID3 ứng với cây quyết định trên hình H.6 tương ứng với 7 luật phân loại sau:

```
1. If (x1003 ≤ 0.4218) and (x1< 0.2576) then class=2(BL)
2. If (x1003 ≤ 0.4218) and (x1≥ 0.2576)
3. If(x1003< 0.4218) and (x545≥ 1.89605) then class=0(EWS) and (x4< -21187) and (x1<11877.9) then class=1(RMS)
4.If(x1003<0.4218)and(x545<1.89605)and(x153≥ 0.60365) then class=3(NB)
5.If(x1003<0.4218)and(x545<1.89605)and(PCA153< 0.60365)and (x9 ≥ 1.1613) then class=2(BL)
6.If(x1003<0.4218)and(x545<1.89605)and(x153< 0.60365)and (x9 < 1.1613) and (x1≥2.53265) then class=0(EWS)
7.If(x1003<0.4218)and(x545<1.89605)and(x153< 0.60365)and (x9 < 1.1613) and (x1<2.53265) then class=1(RMS)
```

Như vậy, chỉ cần sử dụng 5 biểu hiện gen (x1003, x545, x153, x9 và x1) có thể phân tách tốt các nhóm bệnh trong bộ dữ liệu này.

Để kiểm chứng cho chất lượng của giải pháp được đề xuất, các biểu hiện gen được lựa chọn (x1003, x545, x153, x9 và x1) được sử dụng làm đầu vào huấn luyện cho mạng MLP. Với bộ dữ liệu học và kiểm tra được chia theo tỷ lệ 54 mẫu học (17 mẫu EWS, 16 mẫu RMS, 8 mẫu BL, 13 mẫu NB) và 29 mẫu kiểm tra (12 mẫu EWS, 9 mẫu RMS, 3 mẫu BL, 5 mẫu NB). Sau khi được huấn luyện với thuật toán học Levenberg - Marquadrt [10] ta có thể thấy kết quả của mạng MLP như trên hình H.7 cho bộ dữ liệu ung thư tế báo xanh thể cầu nhỏ chỉ sau 30 bước học đã có thể quan sát thấy quá trình học đã hội tụ.



***H. 7** Quá trình giảm sai số trong 30 chu kỳ học đầu tiên của bộ dữ liệu ung thư tế bào xanh thể cầu nhỏ*

Kết quả phân loại đạt độ chính xác 100%. So với kết quả đạt được trong [11] các tác giả sử dụng phân tích thành phần chính kết hợp với biến đổi Wavelet và sau đó cũng cho qua mạng MLP để phân loại thì đối với bộ số liệu ung thư tế bào xanh thể cầu nhỏ thì chỉ đạt độ chính xác 90,36%, thấp hơn so với phương pháp đề xuất trong bài báo. Với công trình [12], các tác giả cũng đạt được độ chính xác 100% nhưng tỷ lệ mẫu học: mẫu kiểm tra lại lơn hơn so với phương án của bài báo đề xuất.

5. Kết luận

Bài báo đã giới thiệu khái quát về công nghệ microarray và ứng dụng cây quyết định lựa chọn thành phần gen đặc trưng cho bộ cơ sở dữ liệu microarray về bệnh ung thư tế bào xanh thể cầu. Cây quyết định và thuật toán ID3 có ý nghĩa lớn trong việc xác định những đặc điểm để phân loại gen đồng thời cho phép lựa chọn các gen đặc trưng có khả năng phân tách tốt các nhóm số liệu rõ ràng hơn. Các kết quả so sánh cho thấy phương pháp đề xuất có độ chính xác cao hơn hoặc tương đương với các công trình đã có nhưng có ưu điểm là sử dụng số lượng đặc tính ít hơn.

Tài liệu tham khảo

[1] Rampal, J.B., ed. *DNA Array Menthod and Protocol.* Vol. 170. 2001. 229-230.

[2] http://learn.genetics.utah.edu/content/labs/microarray/. [cited 2015 9/8].

[3] Fryer, R.M., et al., *Global Analysis of Gene Expression:Methods*, Interpretation, and Pitfalls. Experimental Nephrology, 2002. 10: p. 64-74.

[4] Cho, S. and H. Won, *Machine learning in DNA microarray analysis for cancer classification*. In APBC, 2003. 34: p. 189-198.

[5] Prabakaran, S., R. Sahu, and S.Verma, *Genomic signal processing using micro arrays*, submitted to hybrid system. 2005.

[6] Monson.L, *Algorithm Alley Column: C4.5*. Dr. Dobbs Journal, 1997.

[7] Ross Quinlan.J, *C4.5 Programs for Machine Learning*. Morgan Kaufmann, 1993.

[8] Khan, J., J.S. Wei, et al., *Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks*. Nature Medicine, 2001. 7: p. 673-679.

[9] http://research.nhgri.nih.gov/microarray/Supplement/. [cited 2015 9/8].

[10] Linh, T.H., ed. *Mạng nơ-rôn và ứng dụng trong xử lý tín hiệu*. ed. 1. 2014, Nhà xuất bản Bách Khoa Hà Nội.

[11] Jayakishan, M., *Mixed PCA and Wavelet Transform based Effective Feature Extraction for Efficient Tumor Classification using DNA Microarray Gene Expression Data*.

International Journal of Advanced Research in Science and Technology 2013. 2(1): p. 110-116.

[12] Zainuddin, Z. and P. Ong, *Improved wavelet neural network for early diagnosis of cancer patients using microarray gene expression data*. International Joint Conference on Neural Networks Atlanta, Georgia, USA, 2009.

PhạmTrung Dũng nhận bằng thạc sỹ về Tự động hóa và điều khiển từ xa (1997) tại Học viện KTQS, Kỹ thuật không gian (1999) tại SUPAERO (CH Pháp), tiến sỹ ngành rada-dẫn đường (2004) tại ĐH hàng không quốc gia Mat-xcơ-va (LB Nga).
Lĩnh vực quan tâm: điều khiển thiết bị bay, lý thuyết điều khiển hiện đại, xử lý ảnh và xử lý tín hiệu.

Đặng Thúy Hằng sinh năm 1981, tốt nghiệp ĐHBK Hà Nội năm 2004 chuyên ngành Điện tử Y sinh, nhận bằng Thạc sỹ chuyên ngành Tự động hóa năm 2007 (Học viện Kỹ thuật Quân sự). Hiện nay Đặng Thúy Hằng đang công tác tại Khoa Kỹ thuật Điều khiển, Học viện Kỹ thuật Quân sự.
Nghiên cứu chính là y học hạt nhân trong xạ trị.

Trần Hoài Linh sinh năm 1974, tốt nghiệp ĐHBK Vác-sa-va năm 1997 chuyên ngành Tin học ứng dụng, nhận bằng Tiến sỹ chuyên ngành Kỹ thuật điện năm 2000 (ĐHBK Vác-sa-va), bằng Tiến sỹ khoa học chuyên ngành Kỹ thuật điện và Trí tuệ nhân tạo năm 2005 (ĐHBK Vác-sa-va).
Được phong Phó Giáo sư năm 2007.

Hiện nay Trần Hoài Linh đang công tác tại Viện Điện, trường ĐHBK Hà Nội. Các nghiên cứu chính của ông là ứng dụng trí tuệ nhân tạo trong các giải pháp đo lường, điều khiển và tự động hóa, các thiết bị đo thông minh, hệ chuyên gia.