**TRANG THÔNG TIN LUẬN ÁN**

Tên đề tài luận án: Thuật toán di truyền cho bài toán nhận dạng thống kê và ứng dụng

Ngành: Lý thuyết xác suất và thống kê toán học

Mã số ngành: 9460106

Họ tên nghiên cứu sinh: Phạm Toàn Định

Khóa đào tạo: 2019-2022

Người hướng dẫn khoa học:

1. PGS.TS. Võ Văn Tài

2. TS. Lê Thị Xuân Mai

Cơ sở đào tạo: Trường Đại học Khoa học Tự nhiên, ĐHQG.HCM

**1. TÓM TẮT NỘI DUNG LUẬN ÁN**:

Luận án được trình bày trong 5 chương.

Chương đầu tiên đề nghị thuật toán di truyền trong phân tích chùm cho các phần tử rời rạc. Trong thuật toán này, việc xác định số chùm thích hợp, những phần tử trong mỗi chùm và xác suất thuộc vào mỗi chùm của mỗi phần tử được thực hiện cùng lúc. Một đóng góp quan trọng của chương này là sự vận dụng thuật toán phân tích chùm đề nghị để xây dựng mô hình chuỗi thời gian mờ. Mô hình đề nghị đã cho những kết quả nổi bật trong so sánh với nhiều mô hình trước đó, qua nhiều tập dữ liệu đối chứng. Thuật toán phân tích chùm và mô hình chuỗi thời gian đề nghị đã được xây dựng trên phần mềm Matlab nên có thể áp dụng được cho số liệu thực. Nội dung chính của chương này được lấy từ nghiên cứu [CT1] trong danh mục công trình của tác giả.

Chương 2 trình bày thuật toán di truyền mờ trong phân tích chùm cho các hàm mật độ xác suất. Thuật toán đề nghị có ba đóng góp chính trong bài toán phân tích chùm cho đối tượng này. Thứ nhất, đề xuất một chỉ số mới làm hàm mục tiêu có hiệu quả trong xây dựng chùm. Thứ hai, xây dựng các bước tìm xác suất thuộc vào các chùm đã thiết lập của mỗi hàm mật độ xác suất. Cuối cùng là ứng dụng của thuật toán đề nghị trong phân tích dữ liệu ảnh. Thuật toán đề nghị có ưu điểm hơn các thuật toán trước đó khi so sánh trên nhiều tập dữ liệu đối chứng và qua nhiều tham số đánh giá. Nội dung chính trình bày trong chương này dựa vào nghiên cứu [CT2] đã được công bố trong danh mục công trình của tác giả.

Chương 3 đề nghị thuật toán di truyền trong xây dựng chùm cho các khoảng dữ liệu. Dựa vào khoảng cách chồng lấp trong trường hợp một chiều và được cải tiến trong trường hợp nhiều chiều, luận án đã đề xuất một thuật toán mới hiệu quả cho dữ liệu khoảng. Thuật toán đề nghị có những đóng góp chính như sau: (i) Phát triển khoảng cách chồng lấp cho trường hợp nhiều chiều từ trường hợp một chiều và sử dụng nó làm độ đo đánh giá sự tương tự của các khoảng. (ii) Xây dựng thuật toán xác định số lượng chùm thích hợp cho mỗi tập dữ liệu khoảng. (iii) Tìm xác suất thuộc vào các chùm của mỗi khoảng và ứng dụng thuật toán đề nghị trong phân tích chùm cho dữ liệu ảnh khi đặc trưng của nó được trích xuất thành khoảng hai chiều. Kết quả của phần trình bày của chương này được lấy từ nghiên cứu [CT3] trong danh mục công trình của tác giả.

Chương 4 trình bày mô hình phân loại dựa trên phương pháp Bayes và thuật toán di truyền cải tiến (BGA). Mô hình đề xuất có một số đóng góp như: Xác định xác suất tiên nghiệm, lựa chọn hàm mục tiêu và tối ưu sai số Bayes. Trong BGA, tập dữ liệu huấn luyện được lựa chọn một cách tự động cho mỗi lớp đối tượng để tối ưu hóa sai số. Ví dụ số thể hiện tiềm năng và thuận lợi của mô hình BGA trong khi so sánh với các mô hình khác. Ngoài ra, mô hình đề nghị còn được ứng dụng trong một số vấn đề còn rất nhiều thách thức hiện nay. Các ví dụ số và ứng dụng đã cho thấy tính hợp lý và những ưu điểm của mô hình đề nghị so với mô hình hiện có. Nghiên cứu này đã được chỉnh sửa để đăng trên tạp chí “Annals of Operations Research”.

Chương 5 là kết luận về tính mới của luận án và hướng phát triển cho các nghiên cứu trong tương lai.

**2. NHỮNG KẾT QUẢ MỚI CỦA LUẬN ÁN**:

Luận án đã đạt được những kết quả sau đây:

* Đề xuất tiêu chuẩn đánh giá cho sự tương tự của từng đối tượng như phần tử rời rạc, hàm mật độ xác suất và khoảng dữ liệu.
* Xây dựng thuật toán xác định số chùm thích hợp cho các đối tượng khác nhau của bài toán phân tích chùm.
* Phát triển thuật toán di truyền cho bài toán phân tích chùm mờ và không mờ cho phần tử rời rạc, dữ liệu khoảng và hàm mật độ xác suất.
* Xây dựng mô hình mờ hóa chuỗi thời gian dựa vào thuật toán di truyền và bài toán phân tích chùm mờ.
* Đề xuất phương pháp phân loại mới dựa vào thuật toán di truyền và phương pháp Bayes.
* Đề xuất phương pháp trích xuất dữ liệu ảnh thành các đối tượng như khoảng, hàm mật độ, để từ đó áp dụng hiệu quả trong bài toán phân tích chùm và phân loại.

**3.** **CÁC ỨNG DỤNG/ KHẢ NĂNG ỨNG DỤNG TRONG THỰC TIỄN HAY NHỮNG VẤN ĐỀ CÒN BỎ NGỎ CẦN TIẾP TỤC NGHIÊN CỨU**

Trong thời gian tới, chúng tôi sẽ tiếp tục nghiên cứu các vấn đề sau đây:

* Phát triển thuật toán di truyền cho bài toán phân tích chùm đa mục tiêu.
* Phát triển thuật toán di truyền trong phát hiện dữ liệu bất thường, trong xây dựng mô hình dự báo khoảng cho chuỗi thời gian.
* Xem xét cải tiến và tối ưu các code được thiết lập nhằm giảm thời gian thực hiện cho các thuật toán.
* So sánh hiệu quả phân loại của mô hình đề nghị với các thuật toán trí tuệ nhân tạo như kỹ thuật học sâu.
* Ứng dụng những thuật toán phân tích chùm và phân loại đề nghị cho nhiều vấn đề thực tế, đặc biệt liên quan đến dữ liệu ảnh và dự báo.

|  |  |
| --- | --- |
| **TẬP THỂ CÁN BỘ HƯỚNG DẪN**  (Ký tên, họ tên) | **NGHIÊN CỨU SINH**  (Ký tên, họ tên) |

**PGS.TS. Võ Văn Tài TS. Lê Thị Xuân Mai Phạm Toàn Định**

**XÁC NHẬN CỦA CƠ SỞ ĐÀO TẠO**

**HIỆU TRƯỞNG**

**THESIS INFORMATION**

Thesis title: Genetic algorithm for statistical recognition problem and applications

Speciality: Probability theory and Mathematical Statistics

Code: 9460106

Name of PhD Student: Pham Toan Dinh

Academic year: 2019-2022

Supervisor:

1. Assoc. Prof., Dr. Vo Van Tai

2. Dr. Le Thi Xuan Mai

At: VNUHCM - University of Science

**1. SUMMARY**:

The thesis is presented in 5 chapters.

The first chapter proposes the genetic algorithm in clustering for discrete elements. In this algorithm, the problem determines the appropriate number of clusters, assigns elements to each cluster, and calculates the probability of each element belonging to a particular cluster. An important contribution of this chapter is the application of the proposed algorithm to build the fuzzy time series model. The proposed model shown the outstanding results in comparing with many previous models according to some benchmark datasets. The clustering algorithm and time series model have been implemented by Matlab software, allowing for their application to real datasets. The main content of this chapter is taken from the study of [CT1] in the author's article list.

Chapter 2 presents the genetic algorithm in the fuzzy clustering for the probability density functions. The proposed algorithm has three main contributions in the clustering problem. Firstly, a new index is proposed as an effective objective function in clustering for this object. Secondly, the steps to find the probability of belonging to the clusters of each probability density function are established. The final contribution is the application of the proposed algorithm in image data analysis. The proposed algorithm has more advantages than the previous algorithms according to many benchmark datasets and evaluation parameters. The main content of this chapter is taken from the study of [CT2] in the author's article list.

Chapter 3 proposes the genetic algorithms in clustering for interval data. Based on the overlapping distance in the one-dimensional case and improved for the multi-dimensional case, the study has proposed a new efficient algorithm for interval data. The proposed algorithm has the following main contributions: (i) Develop the overlap distance for the multi-dimensional case from the one-dimensional case and use it as a measure to evaluate the similarity of intervals. (ii) Build the algorithm to determine the appropriate number of clusters. (iii) Find the probability of belonging to the cluster of each interval, and apply the proposed algorithm for images. The main content of this chapter is taken from the study of [CT3] in the author's article list.

Chapter 4 shows the classification model based on the Bayesian method and the improved genetic algorithm (BGA). The proposed model has some contributions as follows: Determining the priori probability, choosing the objective function and optimizing the Bayes error. In BGA, the training dataset for each class are automatically selected to optimize the Bayes error. Numerical examples show the potential and advantages of the BGA model in comparison to other models. In addition, the proposed model is also applied in the classification of image data, a problem with many challenges at present. Numerical examples and applications have shown the rationality and outstanding advantages of the proposed model compared to others. This study has been revised for publication in the journal ``Annals of Operations Research''.

Chapter 5 is the conclusion about the novelty of the thesis and development directions for studying in the next time.

**2. NOVELTY OF THESIS**:

The thesis has achieved the following results:

* Propose measures to evaluate the similarity of elements for the different objects , including discrete elements, probability density functions and intervals.
* Build the algorithms to determine the appropriate number of clusters for different objects of clustering problem.
* Develop a genetic algorithm that can be applied to both fuzzy and non-fuzzy clustering problems. This algorithm will be utilized for clustering different types of objects, including discrete elements, probability density functions and intervals.
* Build a fuzzy time series model based on genetic algorithm and fuzzy clustering problem.
* Propose the new classification method based on genetic algorithm and Bayesian method.
* Propose the method to extract image data into objects such as intervals, probability density functions and then apply effectively in clustering and classification problems.

**3**. **APPLICATIONS/ APPLICABILITY/ PERSPECTIVE**

In next time, we will study some problems as follows:

* Developing the genetic algorithm for multi-objective clustering problem.
* Developing the genetic algorithms in anomaly data detection, in building interval prediction models for time series.
* Considering improvement and optimization for code sources to decrease the running time of various algorithms.
* Comparing the classifying performance of the proposed algorithm to algorithms of artificial intelligence such as deep learning technique.
* Applying the proposed clustering and classification algorithms to many real problems, especially related to image and prediction.

|  |  |
| --- | --- |
| **SUPERVISOR** | **PhD STUDENT** |

Assoc. Prof., Dr. Vo Van Tai Dr. Le Thi Xuan Mai Pham Toan Dinh

**CERTIFICATION**

**UNIVERSITY OF SCIENCE**

**PRESIDENT**