**TÓM TẮT THÔNG TIN VỀ LUẬN ÁN**

Tên đề tài luận án: Một số đóng góp trong bài toán phân tích chùm

Ngành: Lý thuyết xác suất và thống kê toán học

Mã số ngành: 62460106

Họ tên nghiên cứu sinh: Nguyễn Trang Thảo

Khóa đào tạo: 2016

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 này được thực hiện nhằm tiếp tục giải quyết các khe hở nghiên cứu đang tồn tại trong các bài toán phân tích chùm cho dữ liệu rời rạc (CDE), phân tích chùm cho các hàm mật độ xác suất (CDF), và phân tích chùm cho dữ liệu khoảng (CID). Cụ thể, luận án đề xuất các thang đo mới và các thuật toán mới hiệu quả hơn các thuật toán đã tồn tại; phân tích tính hội tụ của các thuật toán đã đề xuất; và ứng dụng các thuật toán này vào một số lĩnh vực cụ thể. Nội dung chính của luận án được trình bày trong 3 chương:

* **Chương 1** trình bày một số đề xuất mới trong bài toán CDE như: hệ số tương tự chùm cho dữ liệu rời rạc và phân tích chùm tự động cho dữ liệu rời rạc dựa trên hệ số tương tự chùm. Chương này cũng chứng minh tính hội tụ của thuật toán mới, và trình bày một số ứng dụng.
* **Chương 2** trình bày một số đề xuất mới trong bài toán CDF như: hệ số tương tự chùm cho các hàm mật độ xác suất, phân tích chùm tự động, phân tích chùm dựa trên tiêu chuẩn hệ số tương tự chùm; phân tích chùm mờ, phân tích chùm dựa trên giải thuật di truyền. Ngoài ra, chương này cũng chứng minh tính chất hội tụ của các thuật toán trên và trình bày một số ứng dụng của bài toán CDF.
* **Chương 3** trình bày một số đề xuất mới trong bài toán CID như: phân tích chùm tự động cho dữ liệu khoảng dựa trên khoảng cách chồng lấp và giải thuật di truyền, phân tích chùm tự động cho dữ liệu khoảng có chứa dữ liệu bất thường. Ngoài ra, chương này cũng làm rõ tính chất hội tụ của các thuật toán đề xuất và trình bày một số ứng dụng cụ thể.

**2. NHỮNG KẾT QUẢ MỚI CỦA LUẬN ÁN:**

Luận án đã thu được các kết quả mới sau:

* Đề xuất 01 tiêu chuẩn mới để đánh giá sự tương tự của các phần tử rời rạc và đánh giá kết quả phân vùng của một phương án phân tích chùm các phần tử rời rạc.
* Đề xuất 01 tiêu chuẩn mới để đánh giá sự tương tự của các hàm mật độ xác suất và 01 tiêu chuẩn dùng để đánh giá mức độ tốt của một phương án phân tích chùm các hàm mật độ xác suất. Tiêu chuẩn mới là một thang đo nằm trong đoạn [0,1] nên có thể khắc phục được một số hạn chế hiện có của tiêu chuẩn độ rộng chùm (nằm trong đoạn [0, *k*-1], phụ thuộc vào số chùm *k*).
* Đề xuất mới 01 thuật toán phân tích chùm các phần tử rời rạc. So với các thuật toán đã có, vốn cần xác định trước số chùm bởi người dùng, thuật toán đề xuất có khả năng tự động xác định số chùm hợp lý cho từng bộ dữ liệu cụ thể nhưng sử dụng nhiều chi phí tính toán hơn. Do đó, thuật toán mới nên được sử dụng trong trường hợp phân tích dữ liệu hoàn toàn mới, chưa có các thông tin định trước về số chùm.
* Đề xuất 03 thuật toán phân tích chùm các hàm mật độ xác suất, bao gồm: phân tích chùm sử dụng tiêu chuẩn hệ số tương tự chùm, phân tích chùm mờ, và phân tích chùm sử dụng giải thuật di truyền. Trong đó, thuật toán phân tích chùm mờ cho phép một hàm mật độ xác suất phụ thuộc đồng thời vào nhiều chùm với các mức độ khác nhau. Do đó, kết quả thu được là tổng quát hơn so với các thuật toán hiện có, vốn theo hướng tiếp cận không mờ.
* Đề xuất 02 thuật toán phân tích chùm cho dữ liệu khoảng, bao gồm: phân tích chùm tự động sử dụng giải thuật di truyền và phân tích chùm tự động cho dữ liệu khoảng bất thường. Các thuật toán mới có thể tự động xác định số chùm và tách các dữ liệu khoảng bất thường thành các nhóm riêng biệt với chi phí tính toán thấp.

Các kết quả chính của luận án được tổng hợp từ các bài báo đã được công bố trên các tạp chí uy tín bao gồm 6 bài trên các tạp chí SCIE và 1 bài trên tạp chí ESCI.

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

Các kết quả của luận án có tiềm năng vào một số bài toán thực tế như phân tích dữ liệu, phân nhóm hình ảnh, và nhận diện các hình ảnh bất thường. Tuy nhiên, độ tin cậy của các phương pháp trên các ứng dụng này cần được tiếp tục kiểm chứng thêm. Ngoài ra, một số vấn đề lý thuyết như tính chất hội tụ của một thuật toán vẫn còn bỏ ngỏ.

|  |  |
| --- | --- |
| **TẬP THỂ CÁN BỘ HƯỚNG DẪN** | **NGHIÊN CỨU SINH** |

**XÁC NHẬN CỦA CƠ SỞ ĐÀO TẠO**

**HIỆU TRƯỞNG**

**THESIS INFORMATION**

Thesis title: Some contributions to the clustering problem

Speciality: Theory of probability and mathematical statistics

Code: 62460106

Name of PhD Student: Nguyen Trang Thao

Academic year: 2016

Supervisor:

1. Assoc. Prof. Dr. Vo Van Tai
2. Dr. Le Thi Xuan Mai

At: VNUHCM -University of Science

**1. SUMMARY:**

The purpose of this thesis is to overcome current research gaps in the clustering problem for discrete data (CDE), probability density functions (CDF), and interval data (CID). Specifically, the thesis presents new, more efficient criteria and algorithms; evaluates the convergence of the proposed algorithms; and applies these algorithms to specific domains. The thesis's primary material is delivered in three chapters:

- Chapter 1 introduces new proposals for the CDE problem, such as the similar coefficient for cluster of discrete elements and the automatic clustering for discrete elements based on the similar coefficient. In addition, this chapter proves the convergence of the new algorithm and provides examples of its applications.

- Chapter 2 presents new proposals for the CDF problem, such as the similar coefficient for cluster of probability density functions, the automatic clustering algorithm, the non-hierarchical clustering algorithm based on the similar coefficient; the fuzzy clustering algorithm, and clustering based on the genetic algorithms. In addition, this chapter proves the convergence properties of the above algorithms and illustrates some applications.

- Chapter 3 presents some new proposals in for CID problem, such as the automatic clustering based on the overlapping distance and the genetic algorithm, and the automatic clustering for interval data with outliers. In addition, this chapter clarifies the convergence properties of the proposed algorithms and illustrates some applications.

**2. NOVELTY OF THESIS:**

The novelties of this thesis are listed as follows:

- Proposing 1 new criterion to evaluate the similarity of discrete elements and the partitioning results of a cluster of discrete elements.

- Proposing 1 new criterion to evaluate the similarity of probability density functions and the partitioning results of a cluster of probability density functions. This is a normalized score in the range [0,1], therefore, it can overcome existing limitations of the cluster-width criterion, which lies in the range [0, *k*-1] and depends on *k,* the number of clusters.

- Proposing 1 new clustering algorithm for discrete elements. Compared with existing algorithms, requiring a given number of clusters, the proposed algorithm is capable of automatically determining the reasonable number of clusters for each particular data set, but using more computational cost. Therefore, the new algorithm should be used to analyze new data sets, without predetermined information about the number of clusters.

- Proposing 3 clustering algorithms for probability density functions, including the non-hierarchical clustering algorithm based on the similar coefficient, the fuzzy clustering, and clustering based on the genetic algorithm. The fuzzy clustering approach permits a probability density function to simultaneously and to varying degrees belong to all groups. Consequently, the findings obtained are more general than those of existing non-fuzzy methods.

- Proposing 2 clustering algorithms for interval data, including the automatic clustering based on the genetic algorithm, and the automatic clustering for interval data with outliers. The new algorithm can automatically determine the number of clusters and separate outliers into separate groups with a low computational cost.

The principal findings of the thesis are based on publications published in reputable journals, including 6 SCIE articles and 1 ESCI article.

**3. APPLICATIONS/ APPLICABILITY/ PERSPECTIVE**

The thesis findings have the potential to be applied to a variety of practical issues, including data analysis, image categorization, and abnormal image detection. However, the reliability of the approaches for these applications must be examined further. In addition, the convergence properties of the non-hierarchical clustering algorithms based on the similar coefficient are still open.

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

**CONFIRMATION UNIVERSITY OF SCIENCE**

**PRESIDENT**