Model-Based Clustering and Classification for Data Science

Nguyen Manh Khiem

2023-10-17

Project này nhằm mục đích tóm tắt cuốn sách **Model-Based Clustering and Classification for Data Science** của Charles BOUVEYRON, Gilles CELEUX, T. Brendan MURPHY và Adrian E. RAFTERY

Chapter 1: Introduction

- Nhiều bộ dataset không thể detected bằng mắt người nên nó đòi hỏi automated algorithms.
- Nhiều thuật toán chỉ mang tính **heuristic** (theo kinh nghiệm, trực giác) ít reference to the statistical theory.
- Vào những năm 1960, người ta nhận ra rằng phân tích cụm có thể được áp dụng dựa trên **principled statistical basis** (cơ sở có nguyên tắc) bằng cách chuyển đổi góc nhìn clustering task thành inference (suy luận) for a finite mixture model.

Finite mixture model là mô hình thống kê mà trong đó người cài đặt giả sử dataset là tổ hợp của một số ít các quy tắc, objects nhất định. Ví dụ thu nhập của người dân có thể được chia thành 3 loại: thu nhập cao/trung bình/thấp. Những quy tắc (pattern) này tuân theo một phân phối xác suất nhất định. Do đó, ta có thể xem mỗi cluster tương ứng với một phân phối xác suất. Tuy nó phụ thuộc lớn vào domain knowledge nhưng nó lại giúp đưa nền tảng toán học thống kê vào.

Mục đích cuốn sách:

- Review model-based approach emerged in the past half-century, and active research field.
- Describe the basic ideas, and aim to show the advantages of thinking in this way.
- Review recent developments, particularly for newer types of data such as high-dimensional data, network data, textual data and image data.

1.1 Cluster Analysis

- Find meaningful groups.
- Những điểm chung 1 nhóm sẽ có tính cohesive (liên kết) and từng nhóm sẽ có tính separated (độc lập) với nhau.
- The purpose is to find groups whose members have something in common that they do not share with members of other groups.

1.1.1 From Grouping to Clustering

Giải thích cụm từ "model-based"

Có 2 hướng tiếp cận:

1. **Modern/Model-based approach**: Sử dụng các mô hình toán học như Mô hình tăng trưởng dân số, hoặc mô hình thống kê như Hồi quy tuyến tính, ANOVA, Time Series, v.v. để mô tả và giải thích mối liên hệ giữa các biến. Mô hình này sau đó có thể được sử dụng để dự đoán hoặc suy luận về dữ liệu.
2. **Traditional approach**: Sử dụng các công thức thống kê mô tả, suy diễn, và các phương pháp kiểm định để phân tích dữ liệu. Phương pháp này không nhất thiết phải dựa trên một mô hình cụ thể mà còn có thể tích hợp, áp dụng thêm vào các mô hình.

Tham khảo thêm:

Sự khác biệt giữa mô hình thống kê (statistical models) và mô hình học máy (machine learning model)

https://viblo.asia/p/su-khac-biet-giua-mo-hinh-thong-ke-statistical-models-va-mo-hinh-hoc-may-machine-learning-models-QpmleRqM5rd

Lịch sử:

- 1735-1758: Linnaeus đã tiến hành phân loại thực vật hầu hết dựa trên trực giác.
- 1757-1763: Adanson bắt đầu dựa trên các đặc tính của objects.
- 1990: Greene định nghĩa khái niệm gom nhóm theo triết học, tiếp sau đó là Plato, trong đó, Plato lấy ví dụ về hình tượng của một người thợ rèn phải đi kèm với hình ảnh hammer.
- **1909-1932**: định nghĩa khái niệm cluster dựa trên difference and similarity between objects + định nghĩa hệ thống các systematic numerical methods (các phương pháp ước lượng, xấp xỉ) cho quantitative data.
    - 1931: Hotelling's T-squared distribution
- **1936**:
    - Stephenson use of factor analysis to identify clusters of people. This was seems to be the first book on cluster analysis.
    - Fisher's discriminant analysis or linear discriminant analysis (LDA)
- **1938**:
    - Cluster sử dụng correlation matrix.
    - Fisher - Analysis of variance (ANOVA)
- **1939**:
    - Multiple group factor analysis.
    - Welch - class-conditional distributions in the case of discriminant analysis. Normal distributions, using either Bayes' theorem, (if the prior probabilities of the classes are known) or the Neyman–Pearson lemma (if these prior probabilities have to be estimated).
- 1944: Cattell Graphical clustering.
- **1950s**: multivariate discrete data
- **1957**: Sneath - Hierarchical clustering (single link)
- **1958**:
    - Sokal và Michener average link method and complete link method.
    - Cox: logistic regression for binary classification.
    - Rosenblatt: Perceptron. Not be able to recognize many classes without adding several layers.

Thời điểm này tuy nó là một lĩnh vực cần thiết, nhưng nó lại phát triển rất chậm vì giới hạn của máy tính.

- **1963**:
    - Cuốn sách của Sokal và Sneath led to a rapid expansion of the use and methodology of cluster analysis. The dominant model for clustering continuous-valued data is the mixture of multivariate normal distributions (phân cụm dữ liệu có giá trị liên tục là hỗn hợp các phân phối chuẩn đa biến), first mentioned by Wolfe in his Master's thesis at Berkeley.
    - Cortes and Vapnik - Support vector machines (SVM). Transform the original data in a high-dimensional space, through a nonlinear projection, where they are linearly separable with a hyperplane. Handle data of various types thanks to the notion of kernel.
    - John Wolfe subsequently developed the first real software for estimating this model, called NORMIX, and also developed related theory (Wolfe, 1965, 1967, 1970), so he has a real claim to be called the inventor of model-based clustering for continuous data.
    - Estimating the model by maximum likelihood using the EM algorithm
- 1970s và 1980s:
    - Dempster et al modified EM algorithm, remains the most used estimation approach in model-based clustering (See Section 2.9).
- 1976: variable selection to avoid the curse of dimensionality in discriminant analysis
- 1982: McLachlan and Ganesalingam use unlabeled data to update a classification rule in order to reduce the classification error
- **1988**:
    - Blashfeld và Aldenderfer image analysis.
    - LeCun et al., 1998: Convolutional neural networks
- **1990s**:
    - Larger data. Online learning.
- 1992:
    - SVM first implementation, thanks to the "kernel trick" of Boser et al.
    - Celeux and Mkhadri regularized discriminant analysis technique for high-dimensional discrete data
- 1993 & 1996:
    - Banfield, Raftery & Bensmail, Celeux constrained Gaussian models
    - Hastie and Tibshirani classification non-normal data using mixtures of Gaussians
- 1997: EM Algorithm
- 1999: Fisher's discriminant subspace.

1.1.2 Model-based Clustering

Tuy nhiên khi sử dụng các thuật toán cluster này, ta không thể trả lời được một số câu hỏi như:

1. How many clusters are there?
2. What is the best clustering algorithm?
3. How should we deal with outliers?
4. How sure are we of a clustering partition, and how should we assess uncertainty about it? (Đánh giá thuật toán như thế nào? Độ ổn định là bao nhiêu?)

Những câu hỏi này có thể được trả lời bằng cách sử dụng các độ đo và kiểm định thống kê. Những công thức đã được chứng minh kỹ càng và có độ tin cậy cao như Elbow, Bayesian Information Criterion (BIC) hoặc Akaike Information Criterion (AIC),...

Vậy tóm lại khi nhắc đến model based hay finite mixture model, chúng ta sẽ hiểu rằng:

- Dùng các thuật toán clustering để gom cụm.
- Dùng phương pháp thống kê để tìm xem các cụm này khớp với phân phối xác suất nào đó mà phương pháp đặt ra (ví dụ Maximum likelihood estimation).

1.2 Classification/Discriminant analysis

- Already have the information about the nature of the classes/already been classified by experts.
- Determine which class new objects belong to
- Supervised problem.

1.5 Organization of the Book

- Chapter 2: Introduce the basic ideas of model-based clustering - Unsupervised data.
- Chapter 3: Descripe difficulty of framework, outliers, degeneracies and non-Gaussian mixture components.
- Chapter 4: Describe model-based approaches to classification - Supervised data
- Chapter 5 we extend this to discuss semi-supervised classification, in which unlabeled data are used as part of the training data.
- Chapter 6: model-based clustering for discrete data. Consider ordinal data, and data of mixed type, i.e. that include both discrete and continuous variables.
- Chapter 7: we consider the selection of variables for clustering, for both continuous and discrete data.
- Chapter 8: Describe methods for model-based clustering for high-dimensional data. Dimension reduction, regularization and subspace methods.
- Chapter 9 Describes ways of clustering data where the component distributions are non-Gaussian by modeling them explicitly, in contrast with the component merging methods of Chapter 3.
- Chapter 10: Describe model-based approaches to clustering nodes in networks, with a focus on social network data.
- Chapter 11: treats methods for model-based clustering with covariates, with a focus on the mixture of experts model.
- Chapter 12: Describe model-based clustering methods for a range of less standard kinds of data. These include functional data, textual data and images. They also include data in which both the rows and the columns of the data matrix may have clusters and we want to detect them both in the same analysis. The methods we describe fall under the heading of model-based co-clustering.

Tổng kết:

- Phân biệt được thống kê và Machine Learning. Ưu điểm và hạn chế
- Các thuật toán Clustering thuộc lĩnh vực nào? Ưu điểm và hạn chế.
- Một số cột mốc quan trọng
- Nắm được khái niệm Model-based, finite mixture model.
- Phân biệt được Cluster/Classification.
- Phân biệt supervised/unsupervised problem.