Biên soạn: PGS.TS. Đỗ Thanh Nghị (Chủ biên)
PGS.TS. Phạm Nguyên Khang

GIÁO TRÌNH
KHAI THÁC DỮ LIỆU VỚI
PYTHON

NHÀ XUẤT BẢN ĐẠI HỌC CẦN THƠ
2022

BIÊN MỤC TRƯỚC XUẤT BẢN THỰC HIỆN BỞI
TRUNG TÂM HỌC LIỆU TRƯỜNG ĐẠI HỌC CẦN THƠ

Đỗ, Thanh Nghị
Giáo trình khai thác dữ liệu với Python / Đỗ Thanh Nghị (Chủ biên), Phạm Nguyên Khang.- Cần Thơ: Nxb. Đại học Cần Thơ, 2022.
182 tr. : minh họa ; 24 cm
Sách có danh mục tài liệu tham khảo
ISBN: 9786049656996
1. Python (Computer program language) 2. Khai thác dữ liệu
I. Nhan đề. II. Phạm, Nguyên Khang
005.133– DDC 23 MFN 244891
Ngh300

LỜI GIỚI THIỆU

Nhằm góp phần làm phong phú nguồn tư liệu phục vụ nghiên cứu, học tập cho bạn đọc, sinh viên, học viên và nghiên cứu ngành Công nghệ thông tin và Truyền thông. Nhà xuất bản Đại học Cần Thơ ấn hành và giới thiệu cùng bạn đọc giáo trình "Khai thác dữ liệu với Python" do PGS.TS. Đỗ Thanh Nghị, PGS.TS. Phạm Nguyên Khang biên soạn.

Giáo trình gồm 12 chương, nội dung giới thiệu về khám phá tri thức và khai thác dữ liệu; giải thuật k láng giềng; giải thuật Bayes thơ ngây; cây quyết định; phương pháp tập hợp mô hình; máy học véctơ hỗ trợ; giải thuật luật kết hợp; hiển thị dữ liệu và phương pháp giảm chiều; kết luận và hướng phát triển; ôn tập xác suất thống kê; giới thiệu ngôn ngữ lập trình python.

Nhà xuất bản Đại học Cần Thơ chân thành cám ơn các tác giả và sự đóng góp ý kiến của quý thầy cô trong Hội đồng thẩm định Trường Đại học Cần Thơ để giáo trình "Khai thác dữ liệu với Python" được ra mắt bạn đọc.

Nhà xuất bản Đại học Cần Thơ trân trọng giới thiệu đến học viên, sinh viên, giảng viên và bạn đọc giáo trình này.

NHÀ XUẤT BẢN ĐẠI HỌC CẦN THƠ

LỜI NÓI ĐẦU

Khám phá tri thức và khai thác dữ liệu là chuyên ngành rất quan trọng cho chương trình đào tạo cử nhân, kỹ sư và thạc sĩ công nghệ thông tin, trong thời đại của cuộc Cách mạng công nghiệp 4.0. Khám phá tri thức và khai thác dữ liệu là tập hợp những phương pháp, công cụ từ các chuyên ngành xác suất thống kê, phân tích dữ liệu, máy học, trí tuệ nhân tạo, hiển thị dữ liệu và cơ sở dữ liệu, nhằm khám phá tri thức tiềm ẩn từ dữ liệu lớn. Trong thời gian giảng dạy, chúng tôi nhận thấy có quá ít tài liệu tham khảo bằng tiếng Việt về khám phá tri thức và khai thác dữ liệu. Chính vì lý do đó, chúng tôi đã thực hiện biên soạn quyển giáo trình "Khai thác dữ liệu với Python" nhằm mục đích cung cấp thêm tài liệu tham khảo bằng tiếng Việt đến độc giả là sinh viên nhóm ngành công nghệ thông tin.

Khám phá tri thức và khai thác dữ liệu là lĩnh vực nghiên cứu rộng lớn, chúng tôi chỉ tập trung vào kiến thức cơ bản về quá trình khám phá tri thức và các giải thuật khai thác dữ liệu phổ biến nằm trong top 10 giải thuật hiệu quả nhất được bình chọn từ cộng đồng khám phá tri thức và khai thác dữ liệu. Từng giải thuật khai thác dữ liệu được trình bày ngắn gọn, tiếp theo sau là ví dụ minh họa giải thuật bằng ngôn ngữ **Python**. Việc lựa chọn ngôn ngữ **Python** để minh họa các giải thuật vì những lý do như: **Python** tương thích với giấy phép phần mềm miễn phí nguồn mở, rất dễ học, hỗ trợ nhiều giải thuật khai thác dữ liệu. Chúng tôi hy vọng quyển giáo trình ngoài việc cung cấp các kiến thức cơ bản của môn học còn giúp cho độc giả trau dồi kỹ năng thực hành khám phá tri thức và khai thác dữ liệu qua các ví dụ và bài tập minh họa trong **Python**.

Trong thời gian biên soạn giáo trình, chúng tôi nhận được sự hỗ trợ quý báu về vật chất, tinh thần từ Khoa Công Nghệ Thông Tin và Truyền Thông, Trường Đại học Cần Thơ. Chúng tôi đặc biệt gửi lời cám ơn đến Quý Thầy Cô đã nhiệt tình góp ý cho bản thảo.

Do thời gian có hạn và lần đầu tiên biên soạn, tài liệu không thể tránh khỏi những thiếu sót. Chúng tôi mong nhận được sự góp ý chân thành từ quý độc giả để quyển sách ngày được hoàn thiện hơn.

Cần Thơ, tháng 9 năm 2022

CÁC TÁC GIẢ

MỤC LỤC

DANH MỤC HÌNH

DANH MỤC BẢNG

TỪ VIẾT TẮT VÀ CÁC THUẬT NGỮ

KDD (Knowledge discovery in databases): khám phá tri thức
Classification: phân lớp (học có giám sát)
Clustering: gom cụm, gom nhóm (học không giám sát)
DM (Data mining): khai thác dữ liệu
Ensemble-based method: phương pháp tập hợp mô hình
Feature space: không gian trung gian hay không gian đặc trưng
Free open source software: phần mềm tự do
Gini index: chỉ số Gini
Hierarchical clustering: mô hình gom cụm phân cấp
Hyperplane: siêu phẳng
Impurity: hỗn loạn
Information gain: độ lợi thông tin
Information visualization: hiển thị dữ liệu
Kernel: hàm nhân
***k*NN** (k nearest neighbors): *k* láng giềng
Laplace estimator: ước lượng Laplace
Machine learning: máy học
MDS (Multidimensional scaling): phương pháp giảm chiều MDS
Naive Bayes: Bayes thơ ngây
Overfitting: học vẹt
PCA (Principal component analysis): phương pháp phân tích thành phần chính
Purest: thuần khiết nhất
Random forest: rừng ngẫu nhiên
Regression: hồi quy
Stress function: hàm độ đo biến dạng
Supervised learning: học có giám sát
SVM (Support vector machines): máy học véctơ hỗ trợ
Train, Training, Learn, Learning: huấn luyện, học
Underfitting: học không thuộc
Unsupervised learning: học không giám sát

KHAI THÁC DỮ LIỆU VỚI PYTHON

Chương 1
GIỚI THIỆU VỀ KHÁM PHÁ TRI THỨC VÀ KHAI THÁC DỮ LIỆU

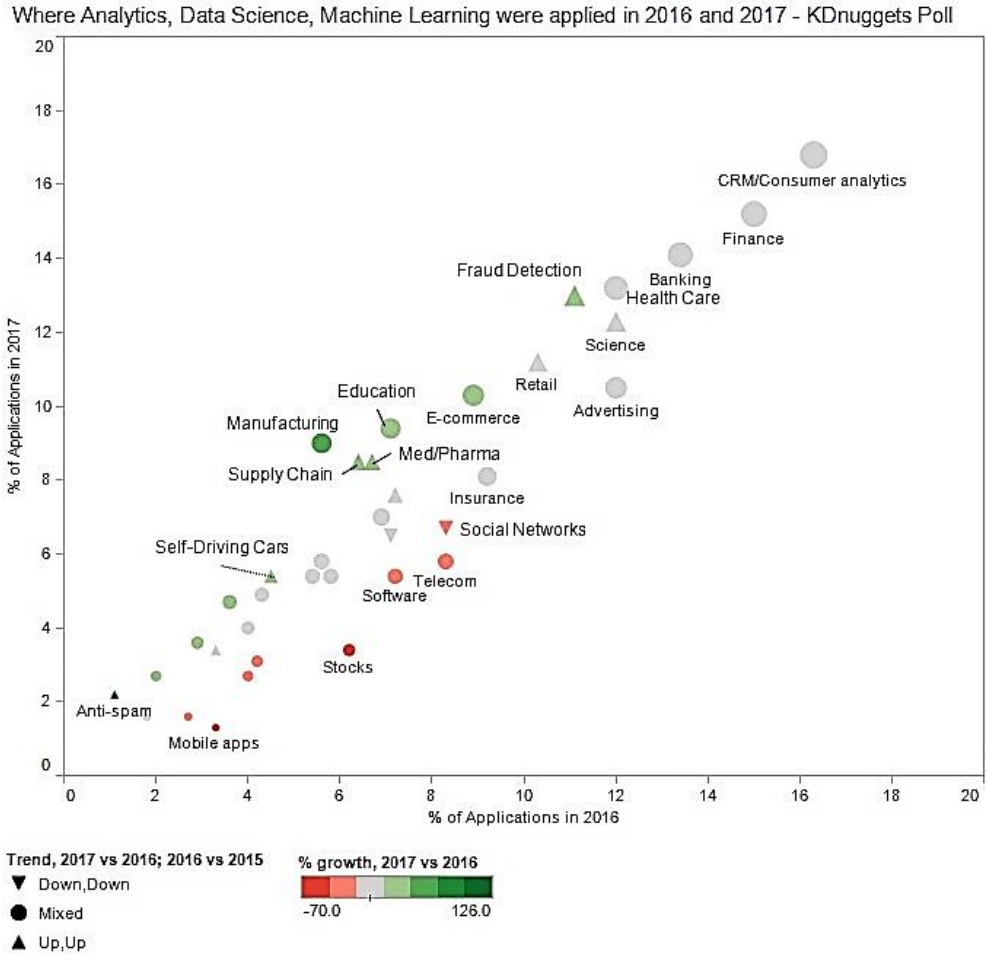
Trong những năm 1990, cuộc cách mạng kỹ thuật số cho phép số hóa thông tin dễ dàng và chi phí thấp, thêm vào đó là sự phát triển của công nghệ thông tin bao gồm cả phần cứng lẫn phần mềm, công nghệ truyền thông, web, internet đã góp phần đưa máy tính vào các sinh hoạt thường nhật của con người. Tất cả các hoạt động kinh doanh, vui chơi giải trí, nghiên cứu khoa học, giáo dục, truyền thông đều có sự hỗ trợ của máy tính. Hệ quả kéo theo khối lượng lớn dữ liệu được sinh ra và lưu trữ trong các cơ sở dữ liệu, thiết bị lưu trữ như băng từ, đĩa từ. Từ năm 1999, Giáo sư P. Lyman và các cộng sự của ông ở Đại học Berkeley đã tiến hành thống kê dữ liệu được sinh ra hằng năm trên toàn cầu. Kết quả chỉ trong năm 2002-2003 (tham khảo ở địa chỉ http://www.sims.berkeley.edu/research/-projects/how-much-info-2003), dữ liệu toàn cầu tăng 5 Exabytes (5.10^{18} bytes). Sự bùng nổ dữ liệu có thể thấy ở các kho dữ liệu như:

- Các vệ tinh quan sát trái đất hiện thời của NASA tạo ra khoảng 1 terabyte (10^{12} bytes) dữ liệu mỗi ngày. Lượng dữ liệu này lớn hơn lượng dữ liệu của tất cả các vệ tinh quan sát trước đây cộng lại.
- Dự án hệ gen người đang lưu trữ hàng ngàn byte dữ liệu cho mỗi cơ số gen trong số hàng tỉ cơ số.
- Rất nhiều công ty hiện đang duy trì các kho dữ liệu khổng lồ về các giao dịch của khách hàng. Một kho dữ liệu tương đối nhỏ thôi cũng có thể lưu trữ hơn một trăm triệu giao dịch.
- Một lượng lớn dữ liệu được ghi lại mỗi ngày dùng các thiết bị ghi nhận tự động như các giao dịch thẻ tín dụng, nhật ký web, v.v.

Tóm lại, U. Fayyad et al. [9] ước lượng dữ liệu toàn cầu tăng gấp đôi trong vòng 9 tháng. Vấn đề đặt ra là làm sao chúng ta có thể rút trích tri thức quan trọng từ các kho dữ liệu khổng lồ. Các tri thức phục vụ cho các tổ chức, cơ quan, công ty bao gồm việc phát hiện quan trọng trong khoa học, các dự báo chính xác về thời tiết và các thảm họa tự nhiên, những tri thức cho phép ta xác định được nguyên nhân và phương pháp điều trị các căn bệnh hiểm nghèo,... Sự ra đời của công nghệ khám phá tri thức và khai thác dữ liệu [7] nhằm đáp ứng nhu cầu cần thiết của các tổ chức, cơ quan, công ty về phát hiện tri thức từ các kho dữ liệu khổng lồ.

Các ứng dụng thành công của công nghệ khai thác dữ liệu có thể tìm thấy trong rất nhiều lĩnh vực như: tiếp thị, ngân hàng, bảo hiểm, y tế, sinh học, phát hiện gian lận, tìm kiếm thông tin, lọc thư rác, phân loại văn bản (xem hình 1.1).

Tạp chí về công nghệ của trường MIT số ra tháng 1-2 năm 2001 cho rằng khai thác dữ liệu là một trong 10 công nghệ nổi bật nhất của thế kỷ XXI.

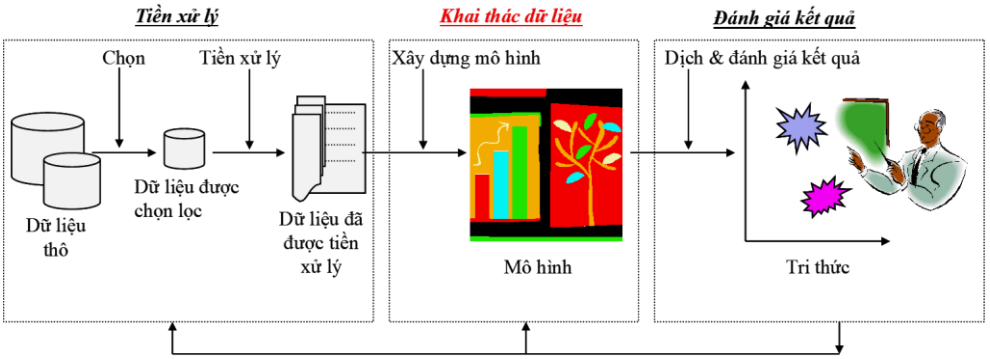


Hình 1.1 Lĩnh vực ứng dụng thành công của khai thác dữ liệu
(Nguồn: KDNuggets)

1.1 KHÁM PHÁ TRI THỨC VÀ KHAI THÁC DỮ LIỆU

Theo Fayyad et al. [7], công nghệ khám phá tri thức từ dữ liệu được định nghĩa là sự trích xuất từ dữ liệu những thông tin hữu ích nhưng tiềm ẩn và chưa được biết đến. Khai thác dữ liệu là một bước quan trọng trong quá trình khám phá tri thức từ dữ liệu. Khai thác dữ liệu thực hiện việc khảo sát,

phân tích tỉ mỉ một lượng lớn dữ liệu nhằm phát hiện ra các mẫu hoặc các luật có ý nghĩa.



Hình 1.2 Quá trình khám phá tri thức

Quá trình khám phá tri thức như mô tả trong hình 1.2 là một quá trình lặp phức tạp, sử dụng nhiều kỹ thuật (hình 1.3) như cơ sở dữ liệu, máy học, phương pháp thống kê trong phân tích dữ liệu, hiển thị dữ liệu, trí tuệ nhân tạo, nhằm tìm ra những tri thức từ kho dữ liệu khổng lồ.

Hình 1.3 Các lĩnh vực liên quan đến khám phá tri thức và khai thác dữ liệu

Quá trình khám phá tri thức bao gồm 3 bước chính: tiền xử lý, khai thác dữ liệu và đánh giá kết quả. Từ mục tiêu đề ra của ứng dụng, ở bước tiền xử lý chúng ta cần thực hiện:

- Tập hợp dữ liệu từ nguồn dữ liệu khác nhau,
- Chọn dữ liệu cần thiết cho mục tiêu đề ra, mẩu tin, trường dữ liệu,

- Biểu diễn dữ liệu, chuyển đổi kiểu sao cho phù hợp với giải thuật khai thác dữ liệu mà bước tiếp theo sử dụng,
- Làm sạch dữ liệu, khắc phục đối với trường dữ liệu rỗng, dư thừa, hoặc dữ liệu không hợp lệ, có thể tinh giảm dữ liệu hơn.

Sau khi đã tiền xử lý dữ liệu xong, đến bước khai thác dữ liệu tiến hành xây dựng các mô hình với sự hỗ trợ của:

- Máy học,
- Trí tuệ nhân tạo,
- Phân tích dữ liệu nhiều chiều bằng phương pháp thống kê
- Hoặc bằng phương pháp trực quan hiển thị dữ liệu.

Các giải thuật khai thác dữ liệu được sử dụng nhiều trong cộng đồng khám phá tri thức bao gồm: luật kết hợp [1], *k* láng giềng [10], phân lớp Bayes thơ ngây [12], cây quyết định [2, 21], Bagging[3], Boosting [11], máy học véctơ hỗ trợ [22], rừng ngẫu nhiên [4], gom cụm *k*Means [19] (tham khảo nguồn KDNuggets, hình 1.4).

Bước khai thác dữ liệu được xem là trung tâm của quá trình khám phá tri thức. Công việc rất phức tạp, lặp đi lặp lại các công việc như: xây dựng mô hình, tạo tri thức từ dữ liệu, kiểm định lại mô hình, nếu chưa đạt thì phải xây dựng mô hình khác.

Khai thác dữ liệu tập trung giải quyết các vấn đề cơ bản như phân lớp (classification, supervised classification), hồi quy (regression), gom nhóm (clustering, unsupervised classification) và luật kết hợp (association rules).

Phân lớp: xây dựng mô hình phân loại dựa trên tập dữ liệu học có nhãn (lớp). Ví dụ như chúng ta có sẵn tập dữ liệu thư điện tử, ***mỗi thư có nhãn là thư rác hay thư bình thường***, mục tiêu là xây dựng mô hình phân lớp tập dữ liệu thư điện tử thành thư rác hay thư bình thường để khi có một thư điện tử mới đến thì mô hình dự báo được thư này có phải là thư rác hay không.

Hồi quy: xây dựng mô hình dự đoán dựa trên tập dữ liệu học có gán nhãn (lớp) là giá trị liên tục. Ví dụ như người ta cần xây dựng mô hình dự báo mực nước sông Mêkông (kiểu số thực) từ các yếu tố như thời tiết, mùa.

Gom nhóm: xây dựng mô hình gom cụm tập dữ liệu học (không có nhãn) sao cho các dữ liệu cùng nhóm có các tính chất tương tự nhau và dữ liệu của 2 nhóm khác nhau có các tính chất khác nhau. Ví dụ như chúng ta cần gom nhóm học sinh trong một lớp sao cho các học sinh cùng nhóm học giỏi cùng ban (khoa học xã hội, khoa học tự nhiên). Gom nhóm cũng được biết đến như vấn đề học không giám sát.

Luật kết hợp: phát hiện mối liên quan giữa các biến của dữ liệu, chẳng hạn luật kết hợp có thể phát hiện quy luật đồng xuất hiện từ đó suy luận được luật như ***nếu một khách hàng mua bơ, bánh mì thì khách hàng cũng mua sữa***.

Sau bước khai thác dữ liệu, tiếp đến là phải đánh giá tri thức sinh ra từ việc tiền xử lý và khai thác dữ liệu: kiểm định kết quả dựa vào mục tiêu ban đầu của ứng dụng. Nghĩa là chỉ có người sử dụng hoặc chuyên gia về lĩnh vực mới có khả năng đánh giá được tri thức sinh ra. Chính vì vậy kết quả sinh ra từ quá trình khám phá tri thức cần dễ hiểu, dễ diễn dịch kết quả để giúp người sử dụng hoặc chuyên gia có thể đánh giá và hiểu được kết quả sinh ra. Nếu kết quả không đạt được so với mục tiêu đề ra, người ta có thể quay lại các bước tiền xử lý hay khai thác dữ liệu để lặp lại quá trình khám phá tri thức.

Trong phạm vi của quyển giáo trình, chúng tôi trình bày các kỹ thuật cơ bản được dùng cho công nghệ khám phá tri thức và khai thác dữ liệu. Tài liệu tham khảo chi tiết có thể được tìm thấy ở [6, 14, 15, 24]. Để thuận lợi cho độc giả, chúng tôi chỉ trình bày ngắn gọn các kỹ thuật khai thác dữ liệu và minh họa với ngôn ngữ lập trình **Python** [13]. Việc lựa chọn ngôn ngữ **Python** trong sách bởi lý do sau: **Python** tương thích với giấy phép phần mềm miễn phí nguồn mở, rất dễ học, phổ biến nhất và có thể phát triển nhanh nhất các ứng dụng khai thác dữ liệu trong thời gian ngắn. **Python** cung cấp nhiều công cụ, thư viện, tích hợp sẵn dùng và khả năng lập trình, hỗ trợ kiểu dữ liệu phong phú, các hàm thống kê, giải thuật học tự động và các giao diện truy vấn dữ liệu, hiển thị dữ liệu.

Chúng tôi tập trung trình bày những giải thuật trong 10 giải thuật quan trọng nhất của khai thác dữ liệu [24] bao gồm:

- *k* láng giềng [10],
- Bayes thơ ngây [12],
- Cây quyết định [2, 21],
- Phương pháp tập hợp mô hình như bagging [3], boosting [11],
- Rừng ngẫu nhiên [3],
- Máy học véctơ hỗ trợ [22]
- Giải thuật gom cụm *k*Means [19],
- Luật kết hợp có tên là Apriori [1].

Ngoài ra, chúng tôi cũng trình bày phương pháp giảm chiều dữ liệu thường dùng là phân tích thành phần chính [20]. Chúng tôi cũng cung cấp thêm các phương pháp hiển thị dữ liệu [5, 8, 16, 17] được sử dụng trong quá trình khám phá tri thức để giúp người dùng rút trích các tri thức một cách trực quan.