TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU

REVIEWS OF DATA MINING

Nguyễn Đức Cường
ndcuong@dit.hcmut.edu.vn

Khoa Công Nghệ Thông Tin, Đại học Bách khoa, Tp. Hồ Chí Minh, Việt nam

---

BẢN TÓM TẮT

This paper introduces fundamentals of Data Mining, including recommended definitions and applied process. Common tasks and applications in Data Mining are also mentioned. Finally, the paper discusses research trends in our faculty.

ABSTRACT

Bài báo giới thiệu những nét cơ bản của Khai phá Dữ liệu, bao gồm các định nghĩa đã được đề nghị và quá trình áp dụng. Các bài toán và các ứng dụng thông dụng trong Khai phá Dữ liệu cũng được đề cập đến. Cuối cùng, bài báo trình bày các hướng nghiên cứu đang được quan tâm phát triển tại khoa chúng tôi.

1. GIỚI THIỆU

Trong thời đại ngày nay, với sự phát triển vượt bật của công nghệ thông tin, các hệ thống thông tin có thể lưu trữ một khối lượng lớn dữ liệu về hoạt động hàng ngày của chúng. Từ khối dữ liệu này, các kỹ thuật trong Khai Phá Dữ Liệu (KPDL) và Máy Học (MH) có thể dùng để trích xuất những thông tin hữu ích mà chúng ta chưa biết. Các tri thức vừa học được có thể vận dụng để cải thiện hiệu quả hoạt động của hệ thống thông tin ban đầu.

Giáo sư Tom Mitchell [15] đã đưa ra định nghĩa của KPDL như sau: "KPDL là việc sử dụng dữ liệu lịch sử để khám phá những qui tắc và cải thiện những quyết định trong tương lai." Với một cách tiếp cận ứng dụng hơn, Tiến sĩ Fayyad [5] đã phát biểu: "KPDL, thường được xem là việc khám phá tri thức trong các cơ sở dữ liệu, là một quá trình trích xuất những thông tin ẩn, trước đây chưa biết và có khả năng hữu ích, dưới dạng các qui luật, ràng buộc, qui tắc trong cơ sở dữ liệu." Nói tóm lại, KPDL là một quá trình học tri thức mới từ những dữ liệu đã thu thập được [7, 8, 12].

Nội dung của bài báo này được tổ chức như sau. Phần 2 trình bày về quá trình KPDL. Các bài toán thông dụng trong KPDL được trình bày trong phần 3. Các môi trường KPDL được giới thiệu trong phần 4. Phần 5 bàn về sự khác biệt và tương đồng giữa KPDL và MH. Các ứng dụng của KPDL được nói trong phần 6. Phần 7 nêu kết luận và những hướng nghiên cứu đang được quan tâm ở khoa chúng tôi.

2. QUÁ TRÌNH KPDL

Một quá trình KPDL bao gồm năm giai đoạn chính sau [3]:

(1) Tìm hiểu nghiệp vụ và dữ liệu
(2) Chuẩn bị dữ liệu
(3) Mô hình hóa dữ liệu
(4) Hậu xử lý và đánh giá mô hình
(5) Triển khai tri thức

Quá trình này có thể được lặp lại nhiều lần một hay nhiều giai đoạn dựa trên phản hồi từ kết quả của các giai đoạn sau. Tham gia chính trong quá trình KPDL là các nhà tư vấn (NTV) và phát triển chuyên nghiệp trong lĩnh vực KPDL.

Trong giai đoạn đầu tiên, **Tìm hiểu nghiệp vụ và dữ liệu**, NTV nghiên cứu kiến thức về lĩnh vực sẽ áp dụng, bao gồm các tri thức cấu trúc về hệ thống và tri thức, các nguồn dữ liệu hiện hữu, ý nghĩa, vai trò và tầm quan trọng của các thực thể dữ liệu. Việc nghiên cứu này được thực hiện qua việc tiếp xúc giữa NTV và người dùng. Khác với phương pháp giải quyết vấn đề truyền thống khi bài toán được xác định chính xác ở bước đầu tiên, NTV tìm hiểu các yêu cầu sơ khởi của người dùng và đề nghị các bài toán tiềm năng có thể giải quyết với nguồn dữ liệu hiện hữu. Tập các bài toán tiềm năng được tinh chỉnh và làm hẹp lại trong các giai đoạn sau. Các nguồn và đặc tả dữ liệu có liên quan đến tập các bài toán tiềm năng cũng được xác định.

Giai đoạn **Chuẩn bị dữ liệu** sử dụng các kỹ thuật tiền xử lý để biến đổi và cải thiện chất lượng dữ liệu để thích hợp với những yêu cầu của các giải thuật học. Phần lớn các giải thuật KPDL hiện nay chỉ làm việc trên một tập dữ liệu đơn và phẳng, do đó dữ liệu phải được trích xuất và biến đổi từ các dạng cơ sơ dữ liệu phân bố, quan hệ hay hướng đối tượng sang dạng cơ sở dữ liệu quan hệ đơn giản với một bảng dữ liệu. Các giải thuật tiền xử lý tiêu biểu bao gồm:

(a) Xử lý dữ liệu bị thiếu/mất: các dữ liệu bị thiếu sẽ được thay thế bởi các giá trị thích hợp.
(b) Khử sự trùng lắp: các đối tượng dữ liệu trùng lắp sẽ bị loại bỏ đi. Kỹ thuật này không được sử dụng cho các tác vụ có quan tâm đến phân bố dữ liệu.
(c) Giảm nhiễu: nhiễu và các đối tượng tách rời (outlier) khỏi phân bố chung sẽ bị loại đi khỏi dữ liệu.
(d) Chuẩn hóa: miền giá trị của dữ liệu sẽ được chuẩn hóa.
(e) Rời rạc hóa: các dữ liệu số sẽ được biến đổi ra các giá trị rời rạc.
(f) Rút trích và xây dựng đặc trưng mới từ các thuộc tính đã có.
(g) Giảm chiều: các thuộc tính chứa ít thông tin sẽ được loại bỏ bớt.

Các bài toán được giải quyết trong giai đoạn **Mô hình hóa dữ liệu**. Các giải thuật học sử dụng các dữ liệu đã được tiền xử lý trong giai đoạn hai để tìm kiếm các qui tắc ẩn và chưa biết. Công việc quan trọng nhất trong giai đoạn này là lựa chọn kỹ thuật phù hợp để giải quyết các vấn đề đặt ra. Các bài toán được phân loại vào một trong những nhóm bài toán chính trong KPDL dựa trên đặc tả của chúng. Các bài toán chính trong KPDL sẽ được trình bày chi tiết trong phần 3 của bài báo.

Các mô hình kết quả của giai đoạn ba sẽ được hậu xử lý và đánh giá trong giai đoạn 4. Dựa trên các đánh giá của người dùng sau khi kiểm tra trên các tập thử, các mô hình sẽ được tinh chỉnh và kết hợp lại nếu cần. Chỉ các mô hình đạt được mức yêu cầu cơ bản của người dùng mới đưa ra triển khai trong thực tế. Trong giai đoạn này, các kết quả được biến đổi từ dạng học thuật sang dạng phù hợp với nghiệp vụ và dễ hiểu hơn cho người dùng.

Trong giai đoạn cuối, **Triển khai tri thức**, các mô hình được đưa vào những hệ thống thông tin thực tế dưới dạng các môđun hỗ trợ việc đưa ra quyết định.

Mối quan hệ chặt chẽ giữa các giai đoạn trong quá trình KPDL là rất quan trọng cho việc nghiên cứu trong KPDL. Một giải thuật trong KPDL không thể được phát triển độc lập, không quan tâm đến bối cảnh áp dụng mà thường được xây dựng để giải quyết một mục tiêu cụ thể. Do đó, sự hiểu biết bối cảnh vận dụng là rất cần thiết. Thêm vào đó, các kỹ thuật được sử dụng trong các giai đoạn trước có thể ảnh hưởng đến hiệu quả của các giải thuật sử dụng trong các giai đoạn tiếp theo.

3. CÁC BÀI TOÁN THÔNG DỤNG TRONG KPDL

Trong KPDL, các bài toán có thể phân thành bốn loại chính [18].

Bài toán thông dụng nhất trong KPDL là **Phân lớp** (Classification). Với một tập các dữ liệu huấn luyện cho trước và sự huấn luyện của con người, các giải thuật phân loại sẽ học ra bộ phân loại (classifier) dùng để phân các dữ liệu mới vào một trong những *lớp* (còn gọi là *loại*) đã được xác định trước. Nhận dạng cũng là một bài toán thuộc kiểu Phân loại.

Với mô hình học tương tự như bài toán Phân loại, lớp bài toán **Dự đoán** (Prediction) sẽ học ra các bộ dự đoán. Khi có dữ liệu mới đến, bộ dự đoán sẽ dựa trên thông tin đang có để đưa ra một giá trị số học cho hàm cần dự đoán. Bài toán

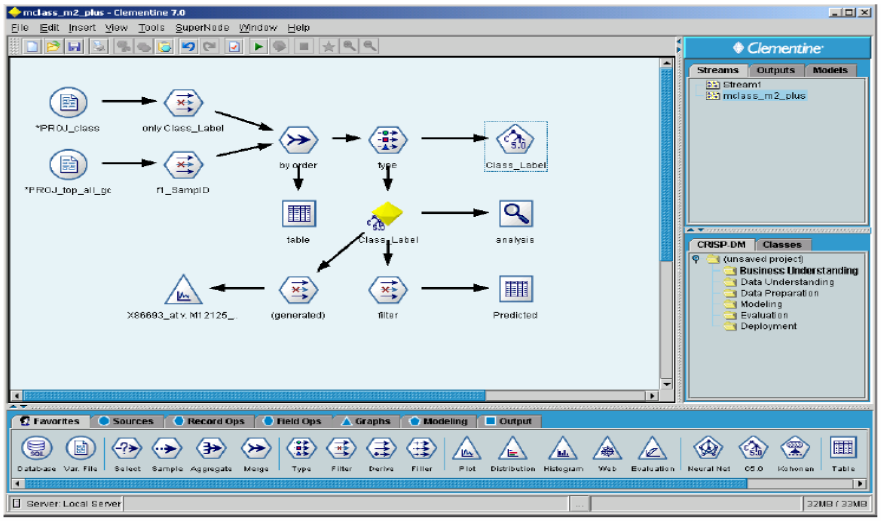
tiêu biểu trong nhóm này là dự đoán giá sản phẩm để lập kế hoạch trong kinh doanh.

Các giải thuật **Tìm luật liên kết** (Association Rule) tìm kiếm các mối liên kết giữa các phần tử dữ liệu, ví dụ như nhóm các món hàng thường được mua kèm với nhau trong siêu thị.

Các kỹ thuật **Phân cụm** (Clustering) sẽ nhóm các đối tượng dữ liệu có tính chất giống nhau vào cùng một nhóm. Có nhiều cách tiếp cận với những mục tiêu khác nhau trong phân loại. Các tài liệu [1, 4, 6, 7, 11] giới thiệu khá đầy đủ và chi tiết về các cách tiếp cận trong Phân cụm. Các kỹ thuật trong bài toán này thường được vận dụng trong vấn đề phân hoạch dữ liệu tiếp thị hay khảo sát sơ bộ các dữ liệu.

4. CÁC MÔI TRƯỜNG KPDL

Do các đặc tính được nêu ra trong phần 2, các công cụ KPDL thường được xây dựng theo dạng môi phát triển, dễ thử nghiệm và thay đổi các tác vụ KPDL. Hình 1 giới thiệu giao diện trực quan của một quá trình KPDL trong môi trường Clementine [10].



Hình 1 – Giao diện trực quan của môi trường KPDL Clementine

Trong các môi trường này, một quá trình KPDL được mô tả như một dòng các tác vụ nối tiếp, bắt đầu bằng việc lấy dữ liệu thực từ nguồn dữ liệu lịch sử, thao tác biến đổi dữ liệu sang dạng thích hợp, học và sinh ra mô hình mới. Mô hình này sau đó được thử nghiệm trên dữ liệu thực để đưa ra các đánh giá. Nếu mô hình được đánh giá chưa thỏa mãn các yêu cầu đề ra, các tác vụ trong quá trình được tinh chỉnh rồi thực hiện lại. Qui trình này được lặp lại cho đến khi nào mô hình sinh ra được đánh giá có hiệu quả tốt. Mô hình sinh ra cuối cùng sẽ được triển khai sử dụng trong thực tế. Các môi trường như vậy rất phù hợp cho quá trình KPDL vì tính chất thử nghiệm và cần thay đổi nhiều của nó.

Việc sử dụng các môi trường thử nghiệm đã thúc đẩy nhanh việc áp dụng KPDL. Thay vì phải bỏ nhiều công sức và thời gian vào việc xây dựng các chương trình hoàn chỉnh và hiện thực các giải thuật, khi dữ liệu sẵn sàng cho việc sử dụng, người vận dụng KPDL chỉ cần phải tìm hiểu các kiến thức cần thiết, khảo sát tính chất dữ liệu, vận dụng các kỹ thuật đã được hiện thực sẵn trên dữ liệu, đánh giá các kết quả tạm thời và vận dụng kết quả cuối cùng. Với phương thức hiện đại như vậy, việc áp dụng KPDL trở nên rất dễ dàng và tiện lợi.

Weka [18] là môi trường thử nghiệm KPDL do các nhà khoa học thuộc trường Đại học Waitako, NZ, khởi xướng và được sự đóng góp của rất nhiều nhà nghiên cứu trên thế giới. Weka là phần mềm mã nguồn mở, cung cấp công cụ trực quan và sinh động cho sinh viên và người ngoài ngành CNTT tìm hiểu về KPDL. Weka còn cho phép các giải thuật học mới phát triển có thể tích hợp vào môi trường của nó.

5. SỰ TƯƠNG ĐỒNG VÀ KHÁC BIỆT GIỮA KPDL VÀ MH

Với cùng mục đích là "học tập từ dữ liệu", các giải thuật trong MH đóng một vai trò nòng cốt trong KPDL. Tuy nhiên, các giải thuật này cần được phát triển để thích hợp với các yêu cầu và thách thức mới của KPDL. Thách thức đầu tiên là mức độ nhiễu cao trong dữ liệu của KPDL. Tiêu chuẩn mạnh mẽ của giải thuật đối với nhiễu trở nên quan trọng hơn trong khi các tiêu chuẩn khác phần nào có thể giảm bớt. Thách thức thứ hai là kích thước lớn của các tập dữ liệu cần xử lý. Các tập dữ liệu trong KPDL thường có kích thước cực kỳ lớn. Khi so sánh các tập dữ liệu chuẩn trong các kho dữ liệu về KPDL [9] và MH [2], các tập dữ liệu trong KPDL thường có số đặc tính lớn hơn 10 lần và số đối tượng lớn hơn 100 lần. Trong thực tế, kích thước của các tập dữ liệu trong KPDL thường ở mức tera-byte (hàng ngàn giga-byte). Với kích thước như thế, thời gian xử lý thường cực kỳ dài. Thêm vào đó, các giải thuật học truyền thống thường yêu cầu tập dữ liệu được tải toàn bộ lên trên bộ nhớ để xử lý. Mặc dù kích thước bộ nhớ trong của máy tính đã gia tăng đáng kể trong thời gian gần đây, việc gia tăng này cũng không thể đáp ứng kịp với việc tăng kích thước dữ liệu. Vì vậy, việc vận dụng các kỹ thuật xác suất, lấy mẫu, đệm, song song và tăng dần vào các giải thuật để tạo ra các phiên bản phù hợp với yêu cầu của KPDL trở nên ngày càng quan trọng.

Các kỹ thuật trong KPDL là hướng tác vụ và hướng dữ liệu. Thay vì tập trung vào xử lý tri thức dạng kí hiệu và khái niệm như trong MH, mọi phát triển trong KPDL thì kết chặt vào các ứng dụng thực tế và đặc tính dữ liệu cụ thể trong các ứng dụng đó. Ví dụ, Luật kết hợp (Association Rules) là kỹ thuật KPDL nhằm tìm kiếm những mối liên kết giữa các món hàng trong các hóa đơn ở siêu thị. Giải thuật học trong kỹ thuật này được phát triển dựa trên đặc tính về dữ liệu rất đặc thù là ở dạng nhị phân và rất thưa.

6. CÁC ỨNG DỤNG CỦA KPDL

KPDL được vận dụng trong nhiều lĩnh vực khác nhau nhằm khai thác nguồn dữ liệu phong phú được lưu trữ trong các hệ thống thông tin. Tùy theo bản chất của từng lĩnh vực, việc vận dụng KPDL có những cách tiếp cận khác nhau.

KPDL cũng được vận dụng hiệu quả để giải quyết các bài toán phức tạp trong các ngành đòi hỏi kỹ thuật cao [18], như tìm kiếm mỏ dầu từ ảnh viễn thám, xác định các vùng gãy trong ảnh địa chất để dự đoán thiên tai, cảnh báo hỏng hóc trong các hệ thống sản xuất,… Các bài toán này đã được giải quyết từ khá lâu bằng các kỹ thuật nhận dạng hay xác suất nhưng được giải quyết với yêu cầu cao hơn bởi các kỹ thuật của KPDL.

Phân nhóm và dự đoán là những công cụ rất cần thiết cho việc qui hoạch và phát triển các hệ thống quản lý và sản xuất trong thực tế [13, 16, 17]. Các kỹ thuật KPDL đã được áp dụng thành công trong việc dự đoán tải sử dụng điện năng cho các công ty cung cấp điện, lưu lượng viễn thông cho các công ty điện thoại, mức độ tiêu thụ sản phẩm cho các nhà sản xuất, giá trị của sản phẩm trên thị trường cho các công ty tài chính hay phân nhóm các khách hàng tiềm năng,…

Ngoài ra, KPDL còn được áp dụng cho các vấn đề xã hội như phát hiện tội phạm hay tăng cường an ninh xã hội [14]. Việc vận dụng thành công đã mang lại những hiệu quả thiết thực cho các hoạt động diễn ra hàng ngày trong đời sống.

7. KẾT LUẬN

KPDL là sự vận dụng học thuật vào các vấn đề thiết thực. Để giải quyết thành công các bài toán KPDL, cần có sự phối hợp và nỗ lực vượt bậc của các chuyên gia KPDL và người sử dụng. Nhà chuyên gia cần nắm vững các kỹ thuật, thấu hiểu các yêu cầu rất thực tế, vận dụng kỹ thuật để giải quyết các bài toán và giải thích kết quả bằng ngôn ngữ thực tế cho người sử dụng. Người sử dụng cần nhận ra những bài toán thiết thực, nắm bắt các kết quả đạt được và vận dụng chúng một cách hiệu quả trong thực tế.

Việc nghiên cứu áp dụng KPDL ở Khoa CNTT đang ở những bước đầu tiên trong việc xây dựng đội ngũ và trang bị những kiến thức và kỹ thuật cần thiết, sẵn sàng đón nhận và vận dụng KPDL vào trong các bài toán thực tế khi nguồn dữ liệu trở nên hiện hữu. Trong giai đoạn này, nhóm đã và đang vận dụng KPDL vào các bài toán tiêu chuẩn được công bố trên Internet,

cũng như các dữ liệu về kết quả học tập của sinh viên Đại học Bách Khoa nhằm nâng cao hiệu quả giảng dạy và học tập, bước đầu đã có những kết quả đáng khích lệ.

Nghiên cứu nhằm xây dựng và cải thiện các kỹ thuật trong KPDL là một lĩnh vực hứa hẹn và phù hợp với điều kiện nghiên cứu ở Việt nam. KPDL là một ngành khá non trẻ, các kỹ thuật của ngành còn chưa có khả năng giải quyết với hiệu quả tốt nhất các bài toán thực tế. Việc nghiên cứu cải thiện các giải thuật nhằm đưa ra các kỹ thuật mới ngang tầm với nền khoa học thế giới là một khả năng có thể thực hiện trong môi trường làm việc còn thiếu thốn ở Việt nam. Một số hướng nghiên cứu về lý thuyết trong KPDL đang được nghiên cứu ở Khoa CNTT-ĐHBK TPHCM:

- Áp dụng các chiến lược tăng dần để cải thiện hiệu quả các giải thuật
- Phát triển các phiên bản mới của các giải thuật có khả năng giải quyết các tập dữ liệu lớn bằng kỹ thuật sử dụng bộ đệm
- Song song và phân bố các giải thuật trong KPDL để tận dụng khả năng tính toán mạnh của tính toán lưới.

TÀI LIỆU THAM KHẢO

1. P. Berkhin: Survey of Clustering Data Mining Techniques. Research paper. Accrue Software, Inc, http://www.accrue.com, (2001).
2. C. Blake, E. Keogh and C. J. Merz: UCI Repository of Machine Learning Databases, Department of Information and Computer Science, University of California Irvine, CA, USA, (1998).
3. P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer and R. Wirth, CRISP-DM 1.0 Process and User Guide, http://www.crisp-dm.org, (2000).
4. R. O. Duda, P. E. Hart and D. G. Stork: Pattern Classification, Second Edition (2001), John Wiley & Sons, Inc, pp. 517-599.
5. U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy: Advances in Knowledge Discovery and Data Mining, AAAI Press, Menlo Park, CA, (1996).
6. J. Grabmeier, and A. Rudolph: Techniques of Clustering Algorithms in Data Mining, Data Mining and Knowledge Discovery, 6 (2002), Kluwer Academic Publishers, Netherlands, pp. 303-360.
7. J. Han and M. Kamber: Data Mining: Concepts and Techniques, Morgan Kaufmann, San Francisco, CA, (2000).
8. D. Hand, H. Mannila and P. Smyth: Principles of Data Mining, The MIT Press, London, England, (2001).
9. S. Hettich and S. D. Bay: The UCI KDD Archive [http://kdd.ics.uci.edu], Irvine, CA. University of California, Department of Information and Computer Science, (1999).
10. ISL: Integral Solutions Ltd., SPSS Clementine Data Mining System, User Guide Version 5 (1998), Basingstoke, Hampshire, UK.
11. A. K. Jain, M. N. Murty and P. J. Flynn: Data Clustering: A Review. ACM Computing Surveys, Vol 31 (1999), No. 3, pp. 264-323.
12. M. Kantardzic: Data Mining: Concepts, Models, Method, and Algorithms, John Wiley & Sons, New York, NY, (2003).
13. R. Mattison: Data Warehousing and Data Mining for Telecommunications, Norwood, MA, (1997).
14 J. Mena: Investigative Data Mining for Security and Criminal Detection, Butterworth Heinemann, New York, NY, (2003).
15. T. Mitchell, Machine Learning and Data Mining, Communications of the ACM, Vol. 42 (1999), No. 11, pp. 30--36.
16. D. Pyle: Business Modeling and Data Mining, Morgan Kaufmann Publishers, San Francisco, CA, (2003).
17. O. P. Rud: Data Mining Cookbook – Modeling Data for Marketing, Risk and Customer Relationship Management, John Wiley & Sons, New York, NY, (2001).
18. I. H. Witten and E. Frank: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann Publishers, New York, NY, (2000).