KỸ THUẬT PHÁT HIỆN TRI THỨC VÀ KHAI PHÁ DỮ LIỆU, ỨNG DỤNG TRONG BÀI TOÁN DỰ BÁO TỪ THÔNG TIN KINH TẾ - XÃ HỘI

ThS. BÙI DƯƠNG HƯNG - ThS. NGUYỄN THỊ THU TRANG*

Sự phát triển của công nghệ thông tin và việc ứng dụng công nghệ thông tin vào các lĩnh vực của đời sống, kinh tế xã hội trong nhiều năm qua cũng đồng nghĩa với lượng dữ liệu được các cơ quan thu thập và lưu trữ ngày một nhiều lên. Các dữ liệu này được dự đoán ẩn chứa những giá trị nhất định trong tương lai. Tuy nhiên, theo thống kê, chỉ có một lượng nhỏ (khoảng từ 5% đến 10%) dữ liệu được phân tích, số còn lại chưa sử dụng nhưng vẫn tiếp tục được thu thập vì e ngại rằng sau này cần dùng đến. Mặt khác, trong môi trường cạnh tranh, cần có nhiều thông tin với tốc độ nhanh để trợ giúp việc ra quyết định, ngày càng có nhiều câu hỏi mang tính chất định tính cần phải trả lời dựa trên một khối lượng dữ liệu khổng lồ đã có. Với những lý do như vậy, các phương pháp quản trị và khai thác cơ sở dữ liệu truyền thống ngày càng không đáp ứng được thực tế. Trước tình hình đó, một khuynh hướng kỹ thuật mới ra đời, đó là Kỹ thuật phát hiện tri thức và khai phá dữ liệu (KDD - Knowledge Discovery and Data Mining).

Kỹ thuật phát hiện tri thức và khai phá dữ liệu đã và đang được nghiên cứu, ứng dụng trong nhiều lĩnh vực khác nhau ở các nước trên thế giới, tại Việt Nam kỹ thuật này cũng đang được nghiên cứu và dần đưa vào ứng dụng. Bài viết trình bày một cách tổng quan về Kỹ thuật phát hiện tri thức và khai phá dữ liệu. Trên cơ sở đó, đưa ra một bài toán dự báo về dân số thế giới và giải quyết bài toán bằng phương pháp hồi qui tuyến tính nhằm cung cấp một cách nhìn khái quát về kỹ thuật mới này cũng như mối tương quan với phương pháp thống kê truyền thống. Nội dung bài báo có tính tương tác, trợ giúp nguồn kiến thức cơ bản về phương pháp dự báo cho sinh viên khối các ngành kinh tế đang được đào tạo tại trường Đại học Công đoàn.

Kỹ thuật phát hiện tri thức (KDD - Knowledge Discovery)

Phát hiện tri thức là gì

Thông thường chúng ta coi dữ liệu như một dãy các bit, hoặc các số và các ký hiệu, hoặc các "đối tượng" với một ý nghĩa nào đó khi được gửi cho một chương trình dưới một dạng nhất định. Chúng ta sử dụng các bit để đo lường các thông tin và xem nó như là các dữ liệu đã được lọc bỏ các dư thừa, được rút gọn tới mức tối thiểu để đặc trưng một cách cơ bản cho dữ liệu. Chúng ta có thể xem tri thức như là các thông tin tích hợp, bao gồm các sự kiện và các mối quan hệ giữa chúng. Các mối quan hệ này có thể được hiểu ra, có thể được phát hiện, hoặc có thể được học. Nói cách khác, tri thức có thể được coi là dữ liệu có độ trừu tượng và tổ chức cao. Phát hiện tri thức trong các cơ sở dữ liệu là một qui trình nhận biết các mẫu hoặc các mô hình trong dữ liệu với các tính năng: hợp thức, mới, khả ích, và có thể hiểu được.

Qui trình phát hiện tri thức

Qui trình phát hiện tri thức được mô tả tóm tắt như sau:

Hình 1. Quy trình phát hiện tri thức

Bước thứ nhất, tìm hiểu lĩnh vực ứng dụng và hình thành bài toán. Bước này quyết định cho việc rút ra được các tri thức hữu ích và cho phép chọn các phương pháp khai phá dữ liệu thích hợp với mục đích ứng dụng và bản chất của dữ liệu.

* Trường Đại học Công đoàn

Bước thứ hai, thu thập và xử lý thô, còn được gọi là tiền xử lý dữ liệu nhằm loại bỏ nhiễu, xử lý việc thiếu dữ liệu, biến đổi dữ liệu và rút gọn dữ liệu nếu cần thiết, bước này thường chiếm nhiều thời gian nhất trong toàn bộ qui trình phát hiện tri thức.

Bước thứ ba, khai phá dữ liệu, hay nói cách khác là trích ra các mẫu hoặc/và các mô hình ẩn dưới các dữ liệu.

Bước thứ tư, hiểu tri thức đã tìm được, đặc biệt là làm sáng tỏ các mô tả và dự đoán. Các bước trên có thể lặp đi lặp lại một số lần, kết quả thu được có thể được lấy trung bình trên tất cả các lần thực hiện.

Khai phá dữ liệu (Data Mining)

Khai phá dữ liệu là gì

Khai phá dữ liệu là một bước trong qui trình phát hiện tri thức gồm có các thuật toán khai thác dữ liệu chuyên dùng dưới một số qui định về hiệu quả tính toán chấp nhận được để tìm ra các mẫu hoặc các mô hình trong dữ liệu. Nói một cách khác, mục đích của phát hiện tri thức và khai phá dữ liệu chính là tìm ra các mẫu và/hoặc các mô hình đang tồn tại trong các cơ sở dữ liệu nhưng vẫn còn bị che khuất bởi hàng núi dữ liệu.

Còn các nhà thống kê thì xem Khai phá dữ liệu như là một qui trình phân tích được thiết kế để thăm dò một lượng cực lớn các dữ liệu nhằm phát hiện ra các mẫu thích hợp và/hoặc các mối quan hệ mang tính hệ thống giữa các biến, sau đó sẽ hợp thức hoá các kết quả tìm được bằng cách áp dụng các mẫu đã phát hiện được cho các tập con mới của dữ liệu. Qui trình này bao gồm ba giai đoạn cơ bản: thăm dò, xây dựng mô hình hoặc định nghĩa mẫu, hợp thức/kiểm chứng.

Các phương pháp khai phá dữ liệu

Với hai đích chính của khai phá dữ liệu là Dự đoán (Prediction) và Mô tả (Description), người ta thường sử dụng các phương pháp sau cho khai phá dữ liệu:

- Phân loại (Classification)
- Hồi qui (Regression)
- Phân nhóm (Clustering)
- Tổng hợp (Summarization)
- Mô hình ràng buộc (Dependency modeling)
- Dò tìm biến đổi và độ lệch (Change and Deviation Dectection)
- Biểu diễn mô hình (Model Representation)
- Kiểm định mô hình (Model Evaluation)
- Phương pháp tìm kiếm (Search Method)

Ứng dụng của Phát hiện tri thức và khai phá dữ liệu

Phát hiện tri thức và khai phá dữ liệu liên quan đến nhiều ngành, nhiều lĩnh vực: thống kê, trí tuệ nhân tạo, cơ sở dữ liệu, thuật toán học, tính toán song song và tốc độ cao, thu thập tri thức cho các hệ chuyên gia, quan sát dữ liệu... Đặc biệt phát hiện tri thức và khai phá dữ liệu rất gần gũi với lĩnh vực thống kê, sử dụng các phương pháp thống kê để mô hình dữ liệu và phát hiện các mẫu, luật... Ngân hàng dữ liệu (Data Warehousing) và các công cụ phân tích trực tuyến (OLAP) cũng liên quan rất chặt chẽ với phát hiện tri thức và khai phá dữ liệu.

Các ứng dụng của Phát hiện tri thức và khai phá dữ liệu

- Thông tin thương mại:
§ Phân tích dữ liệu marketing, khách hàng
§ Phân tích đầu tư
§ Phê duyệt cho vay vốn
§ Phát hiện gian lận...
- Thông tin kỹ thuật:
§ Điều khiển và lập lịch trình
§ Quản trị mạng
§ Phân tích các kết quả thí nghiệm...
- Thông tin khoa học.
- Thông tin cá nhân...

Các thách thức với phát hiện tri thức và khai phá dữ liệu

- Các cơ sở dữ liệu lớn
- Mô hình dữ liệu nhiều chiều
- Thay đổi dữ liệu và tri thức có thể làm cho các mẫu đã phát hiện không còn phù hợp.
- Dữ liệu bị thiếu hoặc nhiễu
- Quan hệ giữa các trường phức tạp
- Giao tiếp với người sử dụng và kết hợp với các tri thức đã có.
- Tích hợp với các hệ thống khác...

Ứng dụng trong bài toán dự báo dân số

Từ lĩnh vực thống kê, có bảng dữ liệu về dân số thế giới (bảng 1).

Bảng 1. *Dân số thế giới tính tại thời điểm giữa năm*

Năm	Dân số thế giới (triệu người)	Năm	Dân số thế giới (triệu người)	Năm	Dân số thế giới (triệu người)	Năm	Dân số thế giới (triệu người)
1980	4,449	1990	5,321	2000	6,128	2010	6,916
1981	4,528	1991	5,409	2001	6,204	2011	6,998
1982	4,609	1992	5,495	2002	6,281	2012	7,080
1983	4,692	1993	5,579	2003	6,358	2013	7,162
1984	4,776	1994	5,661	2004	6,436	2014	7,244
1985	4,864	1995	5,742	2005	6,514	2015	7,325
1986	4,953	1996	5,821	2006	6,593		
1987	5,045	1997	5,899	2007	6,673		
1988	5,138	1998	5,975	2008	6,754		
1989	5,230	1999	6,051	2009	6,835		

(Nguồn GEOHIVE, Population of the entire world)

Bảng 2. *Thông số cơ bản của mô hình hồi qui*

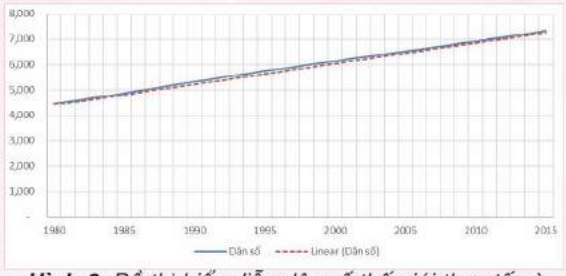
Mô hình	Hệ số xác định (R^2)	Hàm biểu diễn
Linear	0.9993	$y = 81.73x + 4397.4$
Logarit (Ln)	0.8535	$y = 919.63\ln(x) + 3464.2$
Polynomial	0.9998	$y = -0.1904x^2 + 88.775x + 4352.8$
Exponential	0.9918	$y = 4508.8e^{0.014x}$

Bảng 3. *Dự báo dân số thế giới theo các mô hình hồi quy*

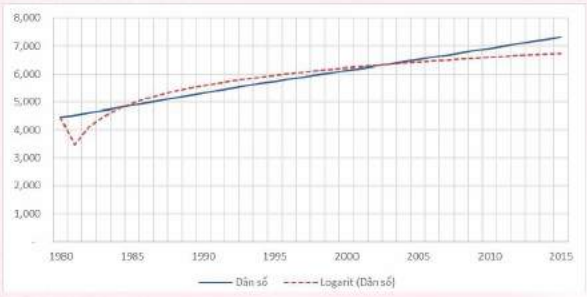
Năm	Kết quả Dự báo dân số (triệu người) theo các mô hình			
	Linear	Logarit (Ln)	Polynomial	Exponential
2016	7,340	6,760	7,302	7,464
2017	7,421	6,785	7,377	7,569
2018	7,503	6,809	7,451	7,675
2019	7,585	6,833	7,525	7,784
2020	7,667	6,857	7,599	7,893
2021	7,748	6,879	7,673	8,005
2022	7,830	6,901	7,745	8,118
2023	7,912	6,923	7,818	8,232
2024	7,994	6,944	7,890	8,348
2025	8,075	6,965	7,962	8,466
2026	8,157	6,985	8,034	8,585
2027	8,239	7,005	8,105	8,706
2028	8,320	7,024	8,175	8,829
2029	8,402	7,043	8,246	8,953
2030	8,484	7,062	8,316	9,080

Như vậy, dựa trên những số liệu thống kê dân số thế giới từ năm 1980 - 2015, sử dụng phương pháp hồi quy (Regression), bài toán dự báo về dân số thế giới đến năm 2030 đã được giải đáp. Mặc dù số lượng các dữ liệu không lớn như trong các dữ liệu kinh tế - xã hội khác, nhưng bài toán này cũng cho ta thấy các mô hình phân tích khác nhau và các kết quả khác nhau khi khai phá những dữ liệu đó. Để đơn giản, ta không đề cập đến bước thu thập và tiền xử lý dữ liệu, các dữ liệu tại bảng dưới được coi là hoàn thiện trong bài toán này. Mặt khác, các dữ liệu thực tế được tính vào giữa các năm do vậy các dữ liệu dân số ta tính toán cũng được hiểu ngầm định là vào giữa năm.

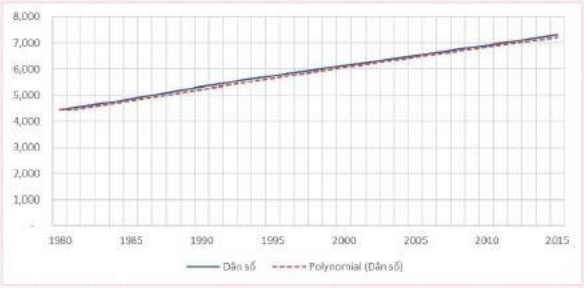
Sau khi thực hiện khai phá dữ liệu dân số bằng phương pháp hồi qui đơn với bốn mô hình khác nhau: Linear (hàm tuyến tính), Logarit (hàm lôgarit tự nhiên), Polynomial (hàm đa thức - trong ví dụ này ta chọn đa thức bậc 2), Exponential (hàm mũ), ta đã xác định được kết quả (Xem bảng 2, 3, hình 2, 3, 4, 5).



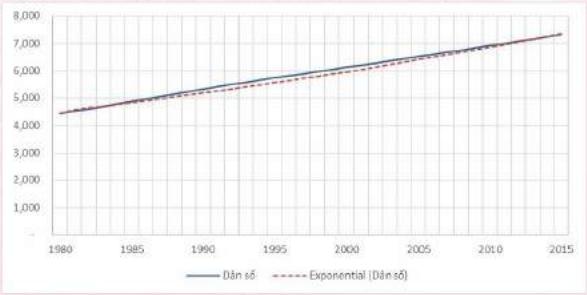
Hình 2. *Đồ thị biểu diễn dân số thế giới thực tế và lý thuyết theo năm với mô hình Linear*



Hình 3. *Đồ thị biểu diễn dân số thế giới thực tế và lý thuyết theo năm với mô hình Logarit (Ln)*



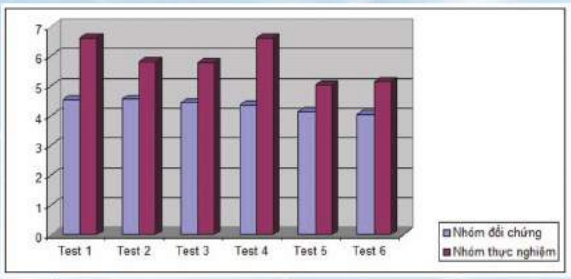
Hình 4. *Đồ thị biểu diễn dân số thế giới thực tế và lý thuyết theo năm với mô hình Polynomial*



Hình 5. *Đồ thị biểu diễn dân số thế giới thực tế và lý thuyết theo năm với mô hình Exponential*

Trong các kết quả trên, ta thấy mô hình đa thức bậc 2 - Polynomial có độ tương quan cao hơn các mô

(Xem tiếp trang 82)



Biểu đồ 2. Nhịp độ tăng trưởng trình độ thể lực của nữ sinh viên nhóm đối chứng và thực nghiệm sau 01 năm học thực nghiệm

độ thể lực của nữ sinh viên năm thứ nhất Trường Đại học Công đoàn.

Để thấy được sự tăng trưởng, đề tài tiến hành so sánh nhịp tăng trưởng trình độ thể lực của nữ sinh viên nhóm đối chứng và nhóm thực nghiệm sau 01 năm học thực nghiệm. Kết quả được trình bày ở bảng 5 và biểu đồ 2.

Qua bảng 5 và biểu đồ 2, cho thấy: Sau 01 năm học thực nghiệm, cả 2 nhóm đều có nhịp độ tăng trưởng trình độ thể lực ở mức tốt. Tuy nhiên nhóm thực nghiệm có nhịp độ tăng trưởng cao hơn so với nhóm đối chứng từ 0.90 - 2.27%.

Như vậy, sau 01 năm thực nghiệm áp dụng các bài tập phát triển thể lực và tiến trình đã xây dựng của đề tài, trình độ thể lực của nữ sinh viên nhóm thực nghiệm tốt hơn hẳn so với nhóm đối chứng. Như vậy, việc áp dụng các bài tập phát triển thể lực đã lựa chọn và tiến trình đã xây dựng của đề tài đã phát huy hiệu quả cao trong việc phát triển trình độ thể lực cho nữ sinh viên năm thứ nhất, Trường Đại học Công đoàn.

Kết luận

Phân tích thực trạng công tác giáo dục thể chất cho nữ sinh viên năm thứ nhất Trường Đại học Công đoàn còn một số vấn đề bất cập như: Số lượng giảng viên còn hạn chế; cơ sở vật chất phục vụ công tác GDTC còn thiếu cả về số lượng và chất lượng; các bài tập được sử dụng trong phát triển thể lực cho đối tượng nghiên cứu còn ít về số lượng, chưa đa dạng về thể loại, chưa được phân nhóm hợp lý, chưa được nghiên cứu chứng minh tính hiệu quả... nên hiệu quả phát triển thể lực cho sinh viên chưa cao, tỷ lệ sinh viên chưa đạt tiêu chuẩn rèn luyện thân thể còn nhiều, đặc biệt ở các chỉ tiêu đánh giá sức bền và khả năng phối hợp vận động.

Để phát triển thể lực cho nữ sinh viên năm nhất Đại học Công đoàn, nhằm nâng cao chất lượng dạy và học giáo dục thể chất, tác giả đề nghị với Ban giám hiệu Trường và bộ môn Thể dục quân sự lựa chọn và đưa nội dung các bài tập mà tác giả nghiên cứu và đề xuất để áp dụng vào công tác giảng dạy. ❑

TÀI LIỆU THAM KHẢO

1. *Bộ Giáo dục và Đào tạo (2008), Quyết định số 53/2008/QĐ-BGDĐT, ngày 18/9/2008 Ban hành quy định về đánh giá, xếp loại thể lực học sinh sinh viên.*
2. *Dương Nghiệp Chí (1991), Đo lường thể thao, Nxb TDTT, Hà Nội.*
3. *Lưu Quang Hiệp, Phạm Thị Uyên (1995), Sinh lý học TDTT, Nxb TDTT, Hà Nội.*
4. *Lưu Quang Hiệp, Vũ Đức Thu (1989), Nghiên cứu về sự phát triển thể chất sinh viên các trường Đại học, Nxb TDTT Hà Nội.*
5. *Nguyễn Xuân Sinh (1999), Giáo trình phương pháp NCKH TDTT, Nxb TDTT, Hà Nội.*
6. *Nguyễn Toán, Phạm Danh Tốn (2000),* Lý luận và phương pháp TDTT, *Nxb TDTT, Hà Nội.*
7. *Phạm Ngọc Viễn, Phạm Xuân Thành (2010),* Tâm lý học TDTT, *Nxb TDTT, Hà Nội.*

KĨ THUẬT PHÁT HIỆN TRI THỨC...

(Tiếp theo trang 79)

hình khác, do vậy, trong trường hợp cụ thể này ta có thể sử dụng các kết quả dự báo của mô hình này. Bài báo không đi sâu phân tích tiếp việc áp dụng dữ liệu đã được dự báo vào các lĩnh vực khác nhau.

Kết luận

Qua các vấn đề đã được trình bày trong bài viết, chúng ta nhận thấy với một lượng dữ liệu thực tế nhỏ và với mục đích bài toán cụ thể nhưng ta có thể tiếp cận theo nhiều hướng khác nhau của cùng một phương pháp khai phá dữ liệu và đạt được kết quả khác nhau, điều đó càng làm sáng tỏ khả năng ứng dụng thực tế to lớn đồng thời với những thách thức đối với kỹ thuật phát hiện tri thức và khai phá dữ liệu trong các bài toán kinh tế - xã hội và trong nhiều lĩnh vực khác. ❑

Tài liệu Tham khảo

[1] Knowledge Discovery Nuggets: http://www.kdnuggets.com/
[2] Ho Tu Bao: Introduction to Knowledge Discovery and Data Mining, Institute of Information Technology.
[3] TS. Hàn Viết Thuận - Chủ biên: Giáo trình Tin học ứng dụng, NXB Thống kê, 1999
[4] Dr. Dang Quang A and Dr. Bui The Hong, Statistical data analysis, Institute of Information Technology.
[5] Do thien Anh Tuan: Tin Học ứng dụng, NXB Lao động, 2012
[6] Giáo trình Tin học ứng dụng (lưu hành nội bộ) Bộ môn Tin học, Trường ĐH Công đoàn.