**DANH SÁCH THAM GIA ĐỀ TÀI**

**THÀNH VIÊN THAM GIA ĐỀ TÀI:**

|  |  |  |
| --- | --- | --- |
| **Họ và Tên** | **Đơn vị công tác** | **Nhiệm vụ** |
| Nguyễn Thành Trung | Trung tâm Đào tạo E-Learning | Chủ nghiệm đề tài |
| Trần Lan Thu | Trung tâm Đào tạo E-Learning | Nghiên cứu viên |
| Nguyễn Danh Tú | Khoa Toán tin ứng dụng ĐHBKHN | Nghiên cứu viên |
| Bùi Minh Tâm | Trung tâm Đào tạo E-Leanring | Thư ký |

**ĐƠN VỊ PHỐI HỢP CHÍNH:**

|  |  |
| --- | --- |
| **TÊN ĐƠN VỊ** | **LĨNH VỰC PHỐI HỢP** |
| Trung tâm Phát triển Đào tạo | Tư vấn hành vi người học |
| Công ty CP ĐT & PT ĐT Edutop 64 | Tập hợp dữ liệu từ phần mềm |

Mục lục

[Danh mục hình vẽ và bảng biểu 4](#_Toc311388019)

[Danh mục từ viết tắt 5](#_Toc311388020)

[MỞ ĐẦU 6](#_Toc311388021)

[Chương 1. Tiếp cận vấn đề học viên nghỉ học trong các cơ sở đào tạo 7](#_Toc311388022)

[1. Những tác động tiêu cực của vấn đề học viên nghỉ học 7](#_Toc311388023)

[2. Nguyên nhân dẫn đến tình trạng nghỉ học ở môi trường đào tạo từ xa 9](#_Toc311388024)

[3. Các hướng giải quyết cho vấn đề nghỉ ở HOU E-learning 12](#_Toc311388025)

[Chương 2. Tóm tắt cơ sở lý thuyết và những kết quả nghiên cứu 15](#_Toc311388026)

[1. Khái niệm về khai phá dữ liệu 15](#_Toc311388027)

[2. Các phương pháp khai phá dữ liệu 15](#_Toc311388028)

[3. Quy trình khám phá tri thức (Knowledge discovery in databases – KDD) 16](#_Toc311388029)

[4. Kết quả nghiên cứu từ những trường đại học khác 17](#_Toc311388030)

[5. Phân tích vấn đề nghỉ học ở HOU – Elearning 20](#_Toc311388031)

[Chương 3. Quá trình và kết quả nghiên cứu và phát triển 22](#_Toc311388032)

[1. Thu thập dữ liệu 22](#_Toc311388033)

[2. Các mô hình dự báo 22](#_Toc311388034)

[3. Sản phẩm phần mềm và các bước xử lý 23](#_Toc311388035)

[3.1. Khái niệm tiền xử lý tập dữ liệu 23](#_Toc311388036)

[3.2. Các thao tác tiền xử lý 24](#_Toc311388037)

[3.3. Mã hóa cột dữ liệu văn bản (Encoding Categorical Columns) 25](#_Toc311388038)

[3.4. Mã hóa cột dữ liệu Date/Time (Encoding Date/Time Columns) 26](#_Toc311388039)

[3.5. Sản phẩm phần mềm 27](#_Toc311388040)

[4. Báo cáo phân tích dữ liệu nghỉ học 36](#_Toc311388041)

[KẾT LUẬN VÀ KIẾN NGHỊ 37](#_Toc311388042)

[TÀI LIỆU THAM KHẢO 40](#_Toc311388043)

[PHỤ LỤC 1: CÁC BÀI TOÁN THƯỜNG GẶP TRONG KHAI PHÁ DỮ LIỆU 48](#_Toc311388044)

[1. Bài toán phân lớp 48](#_Toc311388045)

[1.1. Khái niệm phân lớp 48](#_Toc311388046)

[1.2. Các kĩ thuật phân lớp: 48](#_Toc311388047)

[2. Tiền xử lý dữ liệu 48](#_Toc311388048)

[2.2. Làm sạch dữ liệu (Data cleaning) 50](#_Toc311388049)

[2.3. Hồi quy (Regression) 52](#_Toc311388050)

[2.4. Tích hợp dữ liệu (Data integration) 52](#_Toc311388051)

[2.5. Biến đổi dữ liệu (Data transformation) 53](#_Toc311388052)

[3. Lịch sử, nguồn gốc của mạng Nơron nhân tạo 54](#_Toc311388053)

[3.1. Vài nét về hoạt động của não con người 54](#_Toc311388054)

[3.2. Nguồn gốc, ý tưởng của mạng Nơron 56](#_Toc311388055)

[3.3. Lịch sử phát triển của mạng Nơron 57](#_Toc311388056)

[4. Khái niệm và cấu trúc của một Nơron nhân tạo 57](#_Toc311388057)

[4.1. Mô hình một Nơron nhân tạo 57](#_Toc311388058)

[4.2. Liên kết Nơron 58](#_Toc311388059)

[4.3. Hàm gộp 59](#_Toc311388060)

[4.4. Hàm kích hoạt 59](#_Toc311388061)

[5. Mô hình mạng Nơron nhân tạo 60](#_Toc311388062)

[5.1. Khái niệm mạng Nơron 60](#_Toc311388063)

[5.2. Phân loại mạng Nơron 61](#_Toc311388064)

[6. Ứng dụng của mạng Nơron 67](#_Toc311388065)

[6.1. Ưu, nhược điểm của mạng Nơron 67](#_Toc311388066)

[6.2. Các lĩnh vực ứng dụng 68](#_Toc311388067)

[7. Huấn luyện mạng Nơron 69](#_Toc311388068)

[7.1. Phương pháp huấn luyện 69](#_Toc311388069)

[7.2. Thuật toán luyện mạng 70](#_Toc311388070)

[7.3. Cách thức huấn luyện 72](#_Toc311388071)

[7.4. Vấn đề thiết kế cấu trúc mạng 72](#_Toc311388072)

[8. Mạng Perceptron và qui tắc học Perceptron 74](#_Toc311388073)

[8.1. Mô hình của mạng Perceptron 74](#_Toc311388074)

[8.2. Thuật toán luyện mạng Perceptron 75](#_Toc311388075)

[9. Mạng Perceptron nhiều lớp 76](#_Toc311388076)

[9.1. Cấu trúc của mạng Perceptron nhiều lớp 76](#_Toc311388077)

[9.2. Thuật toán lan truyền ngược 77](#_Toc311388078)

# Danh mục hình vẽ và bảng biểu

[**Hình 1.** Thống kê ảnh hưởng của nghỉ học tới trường học, học viên và xã hội. 11](#_Toc311387667)

[**Hình 2.** Nguyên nhân dropout của các chương trình online 11](#_Toc311387668)

[**Hình 3.** Luồng nghỉ học của sinh viên 12](#_Toc311387669)

[**Hình 4.** Thống kê các nguyên dẫn dẫn tới NGHỈ HỌC của học viên 14](#_Toc311387670)

[**Hình 5.** Khái niệm khai phá dữ liệu 17](#_Toc311387671)

[**Hình 6.** Mô hình khám phá tri thức 18](#_Toc311387672)

[**Hình 7.** Quy trình khai phá tri thức 18](#_Toc311387673)

[**Hình 8.** Thống kê các nguồn dữ liệu 24](#_Toc311387674)

[**Hình 9.** Giao diện chính phần mềm 30](#_Toc311387675)

[**Hình 10.** Chức năng phân tích tập dữ liệu 31](#_Toc311387676)

[**Hình 11.** Chức năng tiền xử lý dữ liệu 32](#_Toc311387677)

[**Hình 12.** Tùy chỉnh tham số mạng 33](#_Toc311387678)

[**Hình 13.** Đồ thị hàm lỗi của mạng trên các tập dữ liệu 34](#_Toc311387679)

[**Hình 14.** Thiết lập kiến trúc mạng 34](#_Toc311387680)

[**Hình 15.** Phần trăm đúng CCR trên các tập dữ liệu tham gia luyện mạng 35](#_Toc311387681)

[**Hình 16.** Phần trăn đúng dựa trên dữ liệu xác minh 36](#_Toc311387682)

[**Hình 17.** Phần trăm đúng trên tập dữ liệu Test 36](#_Toc311387683)

[**Hình 18.** Kết quả truy vấn 1 37](#_Toc311387684)

[**Hình 19.** Kết quả truy vấn 2 38](#_Toc311387685)

[**Hình 20.** Làm sạch dữ liệu 51](#_Toc311387686)

[**Hình 21.** Tích hợp dữ liệu 52](#_Toc311387687)

[**Hình 22.** Giảm bớt dữ liệu 52](#_Toc311387688)

[**Hình 23.** Mô hình Nơron sinh học 57](#_Toc311387689)

[**Hình 24.** Mô hình một Nơron nhân tạo 60](#_Toc311387690)

[**Hình 25.** Mô hình mạng Nơron một lớp 64](#_Toc311387691)

[**Hình 26.** Mô hình mạng Nơron nhiều lớp 65](#_Toc311387692)

[**Hình 27.** Mô hình mạng lan truyền tiến với một mức Nơron 66](#_Toc311387693)

[**Hình 28.** Mạng lan truyền tiến kết nối đầy đủ 67](#_Toc311387694)

[**Hình 29.** Mạng hồi quy không có Nơron ẩn và không có vòng lặp tự phản hồi 68](#_Toc311387695)

[**Hình 30.** Mạng qui hồi có các Nơron ẩn 68](#_Toc311387696)

[**Hình 31.** Mô hình của mạng Perceptron 76](#_Toc311387697)

[**Hình 32.** Mô hình mạng Perceptron nhiều lớp 79](#_Toc311387698)

# Danh mục từ viết tắt

|  |  |  |
| --- | --- | --- |
| **STT** | **Nội dung** | **Viết tắt** |
|  | Knowledge discovery in databases | KDD |
|  | School Managent System | CRM |
|  | Customer Relationalship System | CRM |
|  | Cố vấn học tập | CVHT |
|  | Tư vấn tuyển sinh | TVTS |
| 6 | Chương trình đào tạo hệ từ xa theo phương thức E-Learning do Trung tâm Đào tạo E-Learning quản lý | HOU E-Learning |
| 7 | Công nghệ thông tin và truyền thông | ICT |
| 8 | Học viên | HV |

# MỞ ĐẦU

Là một trong những trường đại học hàng đầu tại Việt Nam về đào tạo từ xa, và cũng từ nhiều năm là thành viên chính thức của hiệp hội các trường đại học mở Châu Á (AAOU), Viện Đại học Mở Hà Nội đã thực hiện nhiều nghiên cứu về công nghệ và các xu hướng phát triển đào tạo trong nước và quốc tế. Đội ngũ cán bộ, giảng viên của Viện xác định rằng E-Learning chính là giải pháp quan trọng nhất để đa dạng hóa phương thức đào tạo, thay đổi cơ bản trong cách dạy và học, mở rộng quy mô, đảm bảo chất lượng, đẩy mạnh các chương trình hợp tác quốc tế trong giai đoạn 2010-2020. Sau 03 năm hoạt động, Trung tâm Đào tạo E-Leanring Viện Đại học Mở Hà Nội đã có hơn 5000 học viên theo học.

Cùng với sự phát triển mạnh mẽ về quy mô đào tạo E-Learning, Trung tâm Đào tạo E-Learning cũng như các đơn vị đào tạo từ xa và đào tạo trực tuyến khác trên thế giới đối mặt với tình trạng bỏ học của một bộ phận không nhỏ học viên. Vấn đề này ảnh hưởng trực tiếp đến quy mô phát triển, chất lượng và hiệu quả kinh tế của chương trình đào tạo. Tìm hiểu và dự báo được khả năng nghỉ học của các nhóm học viên, phát hiện được những thành tố liên quan trực tiếp đến quá trình nghỉ học sẽ giúp các nhà quản lý đưa ra các phương pháp phù hợp để giảm bớt tình trạng này mang lại hiệu quả cho Viện, cho Trung tâm trong cả ngắn hạn và dài hạn.

Mặc dù trên thế giới đã có những nghiên cứu về vấn đề học viên nghỉ học (drop-out) trong môi trường người học là người lớn, trong giáo dục mở và từ xa. Chúng ta phải thừa nhận rằng điều kiện xã hội, điều kiện kinh tế, động cơ học tập của người học tại các nước khác nhau có sự khác biệt lớn. Đề tài **Khai phá dữ liệu: dự báo tỉ lệ nghỉ học của các nhóm học viên hệ từ xa theo phương thức E-Learning** có mục đích xây dựng những nền tảng ban đầu trong nghiên cứu vấn đề học viên nghỉ học trong môi trường E-Learning đặc thù của HOU.

Báo cáo này được nhóm đề tài xây dựng nhằm tóm tắt những nét chính trong quá trình nghiên cứu và một số kết quả đạt được ban đầu. Nhóm cũng đưa ra những kiến nghị và hướng phát triển tiếp theo của đề tài.

# Tiếp cận vấn đề học viên nghỉ học trong các cơ sở đào tạo

## Những tác động tiêu cực của vấn đề học viên nghỉ học

Vấn đề nghỉ học của học viên tại các lĩnh vực đào tạo nói chung và các trường đại học nói riêng có những tác động tiêu cực khác nhau. Những tác động này không chỉ ảnh hưởng đến cá nhân học viên, cơ sở đào tạo mà còn có ảnh hưởng trên nhiều đối tượng khác nhau trong xã hội, … Hiện nay, trên thế giới cũng đã có rất nhiều nghiên cứu về tác hại của vấn đề nghỉ học của học viên.

**Ảnh hưởng tiêu cực tới cơ sở đào tạo:**

* Thương hiệu trường học bị ảnh hưởng vì có nhiều học viên bỏ học
* Khó cân đối về tài chính do thay đổi lớn về số lượng học viên
* Thương hiệu trường học bị đánh giá không cao
* Số học viên tuyển sinh bị giảm các năm học sau

**Ảnh hưởng tiêu cực với học viên:**

* Chi phí cho quá trình học tập (tiền, thời gian,…) tăng mà không đạt được mục đích
* Tâm lý bị tiêu cực và chán nản
* Không có bằng cấp và kỹ năng nghề nghiệp

Theo các nghiên cứu của các trường đại học tại Anh và Mỹ về mặt dài hạn, đối tượng học viên cũng chịu nhiều tác động:

* Trong năm 1971 một đàn ông Mỹ nghỉ học trung bình kiếm được 35,087 đô la, đến năm 2002 kiếm được 23,903 đô la sụt giảm 31.87% . Cũng trong khoảng thời gian trên số tiền kiếm được của một người phụ nữ nghỉ học bị giảm từ 19,888 đô la xuống 17,114 đô la (Paul E. Barton. One Third of a Nation Pg. 41 Policy Information Center, Educational Testing Service 2005).

Với phạm vi rộng hơn, trên mặt “Xã hội” cũng chịu một số tác động trong ngắn hạn:

* Nạn thất nghiệp gia tăng: Tỷ lệ NGHỈ HỌC cao của những người mà chưa có việc làm gây ra nạn thất nghiệp và ảnh hưởng tới lực lượng lao động của nghành kinh tế (Civil Rights Project. Harvard University. Civil Rights in Brief DROPOUTS p1).
* Chi phí đầu tư giáo dục tăng

**Ảnh hưởng xã hội:**

* Trong khoảng thời gian từ 25 đến 30 năm một học viên NGHỈ HỌC thiệt hại cho chi phí xã hội là $500,000 cho sự giúp đỡ công cộng, y tế công cộng, và chi phí tù tội.( Indiana’s Education Round Table 2003 p 17)
* Số tù nhân là những người đã từng NGHỈ HỌC gấp 3.5 lần số tù nhân đã tốt nghiệp. (Harlow, C.W. (2003). Education and correctional populations, bureau of justice statistics special report. Washington, DC: US Department of Justice.)

Bảng sau tổng kết một số tác động chính lên trường học, học viên, và xã hội về mặt ngắn hạn và dài hạn:

|  |  |  |
| --- | --- | --- |
| **Đối tượng** | **Ngắn hạn** | **Dài hạn** |
| Trường học | Doanh thu giảm | Thương hiệu bị đánh giá thấp |
| Chi phí tăng | Số học viên tuyển sinh bị giảm các năm học sau |
| Học viên | Chi phí học tập học viên tăng mà không đạt được mục đích | Không có việc làm |
| Tâm lý bị tiêu cực, chán nản | Thu nhập bị sụt giảm |
| Không có bằng cấp và kỹ năng nghề nghiệp | Khả năng phạm tội tăng |
| Xã hội | Nạn thất nghiệp gia tăng | Chi phí công đầu tư cho giáo dục, phúc lợi xã hội tăng |
| Chi phí đầu tư giáo dục tăng | Tỷ lệ phạm tội gia tăng |

##### Thống kê ảnh hưởng của nghỉ học tới trường học, học viên và xã hội.

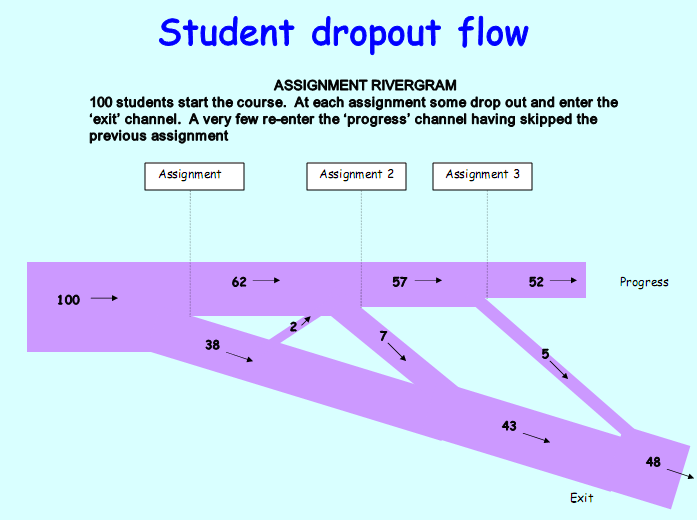
## Nguyên nhân dẫn đến tình trạng nghỉ học ở môi trường đào tạo từ xa

Nghiên cứu quốc tế tại các môi trường đào tạo trực tuyến chỉ ra các nguyên nhân chính như sau:



##### Nguyên nhân dropout của các chương trình online

Đối với Đại học Mở Anh Quốc tỉ lệ học viên nghỉ học thay đổi nhiều theo các thời điểm tham gia vào khóa học của từng nhóm học viên (*Ormond Simson, International Forum on Student Retention in Higher Education, 2009*)



##### Luồng nghỉ học của sinh viên

**Nguyên nhân của vấn đề nghị học tại HOU E-Learning**

Có rất nhiều các nguyên nhân dẫn đến tình trạng nghỉ học của học viên học tại môi trường đào tạo từ xa. Các ảnh hưởng này cũng xuất phát từ nhà trường, học viên và môi trường xã hội. Qua phân tích số liệu và phỏng vấn nhóm nhỏ những nguyên nhân này có nhiềm điểm khác so với nguyên nhân của các học viên tại Mỹ và Anh.

* Do không có động lực học tập (biểu hiện: Bận công việc)
* Đi học trường khác
* Do kinh tế
* Do không hoàn thành được học phần
* Không hài lòng với chất lượng
* Không đăng nhập hoặc không làm bài tập được
* Do điều kiện công tác (ví dụ đi công tác dài hạn ở nước ngoài)

**Tổng hợp các nguyên nhân nghỉ học:**

Chúng tôi đã tổng hợp một số nguyên nhân dẫn đến tình trạng nghỉ học của học viên:

| **Nguyên nhân** | **Trước quá trình học** | **Trong quá trình học** |
| --- | --- | --- |
| Trường học | Tư vấn chưa đúng (khác so với quá trình học)  **Vergidis & Panagiotakopoulos, 2002**  Đề cương chương trình không đầy đủ, thiếu chi tiết  Giảng viên không đủ tiêu chuẩn/thiếu kinh nghiệm | Môi trường học tập và phương pháp giảng dạy không thu hút HV  **Keller,1987**  Mức độ hài lòng thấp của học viên v̀ môi trường học trực tuyến (Vấn đề về cách thức quản lý của CVHT vấn đề về công nghệ,…)  **Gortan & ereb, 2007; Keller; Levy, 2007; Liaw**  Học liệu không: đầy đủ, trực quan và đa tương tác  Thiếu giám sát, quản lý, thiếu sự hộ trợ sinh viên |
| Học viên | Nền tảng kiến thức của HV chưa đáp ứng  Không đủ tài chính để học tập  Kỹ năng công nghệ của học viên không đủ đáp ứng  **Parker; Boyd, 2004**  Mục tiêu giáo dục của học viên sai khác so với mục tiêu giáo dục của chương trình đào tạo  **Parker, 1995**  Kĩ năng mềm của học viên (ghi nhớ, quản lý thời gian, đọc, viết,..) KHÔNG đủ đáp ứng  **Nash,2005; Liaw, 2008** | Học viên không đủ niềm tin, kiên trì để học trong môi trường từ xa **Chacon-Duque, 1987; Liaw, 2008**  Học viên không sắp xếp được thời gian để theo học  Học viên Không đủ điều kiện tốt nghiệp  Mục tiêu giáo dục của học viên thay đổi trong quá trình học  Học viên không thu được kiến thức và kĩ năng mà họ đặt ra  **Running head: DROPOUTS IN E-LEARNING COURSES**  HV không đánh giá đúng thời gian và sự nỗ lực cần thiết cho khóa học  **Nash, 2005; Seifert et al., 2008; Vergidis & Panagiotakopoulos** |
| Môi trường xã hội | Áp lực gia đình, áp lực công việc làm thêm  Xu hướng xã hội không phù hợp | Công nghệ giảng dạy không phù hợp, lỗi thời  Sự cố cuộc sống không mong đợi xảy ra (tai nạn, gia đình, tệ nạn, ...)  Xu hướng ngành nghề xã hội bị thay đổi (tác động giảm) |

##### Thống kê các nguyên dẫn dẫn tới NGHỈ HỌC của học viên

## Các hướng giải quyết cho vấn đề nghỉ ở HOU E-learning

Để giải quyết một phần vấn đề này Trung tâm đào tạo từ xa Elearning đã đưa một số định hướng như sau.

Đối với học viên sắp hoặc đã nghỉ, các biện pháp là:

* Tái tư vấn thông tin
* Bảo lưu kết quả
* Hỗ trợ chuẩn hóa hồ̀ sơ
* Chuyển đổi ngành học cho học viên
* Thay đổi chương trình đào tạo
* Hỗ trợ tư vấn tài chính

Để tránh thụ động trong giải quyết vấn đề nghỉ học này, chúng tôi đã đưa ra một số phương pháp ngăn ngừa.

Đối với các học viên đang theo học, chúng tôi:

* Dự báo tiềm năng NGHỈ HỌC dựa trên thông tin trong quá trình học với các thông tin về tình trạng đóng học phí của học viên (thời hạn đóng học phí, số tiền đóng, …), thông tin quá trình học tập, thông tin tương tác học viên với môi trường học (tình trạng tham gia lớp học qua môi trường mạng, đăng nhập, trao đổi thảo luận trên diễn dàn, H24/72, 3D, làm bài luyện tập, bài tập về nhà, bài tập kỹ năng, nhóm,…), hồ sơ học viên.
* Tổ chức các hoạt động tăng cường tương tác 2 chiều
* Hỗ trợ các hoạt động tăng cường kỹ năng (kỹ năng cá nhân, kỹ năng giao tiếp, kỹ năng làm việc nhóm, kỹ năng học tập và làm việc qua môi trường mạng, kỹ năng phỏng vấn, nhập môn Elearning, ….)

Song song với đó, chúng tôi chủ động giải quyết vấn đề nghỉ học của học viên ngay tại thời điểm trước khi vào học (trong quá trình tư vấn). Một số biện pháp chủ yếu được đưa ra là:

* Tư vấn chính xác
* Chọn lọc đúng đối tượng
* Kiểm tra tư vấn chéo
* **Dự báo tiềm năng NGHỈ HỌC dựa trên hồ̀ sơ thông tin học viên**

Vấn đề dự báo khả năng của học viên hiện tại vẫn đang dựa trên kinh nghiệm của các cố vấn học tập và chưa có một quy trình cụ thể. Chính vì vậy, chúng tôi đã thực hiện nghiên cứu đề tài với nội dung: **KHAI PHÁ DỮ LIỆU: DỰ BÁO TỶ LỆ NGHỈ HỌC CỦA CÁC NHÓM HỌC VIÊN HỆ TỪ XA THEO PHƯƠNG THỨC E-LEARNING**. Kết quả của đề tài nghiên cứu sẽ là một số mô hình cho phép dự báo được khả năng của học viên dựa trên thông tin hồ sơ của học viên.

# Tóm tắt cơ sở lý thuyết và những kết quả nghiên cứu

## Khái niệm về khai phá dữ liệu

Khai phá dữ liệu (Data mining): là việc trích rút ra được các mẫu hoặc tri thức quan trọng từ một lượng dữ liệu lớn.

Một số tên gọi khác: Khám phá tri thức trong các cơ sở dữ liệu (knowledge discovery (mining) in database-KDD), trích rút tri thức (Knowledge extraction), phân tích mẫu, phân tích dữ liệu (Data/pattern analysis) , …

Khai phá dữ liệu khác với: tìm kiếm thông tin (information retrieval), xử lý câu truy vấn (SQL) đối với các cơ sở dữ liệu

CÁC MẪU

KHAI PHÁ DỮ LIỆU

DỮ LIỆU

##### Khái niệm khai phá dữ liệu

## Các phương pháp khai phá dữ liệu

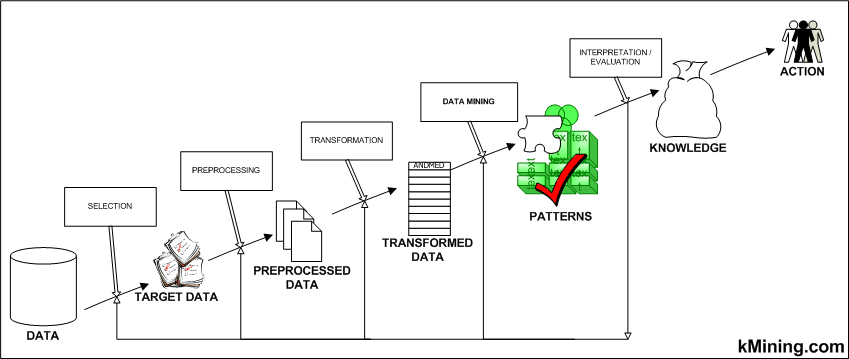
* Phân lớp (classification)
* Phân cụm (Clustering)
* Luật kết hợp (Associations)
* (Visualization)
* Tổng hợp (Summarization)
* Dò tìm độ lệch (Deviation Detection)

## Quy trình khám phá tri thức (Knowledge discovery in databases – KDD)

Khám phá tri thức là sự giao nhau của năm lĩnh vực chính: trí tuệ nhân tạo, thống kê, công nghệ tính toán mềm, tính toán hiệu năng cao, cơ sở dữ liệu.

##### Mô hình khám phá tri thức

Dưới đây là quy trình khám phá tri thức:



##### Quy trình khai phá tri thức

Các bước của quá trình khám phá tri thức KDD:

Bước 1: Tìm hiểu phạm vi của bài toán: mục đích của bài toán, các tri thức cụ thể của bài toán.

Bước 2: Tạo ra một tập dữ liệu mục tiêu: lựa chọn dữ liệu

Bước 3: Làm sạch và tiền xử lý dữ liệu (có thể chiếm 60% kết quả của sự nỗ lực).

Bước 4: Giảm kích thước của dữ liệu và chuyển đổi dữ liệu: tìm những thuộc tính hữu ích, giảm số chiều (số thuộc tính), biểu diễn bất biến.

Bước 5: Lựa chọn chức năng khai phá dữ liệu: tóm tắt hóa, phân loại/phân lớp, hồi quy/dự đoán, kết hợp, phân cụm.

Bước 6: Lựa chọn/phát triển các giải thuật khai phá dữ liệu phù hợp.

Bước 7: Tiến hành quá trình khai phá dữ liệu.

Bước 8: Đánh giá mẫu thu được và biểu diễn tri thức: hiển thị hóa, chuyển đổi, bỏ đi các mẫu dư thừa…

Bước 9: Sử dụng các tri thức được khám phá.

## Kết quả nghiên cứu từ những trường đại học khác

Dưới đây là một số nghiên cứu về vấn đề NGHỈ HỌC của những trường đại học khác trên thế giới chúng tôi đã tổng hợp:

| **Bài báo** | **Tác giả/ Trường** | **Năm** | **Tóm tắt kết quả** |
| --- | --- | --- | --- |
| Flight risk: identifying potential drop outs | Prakash V. Arumugam  School of Business and Admisistration  Wawasan Open University | 2010 | Mô hình dự báo khả năng bỏ học của học viên sau đại học. Đề xuât nghiên cứu làm thế nào để cải thiện tỉ lệ duy trì ở các trường ĐH |
| Improving graduation rates of open and distance learners via online community | Richard Ng  Director, Perak Regional Learning Centre  Open University Malaysia | 2010 | Nghiên cứu tập trung vào đội ngũ trực tuyến giúp học viên cam kết ở lại các chương trình của họ. |
| A review of online course drop out research and suggestions for future research | Youngju Lee, Jaeho Choi, & Patrick McGuire  University of Virginia | 2010 | Xác định được 3 nhóm nhân tố ảnh hưởng tới quyết định bỏ học của học viên.Phân tích mức độ ảnh hưởng của từng nhân tố. Phân tích các chiến lược khắc phục drop-out. |
| An artificial neural network for predicting student graduation outcomes | Stamos T. Karamouzis and Andreas Vrettos | 2008 | Sử dụng mạng Perceptron 3 lớp dự báo khả năng tốt nghiệp. Luyện với 1407 hồ sơ của các học viên trường Waubonsee College; Tập dữ liệu chia làm 2 tập. Kết quả: Trên tập luyện với số mẫu 1,100 là 77% và tập dữ liệu kiểm tra với số mẫu 307 là 68%. |
| Predicting students drop out: a case study | Gerben W. Dekker /  Department of Electrical Engineering,  Eindhoven University of Technology | 2009 | Mô hình dự báo NGHỈ HỌC tại Electrical Engineering Department of Eindhoven University of Technology sử dụng WEKA . Kết quả phân lớp đúng từ 75 đến 80% dựa các dữ liệu trước đại học và dữ liệu kỳ đầu tiên. |
| Student dropout analysis with application of data mining methods | M. Jadrić, Ž. Garača, M. Ćukušić | 2010 | Trình ứng dụng phân tích sinh viên NGHỈ HỌC sử dụng phương pháp phân lớp (hồi qui, mạng nơ ron, cây quyết định). Luyện với 40% tập dữ liệu 2000 mẫu trên 715 sinh viên của Faculty of Economics Information System. |
| Identifying factors that predict student success in a community college online distance learning course | Johnelle Bryson Welsh, A. G. S., B. S., M. S.  / University Of North Texas | 2007 | Xác định các nhân tố để dự báo học viên tốt nghiệp trong chương trình từ xa trực tuyến. Phân tích trên mẫu dữ liệu gồm 926 học viên. |
| Enrollment prediction models using data mining | Ashutosh Nandeshwar; Subodh Chaudhari | 2009 | Mô hình dự báo tuyển sinh dùng khai phá dữ liệu bằng WEKA phân lớp sử dụng cây quyết định. Kết quả dự báo chính xác trong khoảng 83%-84%. Dữ liệu được lấy tại West Virginia University |
| Data mining algorithms to classify students | Cristóbal Romero, Sebastián Ventura, Pedro G. Espejo and César Hervás  / Computer Science Department, Córdoba University, Spain | 2008 | So sánh câc phương pháp khai phá dữ liệu để phân lớp sinh viên dựa vào 438 dữ liệu sinh viên trong 7 khóa học từ hệ thống Moodle của trường Cordoba University .Phát triển và tích hợp công cụ Moodle Mining Tool hỗ trợ các giảng viên trực tuyến. |
| Prediction of student academic performance by an application of data mining techniques | Sajadin Sembiring, M. Zarlis, Dedy Hartama, Ramliana S, Elvi Wani | 2011 | Mô hình dự báo chất lượng học viên của trường University Malaysia Pahang với 300 mẫu dữ liệu kỳ 3 của năm 07-08. Sử dụng phương pháp phân lớp SSVM và phân cụm K-Means. Kết quả tốt nhất là 93.7% và thấp nhất là 61%. |
|  |  |  |  |

## Phân tích vấn đề nghỉ học ở HOU – Elearning

Tập dữ liệu học viên được phân thành bốn lớp: Đang học, Đã tốt nghiệp, Đã NGHỈ HỌC, Đang tư vấn. Nên bản chất bài toán của chúng tôi gặp phải tại HOU-Elearning là bài toán phân lớp. Trong “phân lớp” có những kĩ thuật như:

* Mạng Noron
* Mạng Bayes
* Cây quyết định
* Mô hình Markov ẩn
* Luật cơ sở
* Trí nhớ dựa trên các nguyên nhân
* Giải thuật di truyền

Do mạng noron có những ưu điểm như:

* Cho phép xấp xỉ những ánh xạ phi tuyến tùy ý
* Là hệ thống xử lý song song làm tăng tốc độ tính toán cho phép đáp ứng khả năng tính toán thời gian thực và chính xác
* Là hệ học và thích nghi, khi mạng được huấn luyện từ các dữ liệu quá khứ, đồng thời có khả năng khái quát hóa khi dữ liệu vào bị thiếu hoặc không đầy đủ
* Phù hợp với các hệ thống nhận dạng, chuẩn đoán kỹ thuật...

Nên chúng tôi chọn kĩ thuật mạng nơron để giải quyết bài toán “"**DỰ BÁO TỶ LỆ NGHỈ HỌC CỦA CÁC NHÓM HỌC VIÊN HỆTỪ XA THEO PHƯƠNG THỨC E-LEARNING**.

# Quá trình và kết quả nghiên cứu và phát triển

## Thu thập dữ liệu

Chúng tôi đã thu thập được các dữ liệu cần thiết cho các mô hình dự báo như: Dữ liệu về hồ sơ học viên gồm có 7476 học viên và 30 cột thuộc tính (Tuổi, giới tính, bằng tốt nghiệp cấp 3, dân tộc thiểu số YN, tôn giáo,…) từ ngày bắt đầu tuyển sinh đến ngày 5/10/2011.

Dưới đây là thống kê về các nguồn dữ liệu đã lấy được:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Loại dữ liệu** | **Nguồn** | **SL trường thông tin** | **SL bản ghi thu thập** |
| 1 | Hồ sơ học viên | SCM | 30 | 7,476 |
| 2 | Kết quả thi PTTH | CRM | 4 | 7,476 |
| 3 | Ngành học, đối tượng học | CVHT | 3 | 20 |
| 4 | Thông tin quá trình tư vấn | CRM | 5 | 712,564 |
| 5 | Bằng cấp, chứng chỉ | SCM | 10 | 1,024 |
| 6 | Hồ sơ học tập học viên | SCM | 12 | 642,715 |
| 7 | Lộ trình thu học phí | CVHT | 7 | 160 |
| 8 | Thông tin đóng học phí | SCM | 8 | 21,542 |
| 9 | Qui trình tuyển sinh | TVTS | 4 | 20 |
| 10 | Thông tin quá trình học | LMS | 3 | 614,523 |

##### Thống kê các nguồn dữ liệu

## Các mô hình dự báo

Thông qua quá trình nghiên cứu trên yêu cầu của Trung tâm và các loại dữ liệu thu thập được, chúng tôi đã đề xuất một số mô hình dự báo để tiến hành nghiên cứu và triển khai như sau:

|  |  |  |
| --- | --- | --- |
| **Mô hình** | **Đầu vào** | **Đầu ra** |
| Dự báo tình trạng nghỉ học của học viên | Hồ sơ của những học viên đã tốt nghiệp và những học viên đã NGHỈ HỌC | **Trạng thái học viên** 0 – NGHỈ HỌC 1 – Tốt nghiệp |
| Dự báo tình trạng học tập của học viên qua 1 kỳ xác định (Ký hiệu: kỳ x =2,3,4) | Hồ sơ của những học viên đã học qua kỳ x và những học viên đã NGHỈ HỌC trước kỳ x | **Trạng thái học viên** 0 – NGHỈ HỌC  1 – Học vượt qua kỳ x |
| Dự báo tiềm năng nghỉ học của học viên trong quá trình học | Kỳ x;  Hồ sơ học viên;  Qui mô lớp; Điểm tích lũy trung bình; Số lần đăng nhập diễn đàn; Tỉ lệ NGHỈ HỌC kỳ trước | **Trạng thái học viên** Màu xanh (Học thêm được ít nhất 2 kỳ nữa)  Màu vàng (Kỳ tới sẽ NGHỈ HỌC) Màu đỏ (Kỳ này NGHỈ HỌC) |

## Sản phẩm phần mềm và các bước xử lý

### Khái niệm tiền xử lý tập dữ liệu

1. Tiền xử lý dữ liệu nghĩa là sự thay đổi giá trị của tập dữ liệu trước khi đưa vào một mạng neuro. Tiền xử lý biến đổi dữ liệu để làm cho nó thích hợp với hệ thống mạng neuro. Ví dụ, định tỉ lệ (scaling) giá trị chữ số và chuyển đổi giá trị văn bản thành số.
2. Sau tiền xử lý (post-preprocessing) nghĩa là thay đổi đầu ra của mạng neuro để làm cho nó dễ hiểu bởi người dùng hoặc thích hợp với môi trường thế giới thực. Định lại tỉ lệ giá trị hoặc mã hóa lại các danh mục (categories).
3. Mạng neuro chỉ làm việc với dữ liệu số. Thường thì kiểu dữ liệu ngày tháng và thời gian là giống với dữ liệu văn bản (categories) cần được xử lý. Do đó, chuyển đổi dữ liệu là cần thiết trước khi đưa chúng tới đầu vào của mạng (preprocessing) và sau đó nhận chúng tại đầu ra (post-preprocessing).
4. Giá trị văn bản có thể chỉ được sử dụng nếu chúng đại diện cho một tập giá trị cuối cùng. Cột có các giá trị đó được gọi là categories columns. Chúng cần được mã thành giá trị số trước khi đưa vào luyện mạng. *Xem thêm phần mã hóa cột dữ liệu văn bản.*
5. Giá trị số cần được định lại tỉ lệ trước khi đưa chúng vào luyện mạng, bởi vì các neurons nhân tạo có một khoảng giá trị giới hạn các giá trị tính toán. Do đó, thông tin đầu vào nên được giảm xuống khoảng giá trị này và thông tin đầu ra cũng nhận được từ mạng bằng việc biến đổi ngược lại. *Xem thêm phần định lại tỉ lệ cột dữ liệu số.*
6. Cột với thông tin ngày tháng và thời gian thường được mã hóa theo chu kỳ bởi vì trong nhiều trường hợp ảnh hưởng của các dữ liệu đó vào vấn đề định kỳ - xác định chu kỳ - (ví dụ, đầu ra có thể phụ thuộc vào các ngày trong tuần hoặc tháng). Chu kỳ đó các dữ liệu cần phải được mã hóa thông thường nhất cần được xác định một cách rõ ràng (tường minh). *Xem thêm phần mã hóa cột dữ liệu ngày tháng.*

### Các thao tác tiền xử lý

Định lại tỉ lệ cột dữ liệu số (Scaling Numeric Columns)

1. Cột dữ liệu số được định lại tỉ lệ theo công thức sau



Trong đó,

* X – giá trị thực của dữ liệu
* Xmin – giá trị thực nhỏ nhất của cột dữ liệu (minimum actual value)
* Xmax – giá trị thực lớn nhất của cột dữ liệu (maximum actual value)
* SRmin – giới hạn dưới của khoảng tỉ lệ mới (lower scaling range)
* Srmax – giới hạn trên của khoảng tỉ lệ mới (upper scaling range)
* SF – Hệ số tỉ lệ

1. Khoảng định tỉ lệ

* Đối với cột dữ liệu vào được định lại tỉ lệ trong khoảng [-1 .. 1].
* Đối với cột đích khoảng định tỉ lệ phụ thuộc vào hàm kích hoạt sử dụng ở lớp xuất (output layer):

|  |  |
| --- | --- |
| **Output layer activation function** | **Scaling range** |
| Linear | **[-1 .. 1]** |
| Logistic | **[0 .. 1]** |
| Hyperbolic Tangent | **[-1 .. 1]** |
| Softmax | **[0 .. 1]** |

* Cột đích sẽ được tiền xử lý lại tự động nếu hàm kích hoạt thay đổi trong bước tinh chỉnh mạng ở lớp xuất (output layer).

1. Mặc định, Xmin và Xmax được lấy từ tập dữ liệu đầu vào.

### Mã hóa cột dữ liệu văn bản (Encoding Categorical Columns)

1. Cột dữ liệu văn bản được tự động mã hóa trong quá trình tiền xử lý dữ liệu sử dụng mã hóa dạng One-of-N, Binary hoặc Numeric.
2. Mã hóa dạng **One-of-N** nghĩa là 1 cột với N giá trị văn bản (categories values) khác nhau được mã thành tập N cột số, với mỗi một cột tương ứng mỗi giá trị trong N giá trị. Ví dụ, có cột Capacity với các giá trị “Low”, “Medium” and “High”. Vậy “Low” sẽ được biểu diễn (mã hóa) thành {1, 0, 0}, “Medium” thành {0, 1, 0}, “High” thành {0, 0, 1}.
3. Mã hóa dạng **Binary** nghĩa là 1 cột với N giá trị văn bản (categories values) khác nhau được mã thành tập M cột nhị phân, trong đó M bằng với độ dài số nhị phân được dùng có thể biểu diễn được N giá trị khác nhau. Ví dụ, cột Color có các giá trị “Red”, “Yellow”, “Green”, “Blue”, và “White” sẽ được mã thành 3 cột nhị phân. Và Red sẽ được biểu diễn thành {0, 0, 0}, Yellow là {0, 0, 1}, Green là {0, 1, 0}, Blue là {0, 1, 1}, và White là {1, 0, 0}. Ta luôn có được M <= N (dấu bằng xảy ra khi M = N = 1).
4. Mã hóa dạng **Numeric** nghĩa là 1 cột với N giá trị văn bản (categories values) khác nhau được mã thành 1 cột số, với 1 số nguyên sẽ được gán cho 1 giá trị văn bản. Ví dụ, cột Capacity có các giá trị “Low”, “Medium”, và “High”. Vậy Low được đại diện bởi giá trị {1}, Medium là {2}, và High là {3}.
5. Kiểu mã hóa dạng One-of-N được khuyến cáo sử dụng bởi tính trực quan hóa. Mã hóa dạng Binary là lựa chọn tốt để giảm số nút nhập và mã hóa dạng Numeric có thể được sử dụng cho dữ liệu văn bản có thứ tự (ordered categories). Chú ý: trong mã hóa dạng One-of-N và Binary thì -1 có thể được sử dụng thay cho 0 phụ thuộc vào hàm kích hoạt.

### Mã hóa cột dữ liệu Date/Time (Encoding Date/Time Columns)

1. Thường thì cột date/time chỉ là các trường hợp định danh và không hữu ích cho việc nâng cao tính chính xác dự báo. Tuy nhiên, đôi khi chúng lại thể hiện thống tin quan trọng hữu ích cho việc giảm lỗi dự báo. Ví dụ, doanh số bán hàng có thể phụ thuộc vào ngày trong tuần hoặc theo mùa.
2. Nếu như chúng ta mã giá trị date/time bằng những con số, thì chúng ta mất thông tin quan trọng về tính chu kỳ. Ví dụ, nếu chúng ta mã Sunday là 1 và Saturday là 7, mạng neuro sẽ “nghĩ” rằng Sunday và Saturday một cách mơ màng, điều đó là sai. Chúng ta cần mã hóa chu kỳ để đưa giá trị date/time một cách chính xác cho mạng neuro sử dụng.
3. Các cột Date và time được tự động mã hóa trong quá trình tiền xử lý. Kiểu dữ liệu Date/Time được mã trong 2 cột. Các giá trị trong các cột mới được tính bởi công thức sau:



Trong đó,

X – cột giá trị

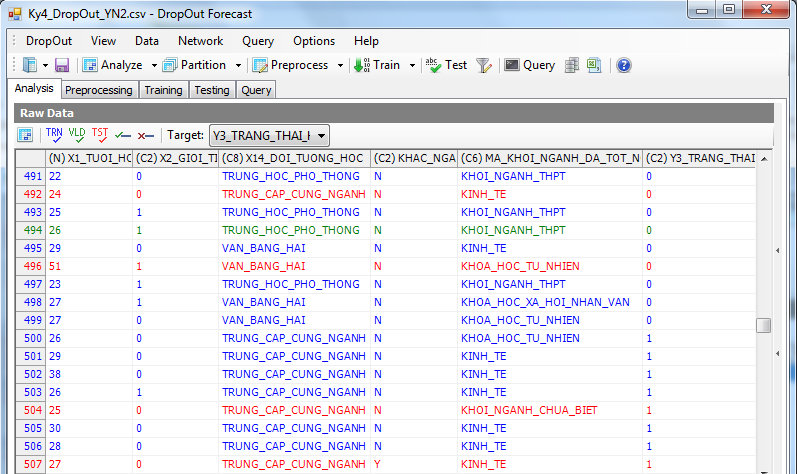
P – tính chu kỳ

Ví dụ, để mã hóa một chu kỳ theo tuần P nên được gán là 7. Đối với cột Time thì P luôn được gán là 24.

Các cột dữ liệu kiểu Date cũng có thể được mã như kiểu dữ liệu văn bản (categories) bằng cách sử dụng mã hóa dạng One-of-N. Với kiểu

### Sản phẩm phần mềm

Trong quá trình thực hiện đề tài chúng tôi đã xây dựng thành công sản phẩm phần mềm mô phỏng kỹ thuật mô hình mạng Neuro cài đặt luật học thích nghi. Mục đích của sản phẩm phần mềm là chạy kiểm thử tập dữ liệu thu thập được, kiểm chứng độ chính xác kết quả các mô hình thực hiện trong đề tài. Hỗ trợ truy vấn kết quả của các mô hình với tập dữ liệu mới tương ứng.

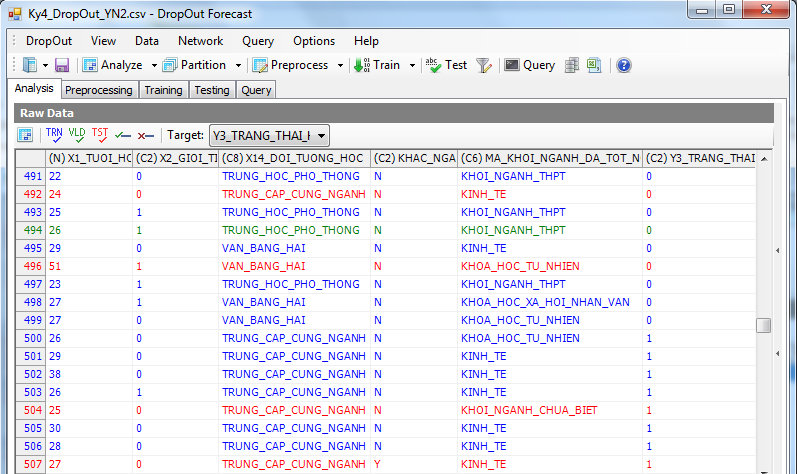


##### Giao diện chính phần mềm

Nhằm mục đích giải quyết vấn đề thế giới thực, chương trình chỉ bao gồm một số tính năng cơ bản và khá dễ dàng sử dụng. Sản phẩm phần mềm dự báo được chúng tôi xây dựng có các chức năng sau( Hình ảnh minh họa là mô hình “Kỳ 4 NGHỈ HỌC Có/Không”:

#### Chức năng: phân tích tập dữ liệu

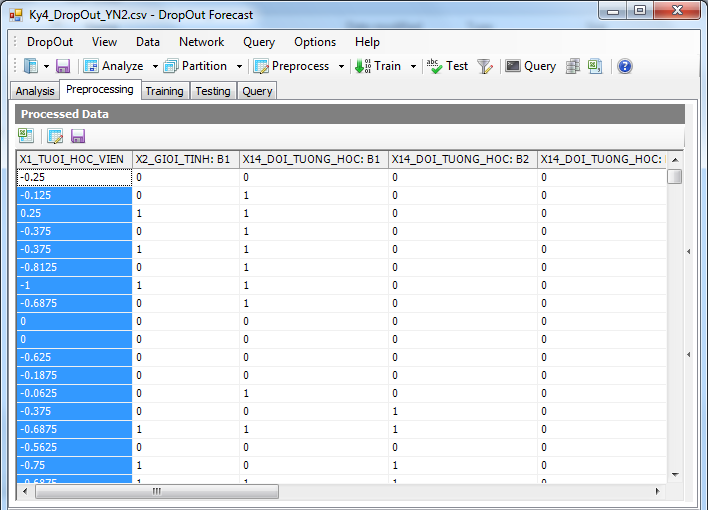
Dữ liệu thô được đưa vào chương trình tại tab Analysis, các bản ghi nhiễu, thiếu hoặc dữ liệu dị thường sẽ được tự động lọc tại đây trước khi chuyển sang bước tiền xử lý dữ liệu. Tại tab này các cột (thuộc tính) dữ liệu được tự động nhận dạng kiểu dữ liệu (có thể tùy chỉnh bằng tay). Đồng thời tập dữ liệu được phân chia làm ba phần dữ liệu tương ứng với tập dữ liệu luyện mạng (blue), tập dữ liệu đánh giá mạng (green), tập dữ liệu kiểm tra mạng (red) và mặc định được chia tỉ lệ 0.68:0.16:0.16.



##### Chức năng phân tích tập dữ liệu

#### Chức năng: Tiền xử lý dữ liệu

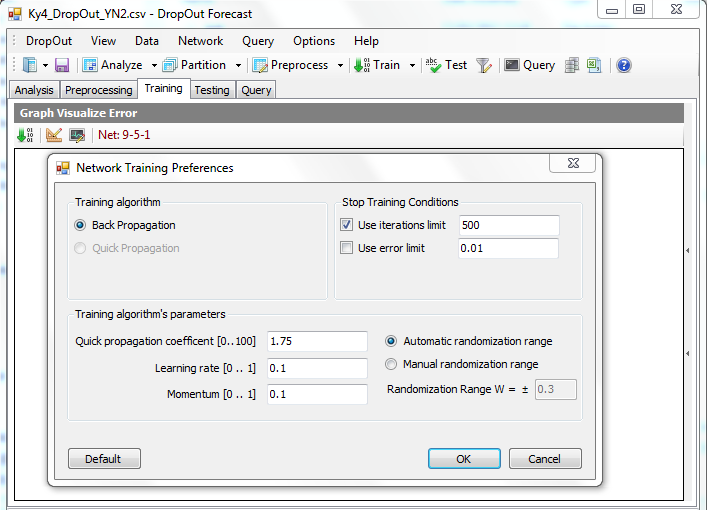
Tại Analysis tab click Preprocess button dữ liệu được tự động xử lý sau quá trình phân tích và phân chia tập dữ liệu tương ứng các kiểu dữ liệu đưa tại bước trước. Các dữ liệu được chuyển đổi từ kiểu dữ liệu phổ thông được trình bày ở phần trước sang kiểu dữ liệu phù hợp với mạng như định lại tỉ lệ kiểu dữ liệu có thứ tự, mã hóa dữ liệu kiểu văn bản, kiểu dữ liệu ngày tháng, …



##### Chức năng tiền xử lý dữ liệu

#### Chức năng: Luyện mạng

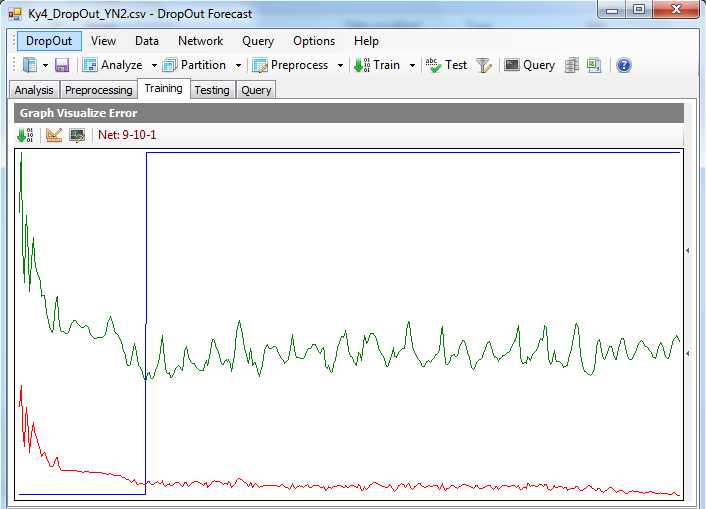
Với bộ dữ liệu được phần tích và xử lý. Tại tab Training chúng ta thấy hệ thống hỗ trợ tự động thiết kế cấu trúc mạng mặc định, trước khi luyện mạng các tham số mạng và kiến trúc mạng có thể được thay đổi bởi người sử dụng để mạng có thể hoạt động được tốt nhất.



##### Tùy chỉnh tham số mạng

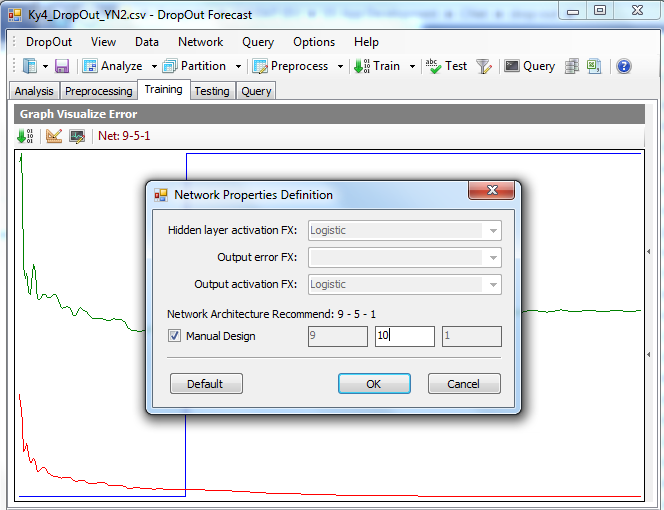
Sau khi hoàn tất bước trên mạng được tự động luyện với bộ dữ liệu vào và các tham số tùy chọn người dùng. Kết quả luyện mạng được trực quan bằng đồ thị hàm lỗi của mạng trên các tập dữ liệu trực tiếp tham gia trong quá trình luyện mạng tương ứng (training set, validation set).

Nhận thấy rằng, đồ thị hàm lỗi của mạng trên tập dữ liệu luyện mạng (training set) là thấp hơn rất nhiều so với đồ thị hàm lỗi của mạng trên tập dữ liệu kiểm tra mạng (validation set), lý do là mạng có năng lực để làm cho các bộ trọng của mạng khớp nhất với những gì mà nó được học. Đó là lý do tại sao ta đưa thêm vào mạng bộ dữ liệu kiểm tra mạng. Vậy tại thời điểm mà lỗi trên tập dữ liệu kiểm tra mạng không giảm nữa bắt đầu tăng ta có được mạng với bộ trọng tốt nhất cho dự báo với độ chính xác tốt nhất.



##### Đồ thị hàm lỗi của mạng trên các tập dữ liệu

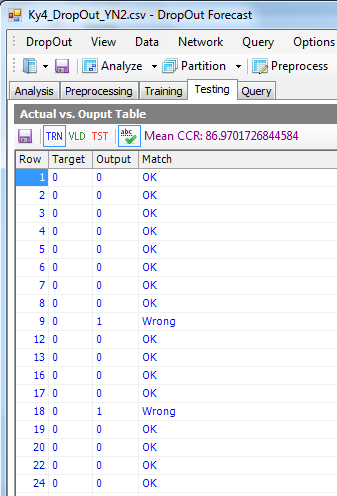
#### Chức năng: Thiết lập kiến trúc mạng



##### Thiết lập kiến trúc mạng

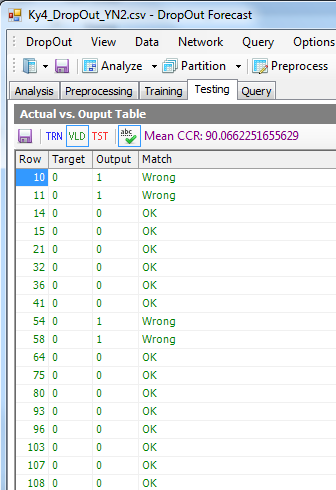
#### Chức năng: Kiểm tra kết quả luyện mạng

Kết quả luyện mạng ta có thể thấy được tại tab Test. Rõ ràng với mạng được chọn kết quả dự đoạn trên các tập dữ liệu là khác nhau. Và kết phần trăm đúng CCR trên tập dữ liệu Test được sử dụng để dự báo trên dữ liệu học viên mới



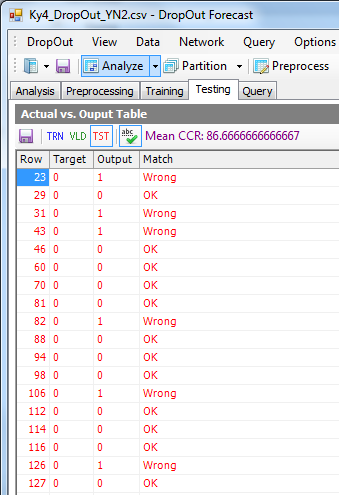
##### Phần trăm đúng CCR trên các tập dữ liệu tham gia luyện mạng

#### Chức năng: kiểm tra tính phù hợp của mạng



##### Phần trăn đúng dựa trên dữ liệu xác minh

#### Chức năng: *Kiểm tra*



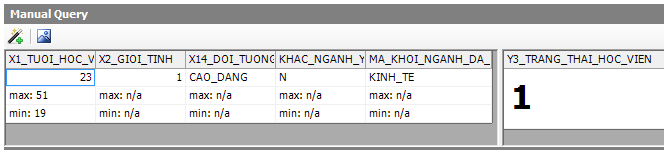
##### Phần trăm đúng trên tập dữ liệu Test

#### Chức năng: Truy vấn kết quả

Hỏi với với hồ sơ của một học viên thì trạng thái của học viên này tại kỳ 4 là như thế nào?

Ví dụ1: ta có hồ sơ của một học viên như sau.

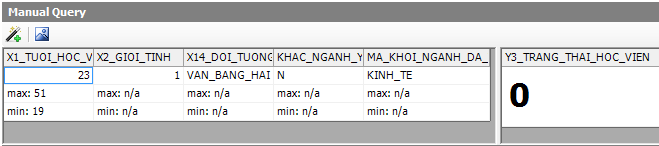
* Tuổi: 23
* Giới tính: 1 (nam)
* Đối tượng học: Cao đẳng
* Khác ngành: Yes
* Mã khối ngành đã tốt nghiệp: Kinh tế



##### Kết quả truy vấn 1

Ví dụ2: Ta có hồ sơ của một học viên như sau.

* Tuổi: 23
* Giới tính: 1 (nam)
* Đối tượng học: Văn bằng hai
* Khác ngành: N
* Mã khối ngành đã tốt nghiệp: Kinh tế



##### Kết quả truy vấn 2

## Báo cáo phân tích dữ liệu nghỉ học

Căn cứ trên 10 loại dữ liệu thập được gồm 87 trường thông tin và khoảng 2 triệu bản ghi dữ liệu, thông qua quá trình tiền xử lý (trích lọc thông tin, loại nhiễu,…) và đưa vào mô hình Mạng Noron để luyện, chúng tôi thu được kết quả như sau:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Dữ liệu mẫu** | **Tỷ lệ mẫu** | **NGHỈ HỌC**  **YN** | **NGHỈ HỌC SAU KỲ** | | |
| **Kỳ 2** | **Kỳ 3** | **Kỳ 4** |
| **Tập dữ liệu thu thập** | **100%** | **78** | **1364** | **1351** | **938** |
| **Tập dữ liệu luyện** | **68%** | **53** | **927** | **918** | **637** |
| **Tập dữ liệu kiểm tra phù hợp** | **16%** | **12** | **218** | **216** | **150** |
| **Tập dữ liệu kiểm tra** | **16%** | **12** | **218** | **216** | **150** |
| **Kết quả dự báo đúng** |  | **91.67%** | **68.81%** | **77.68%** | **86.00%** |

KẾT LUẬN VÀ KIẾN NGHỊ

Giảm thiểu tỉ lệ học viên bỏ học là thách thức lớn trong môi trường đào tạo E-Learning. Mặc dù đã có nhiều nghiên cứu và đánh giá tại các cơ sở giáo dục ứng dụng E-Learning trên thế giới, HOU E-Learning vẫn phải tìm ra những hướng đi mới của mình để có thể áp dụng hiệu quả trong môi trường Việt Nam nhằm giảm thiểu tỉ lệ bỏ học tại HOU E-Learning.

**Kết quả đạt được**

Phương pháp dự báo bằng khai phá dữ liệu bằng phần mềm cho phép tự động hóa quá trình phát hiện học viên có xác xuất bỏ học cao trong tương lai. Kết quả của phương pháp này cho phép Trung tâm đào tạo E-Learning triển khai các biện pháp chủ động như tư vấn sớm cho học viên về lựa chọn tiến độ học tập cũng như định hướng cuộc sống phù hợp để có thể kết thúc được khóa học.

Phương pháp này cũng dự đoán được các tiêu chí có độ nhạy lớn nhất với xác xuất nghỉ học. Đây là một trong những phát kiến quan trọng của đề tài. Tuy nhiên mối liên hệ trực tiếp giữa các tiêu chí này với các nguyên nhân bỏ học chính vẫn còn đòi hỏi các nghiên cứu tiếp theo.

Về mặt mô hình dự báo đã xác định được bước đầu mô hình dựa trên các yếu tố địa lý, nhân khẩu học và thông tin từ hồ sơ nhập học.

Về sản phẩm phần mềm nhóm nghiên cứu đã có được phần mềm trên nền tảng .NET cho phép tự động hóa phân tích dữ liệu được tập hợp từ phần mềm quản lý học viên SCM của Trung tâm.

**Hướng nghiên cứu**

Tiếp tục mở rộng mô hình phân tích là điểm quan trọng cần quan tâm trong thời gian tiếp theo. Đối tượng chính tiếp theo của mô hình sẽ là các kết quả học tập của học viên và hoạt động của học viên trong khóa học.

Để có thể thực hiện các phân tích một các thường xuyên hơn, phần mềm dự báo khả năng nghỉ học cần được tích hợp chặt chẽ vào các hệ thống thu thập dữ liệu như CRM, SCM, LMS,

Để có thể xác định kết quả nghiên cứu rõ ràng hơn các biện pháp kiểm định cần được xác định bằng các phương pháp khác như phỏng vấn, phỏng vấn nhóm nhỏ.

Vì là chương trình mới, phương pháp đào tạo của Trung tâm Đào tạo E-Learning có sự thay đổi theo từng thời gian, cùng thời gian đó uy tín của chương trình E-learning cũng có thay đổi (ví dụ có học viên tốt nghiệp), điều này dẫn đến mô hình và phương pháp phân tích cần có sự điều chỉnh phù hợp. Nhóm đề tài dự kiến sẽ phân tích và đưa ra mô hình mới theo từng 6 tháng.

**Kiến nghị**

Để xác định được tính chính xác và có thể điều chỉnh, mở rộng mô hình phân tích một cách hiệu quả, Trung tâm đào tạo E-Learning cần có kế hoạch ứng dụng kết quả nghiên cứu vào từng nhóm học viên. Thực hiện so sánh, đánh giá hiệu quả và đưa ra những góp ý với nhóm nghiên cứu.

Để mở rộng và tăng cường tính chính xác của mô hình phân tích, chúng ta cũng cần có sự đơn giản và đồng bộ hóa các dữ liệu đầu vào. Trên quan điểm quản trị đây chính là đòi hỏi về tính đồng đều trong chất lượng tư vấn và chất lượng dịch vụ học viên tại các Trung tâm địa phương.

Trong khuôn khổ vấn đề nghỉ học của học viên có những nguyên nhân mới mà có thể chỉ gặp phải tại Việt Nam và một số nước có trình độ phát triển tương đương như hiểu biết của người học về tin học, khả năng kết nối. Những điều kiện cơ bản này cần phải được cải thiện thông qua quá trình xây dựng chương trình đào tạo, chuẩn đầu ra và hỗ trợ cộng đồng …Những vấn đề này cần được các cấp quản lý quan tâm và giải quyết đồng bộ trong chiến lược phát triển kinh tế xã hội như các đề án xây dựng xã hội học tập, phát triển văn hóa nông thôn, phổ cập internet ở vùng sâu vùng xa…

Công nghệ E-Learning, với cách vận dụng thích hợp, chính là chìa khóa để vừa tăng trưởng về quy mô vừa đảm bảo được chất lượng đào tạo. Đối với các tổ chức giáo dục mở và từ xa như Viện Đại học Mở Hà Nội thì E-Learning sẽ biến người học thành người tự học thật sự và đây cũng chính là điểm nhấn quan trọng trong đổi mới phương pháp dạy và học tại Việt Nam trong giai đoạn tới.

# TÀI LIỆU THAM KHẢO

Abas, Z. W. & Fadzil, M.(2009). Towards a more effective engagement of learners and tutors in online discussion forums. AAOU Journal, 4(1), 60 – 67.

http://aaou.ut.web.id/jurnal\_aaou/2009/zoraini.pdf

Ali, A. (2008). Moving forward with e-learning: The experience of Open University Malaysia. Retrieved: April 23, 2010 from

http://iqri.oum.edu.my/v2/pdf/tan%20sri/Moving%20Forward%20with%20E-learning\_1.pdf

Andreas Becker.(2007). Drop Out Prevention using eLearning and Homeschooling No One Left Behind – Education for All – Well Educated People without Schools. IMCT – International Montessori Charter and Training LisAndi – Homeschooling and E-Learning Kultur- und Familienwerkstatt Stuttgart, ermany – Rawai, Phuket, Thailand

AshutoshNandeshwar, SubodhChaudhari.(2009). Enrollment Prediction Models Using DataMining

Astin, A. W.(1993). What matters in college. San Francisco: Jossey-Bass.

Bandura, A. (1973). Aggression: A social learning analysis. Englewood Cliffs: N.J., Prentice-Hall.

Bean, J. P., & Metzner, B. S.(1985). A conceptual model of non-traditional undergraduate learner attrition. Review of Educational Review, 55, 485–540.

Bell, F. (2003, September 17-18). Learning communities-reality or feel-good factor. Paper presented at the Education in a Changing Environment

Berge, Z. & Huang, Y.(2004). A Model for Sustainable Student Retention: A Holistic Perspective on the Student Dropout Problem with Special Attention to e-Learning. DEOSNEWS, Volume 13(5). Retrieved November 1, 2007, from

http://www.ed.psu.edu/acsde/deos/deosnews/deosnews13\_5.pdf

Boyd, D. (2004). The characteristics of successful on-line students. New Horizons in Adult Education, 18(2), 31-39. Retrieved March 29, 2009, from

http://www.nova.edu/~aed/newhorizons.html

Burgess, D. L. A., & Strong, D. S. D. (2003). Trends in online education: Case study at Southwest Missouri State University. Journal of Industrial Technology, 19(3), 1-5.

C.Romero and S.Ventura.(2007). Educational data mining:a survey from 1995 to 2005 , Expert Systems with Applications , no.33, pp.135 - 146

Carr, S. (2000). As distance education comes of age, the challenge is keeping students. Chronicle of Higher Education (online archives), 46(23). Retrieved February 15, 2009, from

http://www.chronicle.com.libproxy.library.wmich.edu/weekly/v46/i23/23a00101.htm

Carroll, J. B. (ed.)(1997). Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf. Cambridge, Mass.: Technology Press of Massachusetts Institute of Technology

Chacon-Duque, F. J.(1987). A multivariate model for evaluating distance higher education. College Park: Pennsylvania State University Press.

Clay Olivio J. et al (2009). Visual Function and Cognitive Speed of Processing Mediate Age- Related Decline in Memory Span and Fluid Intelligence. Journal of Aging and Health Volume 21 Number 4 June 2009 547-566

Coppola, N., Hiltz, S., Rotter, N.(2002). Becoming a virtual professor: Pedagogical roles and asynchronous learning networks. Journal of Management Information Systems, 18(4), 169. Retrieved January 31, 2009, from http://web.njit.edu/~cs222/technology/PTC%20606/Articles/Becoming%20a%20Virtual%20Professor.pdf

Dagger, D., & Wade, V. P. (2004). Evaluation of adaptive course construction toolkit ACCT. Retrieved March 1, 2009, from

De Simone, C. (2006). Preparing our teachers for distance education. College Teaching, 54(1), 183-184.

Deci, E. L. & Ryan, R. M. (1986). Self determination theory and the facilitation of intrinsic motivation, social development and well-being. American Psychologist. 55(1), 68 – 78.

Flood, J. (2002). Read all about it: Online learning facing 80% attrition rates. Tukish Online Journal of Distance Education, 3(2), 79-84.

Flowers, J. (2001). Online learning needs and technology education. Journal of Technology Education, 13(1), 17-30.

Juan R. Rabuñal, Julián Dorado.(2006). Artificial Neural Networks in Real - Life Applications. Published in the United States of America by Idea Group Publishing (an imprint of Idea Group Inc.)

Kanuka, H., Collett, D. & Caswell, C. (2002). University instructor perceptions of the use of asynchronous text-based discussion in distance courses. American Journal of Distance Education, 16(3), 151-167. Retrieved February 21, 2009, from

http://www.informaworld.com.libproxy.library.wmich.edu/10.1207/S15389286AJDE1603\_3

Kaur, A.(2004). The effectiveness of online discussion forum at OUM. Research Paper funded by OUM. Kuala Lumpur: Open University Malaysia.

Kearsley, G.(2002). Is online learning for everybody? Education Technology, 42(1), 41 - 44

Keller, J. (1987). Development and use of the ARCS model of instructional design. Journal of

Instructional Development, 10(3), 2-10.

Kember, D.(1995). Open learning courses for adults. Englewood Cliffs, NJ: Educational Technology.

Kember, D. (1989). A Longitudinal - Process Model of Drop - out from Distance Education. Higher Education. 60, 3.

Kraut, R., Patterson, M., Lundmark, V., Kiesler, S., Mukopadhyay, T., & Scherlis, W. (1998). Internet paradox: A social technology that reduces social involvement and psychological well-being? A roundbreaking study of families, empirically examines the social effects of home Internet use. American Psychologist, 53(9), 1017-1031. Retrieved March 29, 2009, from http://www.cs.cmu.edu/~kraut/RKraut.site.files/articles/kraut98-InternetParadox.pdf

L.P.J. Veelenturf.(1995). Analysis and Applications of Artificial Neural Networks. Prentice Hall

Lakhmi Jain, Ph.D. Anna Maria Fanelli, Ph.D.(2000). Recent Advances In Artificial Neural Networks Design and Applications. CRC Press LLC.

Latif, L. A., Sungsri, S., & Bahroom, R.(2009). Managing retention in ODL institutions: A case study on Open University Malaysia and Sukhothai Thammathirat Open University. ASEAN Journal of Open and Distance Education. 1(1).

Lee, H. (2008). The determinants of the effectiveness of online discussion board systems in eLearning: A case study. Lecture Notes in Computer Science, 5288 LNAI, 271.

Levy, Y.(2007). Comparing dropouts and persistence in e-learning courses. Computers & Education, 48(2), 185-204.

Lewis, T. (1999). Research in technology education-some areas of need. Journal of Technology Education, 10(2), 41-56.

Liaw, S. (2008). Investigating students’ perceived satisfaction, behavioral intention, and effectiveness of e-learning: A case study of the blackboard system. Computers & Education, 51(2), 864-873.

Max Bramer.(2007). Principles of Data Mining. Digital Professor of Information Technology, University of Portsmouth, UK.

McEwen, B. C. (2001). Web-assisted in online learning. Business Communication Quarterly: A Publication of the Association for Business Communication, 62(2), 98-103.

Meyer, J. P. & Allen, N. J. (1991). A three-component conceptualisation of organisational

commitment. Human Resource Management Review. 1, 61 - 69

Michigan Community College Virtual Learning Collaborative.(2009). Is online learning right for me? Retrieved March 29, 2009, from

http://vcampus.mccvlc.org/index.asp?dir='welcome'&content='YouReady.asp

Moore, M. (1993). Three types of interaction. Routledge, London: K. Harry, M. Holm and D. Keegan.

Morgan, C. K., & Tam, M. (1999). Unraveling the complexities of distance education student attrition. Distance Education, 20(1), 96 – 108

Nikola K. Kasabov.(1998). Foundations of Neural Networks, Fuzzy Systems, and Knowledge ngineering. A Bradford Book The MIT Press Cambridge, Massachusetts, London, England

Olgren, C. (2004). Reasons for attrition. Distance education certificate program, University of isconsin-Madison. Retrieved: December 3, 2007 from http://uwmad.courses.wisconsin.edu/d2l/tools/files/accessFile.asp?file-6847687&code=641915067

P.-N.Tan, M.Steinbach, and V.Kumar.(2006). Introduction to datamining.Boston: Pearson Addison-Wesley.

Palloff, R. M., & Pratt, K. (1999). Building learning communities in cyberspace. San Francisco: Jossey-Bass.

Park, J.-H., & Choi, H. J. (2009). Factors Influencing Adult Learners' Decision to Drop Out or Persist in Online Learning. Educational Technology & Society, 12 (4), 207–217.

Parker, A. (1999, December). A study of variables that predict dropout from distance education. International Journal of Educational Technology, 1(2). Retrieved March 29, 2009.

Parsad, B., & Lewis, L. (2008). Distance education at degree-granting postsecondary institutions: 2006–07 (NCES 2009–044). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. Retrieved January 18, 2009.

Pascarella, E. T. & Terenzini, P. T. (1991). How college affects students. San Francisco: Jossey-Bass

Picciano, A. (2002). Beyond student perceptions: issues of interaction, presence, and performance in an online course. Journal of Asynchronous Learning Networks, 6(1), 21-40. Retrieved February 15, 2009, from

http://www.aln.org/publications/jaln/v6n1/pdf/v6n1\_picciano.pdf

Prakash V. Arumugam.Flight Risk: Identifying Potential Drop outs. School of Business and ministration, Wawasan Open University 54, Jalan Sultan Ahmad Shah, 10050 Georgetown, Penang Malaysia

Reisman, S.(2003). Electronic learning communities: Issues and best practices. Greenwich, CT: Information Age Publishing. In Books24x7 (OUM Digital Collection)

Richard Ng. Improving Graduation Rates of Open and Distance Learners via Online Community. Open University Malaysia 1, Jalan Lim Bo Seng, 30300 Ipoh, Perak, Malaysia.

Robertson, J., Fuller, M., Midon, N., Smith, B., Sadera, W. & Song, L. (2008). Relationships between community and student success in online learning. In K. McFerrin et al. (Eds.), Proceedings of Society for Information Technology & Teacher Education International Conference. Chesapeake, VA: AACE. Retrieved: May 10, 2010 from http://www.editlib.org/p/27246.

Rovai, A. P., Lucking, R., & Cristol, D. (2001). Building classroom community at a distance. Paper presented at the American Educational Research Association Annual Meeting, Seattle, Washington.

Sajadin Sembiring, M. Zarlis, Dedy Hartama, Ramliana S, Elvi Wani .(2011). Pridiction of student academic performance by an application of data mining techniques. 2011 International Conference on Management and Artificial Intelligence IPEDR vol.6 (2011) © (2011) IACSIT Press, Bali, Indonesia

Seidman, A. (Ed.). (2005). College student retention: Formula for student success. Westport, CT: ACE/Praeger

Sharma, P., & Maleyeff, J. (2003). Internet education: Potential problems and solutions. The International Journal of Educational Management, 17(1), 19.

Smith, G. G., & Ferguson, D. (2002). Teaching over the web versus in the classroom: Differences in the instructor experience. International Journal of Instructional Media, 29(1), 61 (68 pages).

Smith, P. C., Kendall, L. M., & Hulin, C. L.(1969). The measurement of satisfaction in work and retirement. Chicago: Rand McNally.

Stamos T. Karamouzis and Andreas Vrettos.(2008). Proceedings of the World Congress on Engineering and Computer Science 2008 WCECS 2008, October 22 - 24, 2008, San Francisco, USA

Stearns E. and Glennie E. J. (2006). When and Why Dropouts Leave High School

Stratton L. S., O’Toole D. M. and Wetzel J. N. (2005). A Multinomial Logit Model of College Stopout and Dropout Behavior. Discussion Paper No. 1634 June 2005 The Institute for the Study of Labor, Bonn.

http://www.kent.edu/rcet/Publications/upload/learning%20effectiveness4.pdf

Tinto V. (1975). Dropout from Higher Education: A Theoretical Synthesis of Recent Research

Review of Educational Research, Vol. 45, No. 1 (Winter, 1975), pp. 89-125

Tinto, V. (1982). Limits of theory and practice in student attrition. Journal of Higher Education, 53(6), 687-700.

Tinto, V. (1993). Leaving College: Rethinking the Causes and Cures of Student Attrition (2nd Ed.). Chicago: University of Chicago Press.

Tyler-Smith, K. (2006) Early attrition among first-time elearners: A review factors that contribute to dropout, withdrawal and non-completion rates of adult learners undertaking e-learning programmes. MERLOT Journal Online Learning and Teaching, 2(2), 73-85.

Verduin, J., & Clark, T.(1991). Distance education: The foundations of effective practice. San Francisco, CA: Jossey-Bass.

# PHỤ LỤC 1: CÁC BÀI TOÁN THƯỜNG GẶP TRONG KHAI PHÁ DỮ LIỆU

## Bài toán phân lớp

### Khái niệm phân lớp

Cho một tập các bản ghi (gọi là tập huấn luyện): mỗi bản ghi chứa một tập các thuộc tính trong đó có một thuộc tính phân lớp (class attribute)

Tìm/học một hàm cho thuộc tính phân lớp (hàm phân lớp) đối với các giá trị của các thuộc tính khác.

Tập kiểm thử: được sử dụng một tập các ví dụ khác với các ví dụ học để kiểm tra độ chính xác của hàm phân lớp học. Thông thường tập dữ liệu ban đầu chia thành hai tập không giao nhau: một tập để học hàm phân lớp, một tập để kiểm thử hàm phân lớp học được.

### Các kĩ thuật phân lớp:

* Mạng Noron
* Mạng Bayes
* Cây quyết định
* Mô hình Markov ẩn
* Luật cơ sở
* Trí nhớ dựa trên các nguyên nhân
* Giải thuật di truyền
* SVM
* …

## Tiền xử lý dữ liệu

Những tác vụ chính trong tiền xử lý dữ liệu

**Làm sạch dữ liệu (Data cleaning)**: Gán các giá trị thuộc tính còn thiếu, sửa chữa các dữ liệu nhiễu/lỗi, xác định hoặc loại bỏ các ngoại lai, giải quyết mâu thuẫn dữ liệu.

**Tích hợp dữ liệu (Data integration)**: Tích hợp nhiều cơ sở dữ liệu, nhiều khối dữ liệu (data cubes) hoặc nhiều tập tin dữ liệu.

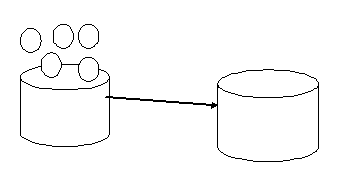
**Biến đổi dữ liệu (Data transformation)**: Chuẩn hóa (normalize) và kết hợp (aggregate) dữ liệu.

**Giảm bớt dữ liệu (Data reduction)** :Giảm bớt về biểu diễn (các thuộc tính) của dữ liệu, giảm bớt kích thước dữ liệu nhưng vẫn đảm bảo các kết quả của khai phá dữ liệu tương đương hoặc xấp xỉ.

**Rời rạc hóa dữ liệu (Data discretization)**: Là một thao tác trong giảm bớt dữ liệu, được sử dụng với các thuộc tính kiểu số.

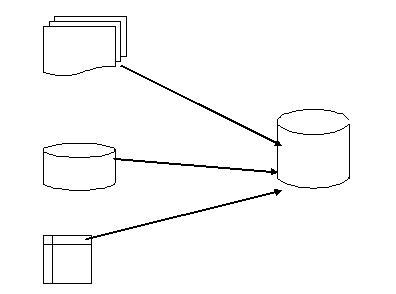
Sau đây là một số hình ảnh minh họa của quá trình tiền xử lý:

Làm sạch dữ liệu:



##### Làm sạch dữ liệu

Tích hợp dữ liệu:

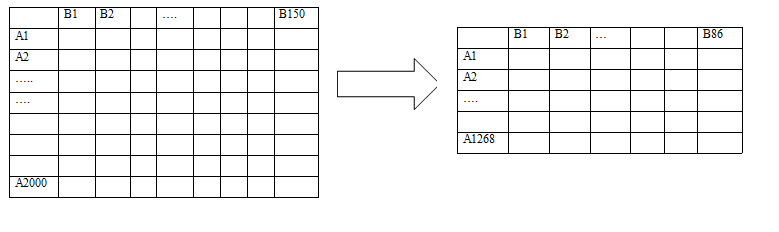


##### Tích hợp dữ liệu

Chuyển đổi dữ liệu:

-2, -6, 100, 2, 6 -0.02, - 0.06, 1.0, 0.02, 0 06

Giảm bớt dữ liệu:



##### Giảm bớt dữ liệu

### Làm sạch dữ liệu (Data cleaning)

Dữ liệu trong thực tế thu được có thể chứa nhiều lỗi, không hoàn chỉnh, có mâu thuẫn.

Không hoàn chỉnh (incomplete): Thiếu các giá trị thuộc tính hoặc thiếu một số thuộc tính. Ví dụ: bằng cấp 3=” ”.

Nhiễu/lỗi (noise/error): chứa đựng các lỗi hoặc các ví dụ bất thường (abnormal instances). Ví dụ: bằng cấp 3 = {k, khá, kha, khả}.

Mâu thuẫn (inconsistent): Chứa đựng các mâu thuẫn không thống nhất. Ví dụ: bằng cấp 3 = {0, 89} (không phù hợp với dữ liệu)

Câu hỏi đặt ra là tại sao phải làm sạch dữ liệu?

*Trả lời:*

Nếu dữ liệu không sạch (có chứa lỗi, nhiễu, không đầy đủ, có mâu thuẫn) thì các kết quả khai phá dữ liệu sẽ bị ảnh hưởng và không đáng tin cậy.

Các kết quả khai phá dữ liệu (các tri thức khám phá được) không chính xác (không đáng tin cậy) sẽ dẫn đến các quyết định không chính xác, không tối ưu.

Giải pháp cho các vấn đề:

* Thuộc tính thiếu giá trị:

Bỏ qua các bản ghi có các thuộc tính thiếu giá trị: Thường được áp dụng trong bài toán phân lớp (classification). Một số người đảm nhiệm việc kiểm tra và gán các giá trị thuộc tính còn thiếu: công việc tẻ nhạt và chi phí cao

Gán giá trị tự động bởi máy tính:

* Một giá trị hằng mặc định
* Giá trị trung bình của thuộc tính đó
* Dữ liệu chứa nhiễu

Phân khoảng (Binning)

* Cách 1: Phân chia với độ rộng (khoảng cách) bằng nhau

Chia khoảng giá trị thành N khoảng với kích thước (độ rộng) bằng nhau. Nếu min và max là giá trị lớn nhất và nhỏ nhất của thuộc tính thì kích thước (độ rộng) của mỗi khoảng = (max-min) /N.

Không phù hợp với các tập dữ liệu lệch hoặc có chứa các ngoại lai (outliers) vì có thể một khoảng sẽ chỉ chứa một hoặc một số các ngoại lai.

* Cách 2: phân chia với độ sâu (tần xuất xuất hiện) bằng nhau.

Chia khoảng giá trị thành N khoảng (không nhất thiết bằng nhau), sao cho mỗi khoảng chứa xấp xỉ bằng nhau số lượng (tần xuất xuất hiện) của các ví dụ.

Hiệu quả hơn cách phân chia với độ rộng (khoảng cách) bằng nhau.

**Ví dụ:** Thuộc tính tuổi: 12, 13, 14, 15, 16, 24, 25, 26, 27, 23, 31, 32, 33, 36, 38.

Bin 1: 12, 13, 14, 15, 16

Bin 2: 24, 25, 26, 27, 23

Bin 3: 31, 32, 33, 36, 38

Biểu diễn khoảng dữ liệu bởi giá trị trung bình:

Bin 1:14, 14, 14, 14, 14

Bin 2:25, 25, 25, 25, 25

Bin 3:34, 34, 34, 34, 34

### Hồi quy (Regression)

Phân cụm (Clustering)

Kết hợp giữa máy tính và kiểm tra của con người

### Tích hợp dữ liệu (Data integration)

Là việc kết hợp dữ liệu từ nhiều nguồn vào một kho dữ liệu thống nhất. Yêu cầu chung là: giảm thiểu các dư thừa và các mâu thuẫn.

Các vấn đề:

* Xác định các thực thể (để tránh dư thừa dữ liệu):
* Phát hiện và xử lý mâu thuẫn đối với giá trị dữ liệu: đối với cùng một thực thể trên thực tế nhưng các giá trị thuộc tính từ nhiều nguồn khác nhau lại khác nhau (Do cách biểu diễn khác, độ đo khác...)
* Xử lý dư thừa dữ liệu (redundant data): thường xuyên xảy ra khi tích hợp dữ liệu từ nhiều nguồn (ví dụ: Nhiều CSDL)
* Dư thừa: định danh đối tượng (cùng một thuộc tính có thể mang các tên khác nhau trong các cơ sở dữ liệu khác nhau), dữ liệu suy ra được (một thuộc tính trong bảng có thể là một thuộc tính được suy ra trong một bảng khác). Các thuộc tính dư thừa có thể được phát hiện bằng phân tích tương quan.

Quá trình tích hợp dữ liệu giúp cải thiện tốc độ của quá trình khai phá dữ liệu và nâng cao chất lượng của các kết quả (tri thức) thu được.

### Biến đổi dữ liệu (Data transformation)

Là việc ánh xạ toàn bộ tập giá trị của một thuộc tính sang một tập mới các giá trị thay thế sao cho mỗi giá trị cũ tương ứng với một trong các giá trị mới.

Các phương pháp biến đổi dữ liệu:

* Làm trơn (Smoothing): loạibỏ nhiễu/lỗi khỏi dữ liệu
* Kết hợp (Aggregation): sự tóm tắt dữ liệu, xây dựng các khối dữ liệu (data cubes)
* Khái quát hóa (Generalization): xây dựng các phân cấp khái niệm (Concept hierarchies)
* Chuẩn hóa (Normalization): đưa giá trị về một khoảng được chỉ định.
* Chuẩn hóa min-max: thành khoảng



* Chuẩn hóa z-score: là giá trị trung bình và độ lệch chuẩn đối với thuộc tính i:
* Chuẩn hóa bởi thang chia 10: j là số nguyên nhỏ nhất sao cho:

Max ({}) <1

* Xây dựng (tạo nên) các thuộc tính mới dựa trên các thuộc tính ban đầu.
* Giảm bớt dữ liệu (Data reduction) Mục tiêu: thu được một biểu diễn thu gọn nhưng vẫn sinh ra cùng các kết quả phân tích (khai phá) như với tập dữ liệu ban đầu.

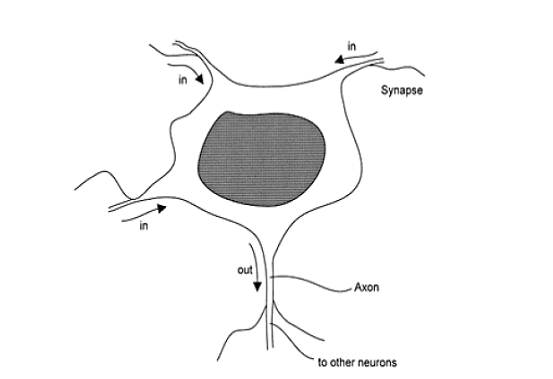
Các chiến lược giảm bớt dữ liệu:

* Giảm số chiều (Dimensionality reduction): loại bỏ bớt các thuộc tính ít quan trọng
* Kĩ thuật: Phân tích thành phần chính (Pricipal component analysis) và lựa chọn tập con các thuộc tính (Feature subset selection)
* Giảm lượng dữ liệu (Data/numerosity reduction)

## Lịch sử, nguồn gốc của mạng Nơron nhân tạo

### Vài nét về hoạt động của não con người

Qua quá trình nghiên cứu về bộ não, người ta thấy rằng: bộ não con người bao gồm khoảng 1011 nơron tham gia vào khoảng 1015 kết nối trên các đường truyền . Mỗi đường truyền này dài khoảng hơn một mét. Các nơron có nhiều đặc điểm chung với các tế bào khác trong cơ thể, ngoài ra chúng còn có những khả năng mà các tế bào khác không có được. Đó là khả năng nhận, xử lý và truyền các tín hiệu điện hóa trên các đường mòn nơron. Các con đường này tạo nên hệ thống giao tiếp của bộ não.



##### Mô hình Nơron sinh học

Mỗi nơron sinh học có 3 thành phần cơ bản:

• Các nhánh vào hình cây ( dendrites)

• Thân tế bào ( cell body )

• Sợi trục ra ( axon )

Các nhánh hình cây truyền tín hiệu vào đến thân tế bào. Thân tế bào tổng hợp và xử lý cho tín hiệu đi ra. Sợi trục truyền tín hiệu ra từ thân tế bào này sang nơron khác. Điểm liên kết giữa sợi trục của nơron này với nhánh hình cây của nơron khác gọi là synapse. Liên kết giữa các nơron và độ nhạy của mỗi synapse được xác định bởi quá trình hóa học phức tạp. Một số cấu trúc của nơron được xác định trước lúc sinh ra. Một số cấu trúc được phát triển thông qua quá trình học. Trong cuộc đời cá thể, một số liên kết mới được hình thành, một số khác bị hủy bỏ. Như vậy nơron sinh học hoạt động theo cách thức sau: nhận tín hiệu đầu vào, xử lý các tín hiệu này và cho ra một tín hiệu output. Tín hiệu output này sau đó được truyền đi làm tín hiệu đầu vào cho các nơron khác.

Có thể tóm tắt hoạt động của một Nơron như sau: Nơron lấy tổng tất cả các điện thế vào mà nó nhận được, và phát ra một xung điện thế nếu tổng ấy lớn hơn một ngưỡng nào đó. Các Nơron nối với nhau ở các synapses. Synapse được gọi là mạch khi nó cho phép truyền dẫn dễ dàng tín hiệu qua các Nơron khác. Ngược lại, một synapse yếu sẽ truyền dẫn tín hiệu rất khó khăn.

Các synapses đóng vai trò rất quan trọng trong sự học tập. Khi chúng ta học tập thì hoạt động của các synapses được tăng cường, tạo nên nhiều liên kết mạnh giữa các Nơron. Có thể nói rằng người nào học càng giỏi thì càng có nhiều synapses và các synapses ấy càng mạnh mẽ, hay nói cách khác, thì liên kết giữa các Nơron càng nhiều, càng nhạy bén.

### Nguồn gốc, ý tưởng của mạng Nơron

Trên cơ sở mạng nơron thần kinh, chúng ta có thể xây dựng mô hình mạng nơ ron nhân tạo mô phỏng hoạt động của mạng nơron thần kinh.

Mô phỏng hoạt động của các nơ ron thần kinh, mạng nơ ron nhân tạo là hệ thống bao gồm nhiều phần tử xử lý đơn giản (Nơron) hoạt động song song. Tính năng của hệ thống này tuỳ thuộc vào cấu trúc của hệ, các trọng số liên kết nơron và quá trình tính toán tại các nơron đơn lẻ. Mạng nơron có thể từ dữ liệu mẫu và tổng quát hoá dựa trên các dữ liệu mẫu học.

Mỗi liên kết gắn với một trọng số, trọng số này được thêm vào trong quá trình tín hiệu đi qua liên kết đó. Các trọng số có thể dương, thể hiện trạng thái kích thích, hay âm, thể hiện trạng thái kiềm chế. Mỗi nơron tính toán mức kích hoạt của chúng bằng cách cộng tổng các đầu vào và đưa ra hàm chuyển. Một khi đầu ra của tất cả các nơron trong một lớp mạng cụ thể đã thực hiện xong tính toán thì lớp kế tiếp có thể bắt đầu thực hiện tính toán của mình bởi vì đầu ra của lớp hiện tại tạo ra đầu vào của lớp kế tiếp. Khi tất cả các nơron đã thực hiện tính toán thì kết quả được trả lại bởi các nơron đầu ra. Tuy nhiên, có thể là chưa đúng yêu cầu, khi đó một thuật toán huấn luyện cần được áp dụng để điều chỉnh các tham số của mạng.

### Lịch sử phát triển của mạng Nơron

Mạng nơron nhân tạo đã có một lịch sử lâu dài. Từ năm 1943 McCulloch và Pitts đã đưa ra khả năng liên kết và một số liên kết cơ bản của mạng nơron. Năm 1958, Rosenblatt đã đưa ra cấu trúc Perceptron. Năm 1969 Minsky và Papert phân tích sự đúng đắn của Perceptron, họ đã chứng minh các tính chất và chỉ rõ sự hạn chế của một số mô hình. Những năm 1967 Grossberg dựa vào tính chất sinh học đã đưa ra một số cấu trúc của hệ động học phi tuyến với các tính chất mới.

Những năm 1970 mạng nơron rơi vào trạng thái im lặng cả về nghiên cứu và ứng dụng. Tuy vậy có thể kể đến nghiên cứu về quá trình học của mạng nhiều lớp, mạng tự tổ chức của Kohonen, mạng kiểu bộ nhớ kết hợp (BAM) cuả Anderson, mạng ART (Adaptive Resonance Theory Neural Networks) của Capenter

Năm 1982 Hoppfield đã đưa ra mạng một lớp qui hồi Hoppfield. Thuật toán lan truyền ngược sai số (Error BackPropagation) được Rumelhart, Hinton, Williams đề xuất có thể huấn luyện mạng nơron nhiều lớp vào năm 1986.

Những năm gần đây, nhiều chương trình nghiên cứu và ứng dụng của mạng nơron trong nhiều lĩnh vực sản xuất nhau như: xử lí điều khiển và lập kế hoạch, điều khiển chất lượng, bảo dưỡng sữa chữa và điều khiển robot...

## Khái niệm và cấu trúc của một Nơron nhân tạo

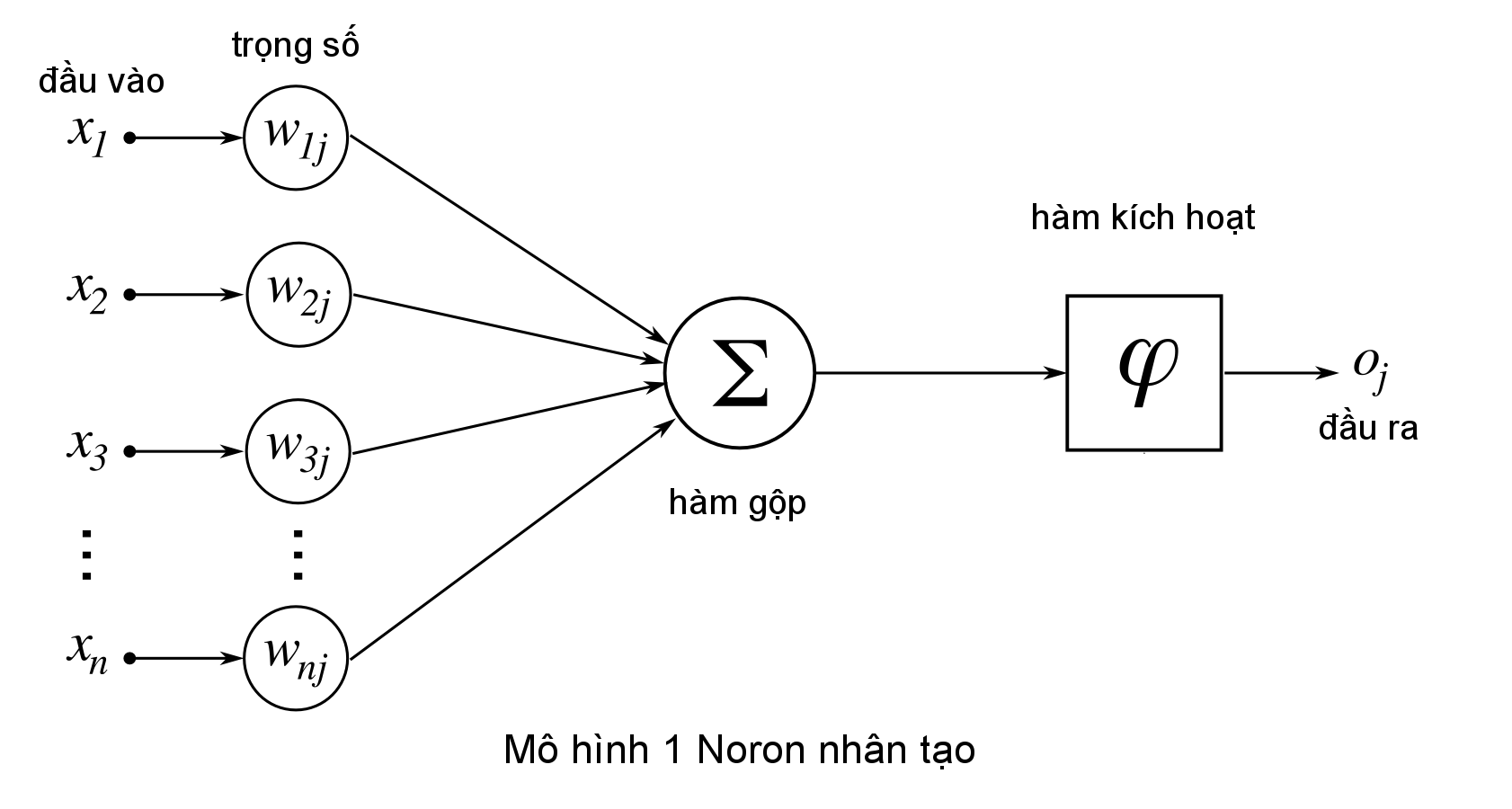
### Mô hình một Nơron nhân tạo

Để mô phỏng các tế bào thần kinh và các khớp nối thần kinh của bộ não người, mạng Nơron nhân tạo cũng có các thành phần có vai trò tương tự là các Nơron nhân tạo và các kết nối.

#### Khái niệm 1.1. Nơron nhân tạo

*Một Nơron nhân tạo là một đơn vị xử lý thông tin cơ sở cho hoạt động của một mạng Nơron. Nó là một đơn vị tính toán có nhiều đầu vào và một đầu ra, mỗi đầu vào đến từ một liên kết. Đặc trưng của Nơron là một hàm kích hoạt phi tuyến chuyển đổi tổ hợp tuyến tính của tất cả các tín hiệu đầu vào thành tín hiệu đầu ra. Hàm kích hoạt này đảm bảo tính chất phi tuyến cho tính toán của mạng Nơron.*

Dưới đây là mô hình một Nơron nhân tạo:



##### Mô hình một Nơron nhân tạo

Một Nơron được cấu tạo gồm các thành phần chính: liên kết Nơron, bộ cộng, hàm kích hoạt.

### Liên kết Nơron

Liên kết Nơron là một thành phần của mạng Nơron nhân tạo để liên kết giữa các Nơron, nó nối đầu ra của Nơron lớp này với đầu vào của một Nơron trong lớp khác. Đặc trưng của thành phần liên kết là một trọng số mà mỗi tín hiệu đi qua đều được nhân với trọng số này. Các trọng số liên kết chính là các tham số tự do cơ bản của mạng neuron, có thể thay đổi được nhằm thích nghi với môi trường xung quanh.

Có thể xem các trọng số là phương tiện để lưu trữ thông tin dài hạn trong mạng Nơron và nhiệm vụ của quá trình huấn luyện của mạng là cập nhật các trọng số khi có thêm thông tin về mẫu học. Hay nói một cách khác, các trọng số đều được điều chỉnh sao cho dáng điệu vào ra của mạng sẽ mô phỏng hoàn toàn phù hợp với môi trường đang xem xét.

### Hàm gộp

Bộ cộng dùng để tính tổng các tín hiệu đầu vào của Nơron, đã được nhân với các trọng số liên kết tương ứng. phép toán được mô tả ở đây tạo nên một bộ hợp tuyến tính.

|  |  |  |
| --- | --- | --- |
| **STT** | **Tên hàm** | **Công thức tính** |
|  | Hàm tuyến tính có ngưỡng |  |
|  | Hàm dạng toàn phương |  |
|  | Hàm đa thức |  |

***Bảng 1.1****. Một số hàm gộp thông dụng*

### Hàm kích hoạt

Hàm kích hoạt hay còn gọi hàm kích hoạt phi tuyến, hàm truyền, chuyển đổi một tổ hợp tuyến tính của tất cả các tín hiệu đầu vào thành tín hiệu đầu ra. Hàm kích hoạt này đảm bảo tính chất phi tuyến cho tính toán mạng Nơron. Nó được xem như là một hàm giới hạn, nó giới hạn phạm vi biên độ cho phép của tín hiệu đầu ra trong một khoảng giá trị hữu hạn hay nói cách khác: hàm kích hoạt có tác dụng “nén” giá trị tổng vào một miền giới hạn nào đó.

Lựa chọn hàm kích hoạt : tùy theo khoảng cần thiết của giá trị kết xuất mà ta lựa chọn hàm kích hoạt phù hợp. Sau đây là một số hàm kích hoạt thông dụng:

|  |  |  |
| --- | --- | --- |
| **STT** | **Tên hàm** | **Công thức tính** |
|  | Hàm nhảy(hardlim) |  |
|  | Hàm ngưỡng(hardlims) |  |
|  | Hàm dốc(Satlin) |  |
|  | Hàm Sigmoidal 1 cực |  |
|  | Hàm Sigmoidal 2 cực |  |
|  | Hàm Gauss |  |
|  | Hàm Logistic |  |
|  | Purelin | a = n |
|  | Satlins | a = -1 với n < 0  a = n với 0 ≤ n ≤ 1  a = 1 với n > 1 |
|  | Tansig |  |
|  | Poslin | a = 0 với n < 0  a = n với n ≥ 0 |
|  | Compet | a = 1 với Nơron có n lớn nhất  a = 0 với các Nơron còn lại |

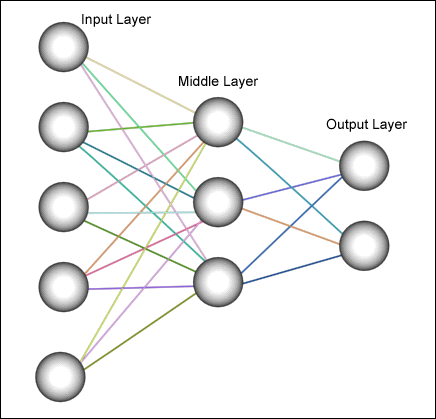
***Bảng 2.2****. Một số hàm kích hoạt thông dụng*

## Mô hình mạng Nơron nhân tạo

### Khái niệm mạng Nơron

Mạng Nơron nhân tạo là một mô hình toán học được xây dựng dựa trên mo phỏng quá trình hoạt động và tính chất của các mạng Nơron sinh học. Nó gồm một nhóm các Nơron nhân tạo (nút) nối với nhau, và xử lý thông tin bằng cách truyền theo các kết nối và tính giá trị mới tại các nút. Trong nhiều trường hợp, mạng Nơron nhân tạo là một hệ thống thích ứng tự thay đổi cấu trúc của mình dựa trên các thôngt in bên ngoài hay bên trong chảy qua mạng trong quá trình học.

**Mô hình mạng Nơron**:

****

***Hình 1.3****. Sơ đồ đơn giản về một mạng Nơron nhân tạo*

Mô hình mạng Nơron ở trên gồm 3 lớp: lớp nhập (input), lớp ẩn(hidden) và lớp xuất (output). Mỗi nút trong lớp nhập nhận giá trị của một biến độc lập và chuyển vào mạng.

Dữ liệu từ tất cả các nút trong lớp nhập được tích hợp ­– ta gọi là tổng trọng số – và chuyển kết quả cho các nút trong lớp ẩn. Gọi là “ẩn” vì các nút trong lớp này chỉ liên lạc với các nút trong lớp nhập và lớp xuất, và chỉ có người thiết kế mạng mới biết lớp này (người sử dụng không biết lớp này).

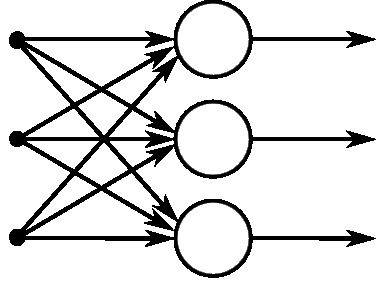
Các nút trong lớp xuất nhận các tín hiệu tổng trọng hóa từ các nút trong lớp ẩn. Mỗi nút trong lớp xuất tương ứng với một biến phụ thuộc.

### Phân loại mạng Nơron

#### Phân loại theo số lượng, số lớp nơron trong mạng

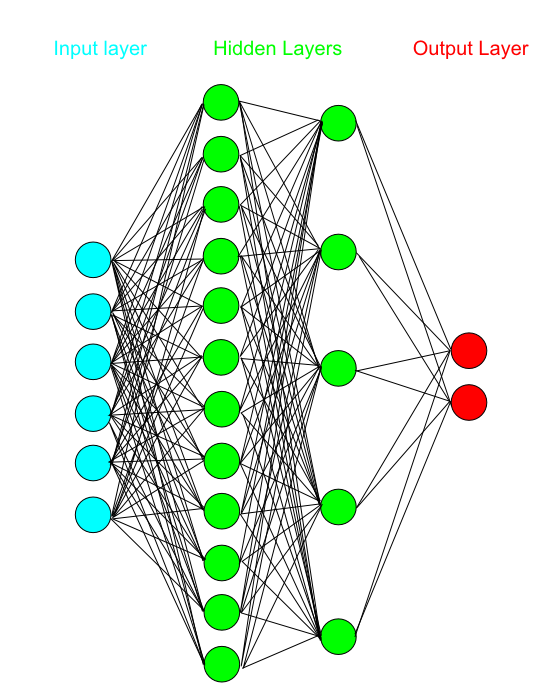
Trong mạng Nơron các Nơron trong cùng một lớp sẽ có cùng chức năng. Dựa theo số lớp Nơron trong mạng mà người ta chia mạng Nơron thành hai loại:

a. Mạng một lớp (Single Layer Neural Network) : loại mạng này chỉ có một lớp vừa đóng vai trò lớp vào, vừa đóng vai trò lớp ra. Mạng này có tính chất tuyến tính nên còn được gọi là mạng tuyến tính.



##### Mô hình mạng Nơron một lớp

b. Mạng nhiều lớp (MLP: Multi Layer Perceptrons Network): Ngược lại với mạng một lớp là mạng nhiều lớp.



##### Mô hình mạng Nơron nhiều lớp

- Lớp vào: là lớp tiếp xúc trực tiếp với thông tin vào, có chức năng thu thập thông tin, không tham gia trực tiếp vào quá trình huấn luyện mạng. Mỗi mạng Nơron chỉ có một lớp vào.

- Lớp ra : có chức năng tổng hợp thông tin và đưa ra kết xuất đầu ra, tham gia trực tiếp vào quá trình luyện mạng. Mỗi mạng Nơron chỉ có một lớp ra.

- Lớp ẩn: là các lớp còn lại, tham gia trực tiếp vào quá trình luyện mạng. Mạng Nơron có thể có nhiều lớp ẩn.

**Khả năng thể hiện**

Các mạng truyền thẳng cho ta một kiến trúc tổng quát thể hiện khả năng ánh xạ hàm phi truyến giữa một tập các biến đầu vào và một tập các đầu ra. Khả năng thể hiện của một mạng có thể được định nghĩa là khoảng mà nó có thể thực hiện ánh xạ khi các trọng số biến thiên.

- Mạng một lớp chỉ có khả năng thể hiện các hàm có khả năng phân tuyến tính hay các miền phân chia được (ví dụ như hàm logic AND có miền giá trị có thể phân chia được bằng một đường thẳng, trong khi miền giá trị của hàm XOR thì không).

- Mạng có hai lớp ẩn có khả năng thể hiện một đường biên phân chia tùy ý với một độ chính xác bất kỳ với các hàm chuyển phân ngưỡng và có thể xấp xỉ bất kỳ ánh xạ mịn nào với độ chính xác bất kỳ với các hàm chuyển có dạng Sigmoid.

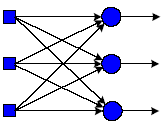
#### Phân loại theo kiểu liên kết Nơron

a. Mạng truyền thẳng (Feelforward)

Có thể nói mạng Nơron truyền thẳng là một kiểu mạng đơn giản trong việc sắp đặt mạng. Trong mạng này thông tin chỉ truyền trên một hướng duy nhất. từ lớp đầu vào xuyên qua lớp ẩn (nếu có) và kết thúc tại lớp đầu ra, không có chu trình hoặc vòng trong mạng.

**Các mạng truyền thẳng đơn mức**

Trong một mạng Nơron phân mức, các Nơron được tổ chức dưới dạng các mức. Với dạng đơn giản nhất của mạng phân mức, chúng ta có một mức đầu vào gồm các nút nguồn chiếu trực tiếp tới mức đầu ra gồm các Nơron.



Mức đầu vào Mức đầu ra

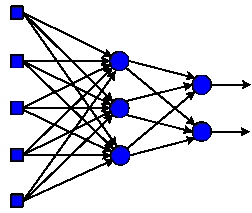
##### Mô hình mạng lan truyền tiến với một mức Nơron

Như vậy, mạng thực sự là không có chu trình. Nó được minh hoạ trong hình 1.6 cho trường hợp ba nút đối với cả mức đầu ra và đầu vào. Một mạng như vậy được gọi là một mạng đơn mức. “Đơn mức” tức là chỉ có một mức, chính là mức đầu ra gồm các nút tính toán (các Nơron). Chúng ta không tính mức đầu vào của các nút nguồn vì không có tính toán nào được thực hiện ở đây.

**Các mạng truyền thẳng đa mức**

Lớp thứ hai của một mạng Nơron dẫn tiến được phân biệt bởi sự có mặt của một hay nhiều mức ẩn, mà các nút tính toán của chúng được gọi là các Nơron ẩn hay các đơn vị ẩn (thuật ngữ “ẩn” ở đây mang ý nghĩa là không tiếp xúc với môi trường). Chức năng của các Nơron ẩn là can thiệp vào giữa đầu vào và đầu ra của mạng một cách hữu hiệu. Bằng việc thêm một vài mức ẩn, mạng có khả năng rút ra được các thống kê bậc cao của tín hiệu đầu vào. Khả năng các Nơron ẩn rút ra được các thống kê bậc cao đặc biệt có giá trị khi mức đầu vào có kích thước lớn.

Mạng Nơron trong hình 1.7 được gọi là kết nối đầy đủ với ý nghĩa là tất cả các nút trong mỗi mức của mạng được nối với tất cả các nút trong mức tiếp sau. Nếu một số kết nối không tồn tại trong mạng, chúng ta nói rằng mạng là kết nối không đầy đủ.



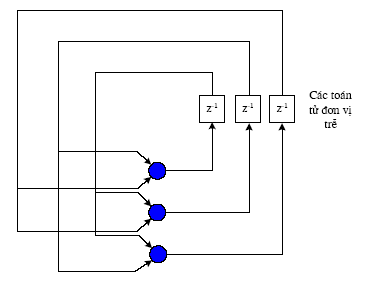
##### Mạng lan truyền tiến kết nối đầy đủ

Trong mạng truyền thẳng các Nơron đi theo một hướng nhất định tạo thành đồ thị không có chu trình, các đỉnh là các Nơron còn các cạnh là các liên kết giữa chúng.

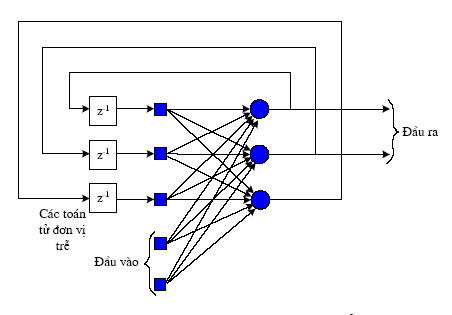
Một số loại mạng truyền thẳng: Functional link network, WTA architecture, Cascade Correlation Architechture, Radial Basis Function nertwork, Feedforward version of the Counterpropagation network.

b. Mạng qui hồi (Feedback)

Trái với mạng Nơron truyền thẳng , mạng nơron qui hồi là những mô hình với hai luồng dữ liệu có hướng. Trong khi mạng truyền thẳng truyền dữ liệu theo một đường thẳng thì những mạng Nơron qui hồi có ít nhất một phản hồi từ những Nơron xử lý sau quay trở lại các Nơron xử lý trước đó.



##### Mạng hồi quy không có Nơron ẩn và không có vòng lặp tự phản hồi



##### Mạng qui hồi có các Nơron ẩn

Các mạng qui hồi cho phép các liên kết Nơron tạo thành chu trình, các thông tin ra của các Nơron được truyền lại cho các Nơron đã góp phần kích hoạt chúng, nên mạng qui hồi còn có khả năng lưu giữ trạng thái trong của nó dưới dạng các ngưỡng kích hoạt ngoài các trọng số liên kết Nơron.

Một số loại mạng hồi qui: Hopfield network, Autoassociative memory, Bidirectional Associative Memories (BAM).

## Ứng dụng của mạng Nơron

### Ưu, nhược điểm của mạng Nơron

#### Ưu điểm của mạng Nơron

#### a. Khả năng xử lý song song và phân tán.

Đặc điểm nổi bật của mạng Nơron là khả năng phát triển các mô hình xử lý phân tán song song. Các mô hình này giả định rằng việc xử lý thông tin được tiến hành thông qua sự tương tác của một số lượng lớn các Nơron, mỗi Nơron gửi các tín hiệu kích thích hay ức chế tới các Nơron khác trong mạng. Bản chất song song đồ sộ của một mạng Nơron làm cho nó rất nhanh trong tính toán đối với một số công việc.

#### b. Khả năng thích nghi và tự tổ chức.

Một mạng Nơron có thể biến đổi các trọng số tùy theo sự thay đổi của môi trường xung quanh. Khi một mạng Nơron đã được tích lũy để hoạt động trong một môi trường xác định có thể được tích lũy lại một cách dễ dàng khi có những thay đổi nhỏ của các điều kiện môi trường hoạt động. Mặt khác, khi hoạt động trong một môi trường không ổn định (các số liệu thống kê thay đổi theo thời gian), một mạng Nơron có thể được thiết kế sao cho có khả năng thay đổi trọng số của nó theo thời gian thực.

Kiến trúc tự nhiên của một mạng Nơron cho việc phân loại mẫu, xử lý tín hiệu và các ứng dụng điều khiển luôn đi đôi với khả năng thích nghi của mạng, tạo cho nó một phương tiện hữa hiệu cho việc phân loại mẫu thích nghi, xử lý tín hiệu thích nghi và các ứng dụng điều khiển thích nghi. Một hệ thống càng có khả năng thích nghi thì tính năng của nó càng mạnh khi hệ thống cần phải hoạt động trong một môi trường không ổn định.

#### c. Khả năng dung thứ lỗi.

Một mạng Nơron được cài đặt dưới dạng phần cứng, vốn có khả năng chấp nhận lỗi hay còn gọi là khả năng tính toán thô (không nhạy cảm lỗi), với ý nghĩa là tính năng của nó chỉ thoái hóa (chứ không đổ vỡ) khi có những điều kiện hoạt động bất lợi. Tuy nhiên, do bản chất phân tán thông tin trong mạng Nơron nên sự hỏng hóc cũng được trải ra trên toàn mạng. Để đảm bảo rằng mạng Nơron thực sự có khả năng chấp nhận lỗi ta cần phải thực hiện những đo đạc hiệu chỉnh trong việc thiết kế thuật toán tích lũy mạng Nơron.

**d. Xử lý các quá trình phi tuyến: xấp xỉ mạng, miễn nhiễu, phân lớp.**

Một Nơron có thể tính toán một cách tuyến tính hay phi tuyến. Vì thế một mạng Nơron cũng có khả năng tính toán phi tuyến khi nó được cấu tạo từ các Nơron phi tuyến. Đây là một thuộc tính quan trọng mà nhờ đó mạng Nơron được ứng dụng rộng rãi vì hầu hết các cơ chế sinh ra các tín hiệu đầu vào đều là phi tuyến.

**Các nhược điểm của mạng Nơron:**

1. Chỉ làm việc được với dữ liệu số.
2. Cần bộ dữ liệu mẫu đủ lớn.
3. Không đưa ra được cơ chế giải thích.
4. Các giải thuật học của mạng đôi khi chưa đảm bảo sự hội tụ cần thiết.

### Các lĩnh vực ứng dụng

Mạng Nơron nhân tạo là phương pháp mô hình hóa tập trung vào các bài toán phi tuyến. Sau đây là một số ứng dụng của mạng Nơron:

1. Các bài toán phân lớp, tách cụm và dự đoán, liên kết.
2. Các bài toán tối ưu: tìm nghiệm cho nhiều lớp bài toán tối ưu toàn cục. Ví dụ: kết hợp mạng Nơron và các thuật toán di truyền.
3. Bài toán hồi qui và tổng quát hóa (thường dùng các thuật học có giám sát): cố gắng tìm các đường thẳng hoặc đường hồi qui phi tuyến trơn sao cho khớp với mẫu.
4. Bài toán hoàn chỉnh dạng: mô hình Markov và các mạng có độ trễ với các mạng Nơron nhiều lớp, máy Bolzmann và mạng Hopfield tĩnh.

Mạng Nơron trong một vài năm trở lại đây đã được nhiều người quan tâm và đã áp dụng thành công trong nhiều lĩnh vực khác nhau, như tài chính, y tế, địa chất và vật lý. Thật vậy, bất cứ ở đâu có vấn đề về dự báo, phân loại và điều khiển, mạng Nơron đều có thể ứng dụng được. Ví dụ như khả năng nhận dạng mặt người trong các hệ thống quản lý thông tin liên quan đến con người (quản lý nhân sự ở các công sở, doanh nghiệp; quản lý học sinh, sinh viên trong các trường trung học, đại học và cao đẳng;… ); các ngành khoa học hình sự, tội phạm; khoa học tướng số, tử vi,…

Kết hợp chặt chẽ với logic mờ, mạng Nơron nhân tạo đã tạo nên cuộc cách mạng thực sự trong việc thông minh hóa và vạn năng hóa các bộ điều khiển kỹ thuật cao cho cả hiện nay và trong tương lai. Ví dụ như ứng dụng tự động điều khiển hệ thống lái tàu, hệ thống dự báo sự cố,…

## Huấn luyện mạng Nơron

### Phương pháp huấn luyện

Có ba kiểu huấn luyện chính, mỗi kiểu mẫu tương ứng với một nhiệm vụ học trừu tượng. Đó là học có giám sát, học không có giám sát và học tăng cường. Thông thường, loại kiến trúc mạng nào cũng có thể dùng được cho các nhiệm vụ trên.

**Huấn luyện có thầy** (học có giám sát)

Mạng được huấn luyện bằng cách cung cấp cho nó các cặp mẫu đầu vào và các đầu ra mong muốn. Các cặp được cung cấp bởi "thầy giáo", hay bởi hệ thống trên đó mạng hoạt động. Sự khác biệt giữa các đầu ra thực tế so với các đầu ra mong muốn được thuật toán sử dụng để thích ứng các trọng số trong mạng. Điều này thường được đưa ra như một bài toán xấp xỉ hàm số - cho dữ liệu huấn luyện bao gồm các cặp mẫu đầu vào x, và một đích tương ứng t, mục đích là tìm ra hàm f(x) thoả mãn tất cả các mẫu học đầu vào.

**Huấn luyện cưỡng bức** (học tăng cường)

Trong học tăng cường, dữ liệu x thường không được cho trước mà được tạo ra trong quá trình một agent tương tác với môi trường. Tại mỗi thời điểm t, agent thực hiện hành động yt và môi trường tạo một quan sát xt và một chi phí tức thời ct, theo một quy trình động nào đó (thường là không được biết). Mục tiêu là tìm một sách lược lựa chọn hành động để cực tiểu hóa một chi phí dài hạn nào đó, nghĩa là chi phí tích lũy mong đợi. Quy trình động của môi trường và chi phí dài hạn cho mỗi sách lược thường không được biết, nhưng có thể ước lượng được. Mạng nơ-ron nhân tạo thường được dùng trong học tăng cường như là một phần của thuật toán toàn cục. Các bài toán thường được giải quyết bằng học tăng cường là các bài toán điều khiển, trò chơi, và các nhiệm vụ quyết định tuần tự (sequential decision making) khác.

**Huấn luyện không thầy** (học không có giám sát)

Trong học không có giám sát, ta được cho trước một số dữ liệu x, và hàm chi phí cần được cực tiểu hóa có thể là một hàm bất kỳ của dữ liệu x và đầu ra của mạng, f. Hàm chi phí được quyết định bởi phát biểu của bài toán. Phần lớn ứng dụng nằm trong vùng các bài toán ước lượng như mô hình hóa thống kê, nén, lọc (filtering), blind source seperation và phân mảnh (clustering).

Với cách học không có thầy, không có phản hồi từ môi trường để chỉ ra rằng đầu ra của mạng là đúng. Mạng sẽ phải khám phá các đặc trưng, các điều chỉnh, các mối tương quan, hay các lớp trong dữ liệu vào một cách tự động. Trong thực tế, đối với phần lớn các biến thể của học không có thầy, các đích trùng với đầu vào. Nói một cách khác, học không có thầy luôn thực hiện một công việc tương tự như một mạng tự liên hợp, cô đọng thông tin từ dữ liệu vào.

### Thuật toán luyện mạng

Quá trình huấn luyện mạng là quá trình học với các tập dữ liệu mẫu để điều chỉnh tập trọng số liên kết.

Có nhiều thuật toán có thể dùng cho việc huấn luyện các mô hình mạng Nơron hầu hết có thể được xem là áp dụng trực tiếp của lý thuyết tối ưu hóa và ước lượng thống kê.

Phần lớn các thuật toán huấn luyện mạng nơ-ron sử dụng một kiểu xuống dốc (gradient descent - tiến dần tới cực tiểu địa phương) nào đó. Điều này được thực hiện bằng cách lấy đạo hàm của hàm chi phí theo các tham số của mạng và thay đổi các tham số đó theo một hướng được tính toán theo độ dốc (gradient-related direction) để tiến dần tới cực tiểu địa phương của hàm chi phí.

Các phương pháp thường dùng cho huấn luyện mạng nơ-ron là: phương pháp tiến hóa, luyện thép (simulated annealing), expectation maximisation (cực đại hóa mong đợi) và các phương pháp không tham số (non-parametric methods).

**Thuật toán 1.1. Thuật toán luyện mạng chung**

|  |
| --- |
| *Bước 1: đặt w=w0 (w0 là trọng số khởi tạo)*  *Bước 2: Tính giá trị hàm gộp F(w,x)*  *Bước 3: Tính giá trị hàm kích hoạt (kết xuất)G(f(w,x))*  *Bước 4: Kiểm tra điều kiện dừng:*  *- Nếu đúng: dừng thuật toán: trọng số tối ưu w\*=w*  *- Nếu sai: chuyển bước 5*  *Thông thường người ta hay dùng cách: tính giá trị hàm lỗi (hàm mục tiêu). Đối chiếu với ngưỡng sai số cho phép.*  *- Nếu nhỏ hơn hoặc bằng: dừng thuật toán: trọng số tối ưu w\*=w*  *- Nếu lớn hơn : chuyển bước 5*  *Bước 5: Cập nhật trọng số:*  *- Tính ∆w*  *- Cập nhật trọng số mới: wmới=w+∆w*  *- Đặt w= wmới và quay về bước 2* |

**Hàm mục tiêu**

Để huấn luyện một mạng và xét xem nó thực hiện tốt đến đâu, ta cần xây dựng một hàm mục tiêu (hay hàm giá) để cung cấp cách thức đánh giá khả năng hệ thống. Việc chọn hàm mục tiêu là rất quan trọng bởi vì hàm này thể hiện các mục tiêu thiết kế và quyết định thuật toán huấn luyện nào có thể được áp dụng. Để phát triển một hàm mục tiêu đo được chính xác cái chúng ta muốn không phải là việc dễ dàng. Một vài hàm cơ bản được sử dụng rất rộng rãi. Một trong số chúng là hàm tổng bình phương lỗi (sum of squares error function):



*Trong đó:*

*p: số thứ tự mẫu trong tập huấn luyện*

*i : số thứ tự của đơn vị đầu ra*

*tpi và ypi : tương ứng là đầu ra mong muốn và đầu ra thực tế của mạng cho đơn vị đầu ra thứ i trên mẫu thứ p.*

### Cách thức huấn luyện

**Huấn luyện đồng thời.**

Kiểm tra điều kiện dừng sau khi đã áp dụng bước 1,2 cho tất cả các mẫu sử dụng luyện mạng.

**Huấn luyện từng mẫu.**

Sau mỗi bước 1,2 cho từng mẫu, ta tiến hành kiểm tra điều kiện dừng ngay sau đó cập nhật trọng số. Kết cục tìm ra được bộ trọng số tối ưu của mẫu này ta mới sử dụng chính bộ trọng số tối ưu này để luyện mẫu tiếp theo.

**Huấn luyện theo mẻ.**

Là phương pháp cải tiến nhằm khắc phục những nhược điểm của hai cách thức huấn luyện trên: đó là tiến hành kiểm tra điều kiện dừng và cập nhật trọng số tối ưu sau khi luyện được một số lượng mẫu nhất định.

### Vấn đề thiết kế cấu trúc mạng

Mặc dù, về mặt lý thuyết, có tồn tại một mạng có thể mô phỏng một bài toán với độ chính xác bất kỳ. Tuy nhiên, để có thể tìm ra mạng này không phải là điều đơn giản. Để định nghĩa chính xác một kiến trúc mạng như: cần sử dụng bao nhiêu lớp ẩn, mỗi lớp ẩn cần có bao nhiêu đơn vị xử lý cho một bài toán cụ thể là một công việc hết sức khó khăn. Một số vấn đề cần quan tâm khi thiết kế mạng:

**Số lớp ẩn**

Người ta đã xác định rằng đối với phần lớn các bài toán cụ thể, chỉ cần sử dụng một lớp ẩn cho mạng là đủ. Các bài toán sử dụng hai lớp ẩn hiếm khi xảy ra trong thực tế. Thậm chí đối với các bài toán cần sử dụng nhiều hơn một lớp ẩn thì trong phần lớn các trường hợp trong thực tế, sử dụng chỉ một lớp ẩn cho ta hiệu năng tốt hơn là sử dụng nhiều hơn một lớp. Việc huấn luyện mạng thường rất chậm khi mà số lớp ẩn sử dụng càng nhiều. Nguyên nhân:

1) Phần lớn các thuật toán luyện mạng cho các mạng nơron truyền thẳng đều dựa trên phương pháp gradient. Các lớp thêm vào sẽ thêm việc phải lan truyền các lỗi làm cho vector gradient rất không ổn định. Sự thành công của bất kỳ một thuật toán tối ưu theo gradient phụ thuộc vào độ không thay đổi của hướng khi mà các tham số thay đổi.

2) Số các cực trị địa phương tăng lên rất lớn khi có nhiều lớp ẩn. Phần lớn các thuật toán tối ưu dựa trên gradient chỉ có thể tìm ra các cực trị địa phương, do vậy chúng có thể không tìm ra cực trị toàn cục. Mặc dù thuật toán luyện mạng có thể tìm ra cực trị toàn cục, nhưng xác suất khá cao là chúng ta sẽ bị tắc trong một cực trị địa phương sau rất nhiều thời gian lặp và khi đó, ta phải bắt đầu lại.

3) Dĩ nhiên, có thể đối với một bài toán cụ thể, sử dụng nhiều hơn một lớp ẩn với chỉ một vài đơn vị thì tốt hơn là sử dụng ít lớp ẩn với số đơn vị là lớn, đặc biệt đối với các mạng cần phải học các hàm không liên tục. Về tổng thể, người ta cho rằng việc đầu tiên là nên xem xét khả năng sử dụng mạng chỉ có một lớp ẩn. Nếu dùng một lớp ẩn với một số lượng lớn các đơn vị mà không có hiệu quả thì nên sử dụng thêm một lớp ẩn nữa với một số ít các đơn vị.

**Số đơn vị trong lớp ẩn**

Một vấn đề quan trọng trong việc thiết kế một mạng là cần có bao nhiêu đơn vị trong mỗi lớp. Sử dụng quá ít đơn vị có thể dẫn đến việc không thể nhận dạng được các tín hiệu đầy đủ trong một tập dữ liệu phức tạp, hay thiếu ăn khớp (underfitting). Sử dụng quá nhiều đơn vị sẽ tăng thời gian luyện mạng, là quá nhiều để luyện khi mà không thể luyện mạng trong một khoảng thời gian hợp lý.

Số lượng tốt nhất của các đơn vị ẩn phụ thuộc vào rất nhiều yếu tố - số đầu vào, đầu ra của mạng, số trường hợp trong tập mẫu, độ nhiễu của dữ liệu đích, độ phức tạp của hàm lỗi, kiến trúc mạng và thuật toán luyện mạng.

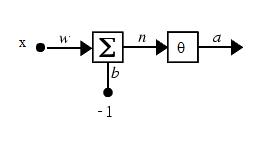
Các luật này chỉ có thể được coi như là các lựa chọn thô khi chọn lựa kích thước của các lớp. Chúng không phản ánh được thực tế, bởi lẽ chúng chỉ xem xét đến nhân tố kích thước đầu vào, đầu ra mà bỏ qua các nhân tố quan trọng khác như: số trường hợp đưa vào huấn luyện, độ nhiễu ở các đầu ra mong muốn, độ phức tạp của hàm lỗi, kiến trúc của mạng (truyền thẳng hay hồi quy), và thuật toán học.

Trong phần lớn các trường hợp, không có một cách để có thể dễ dàng xác định được số tối ưu các đơn vị trong lớp ẩn mà không phải luyện mạng sử dụng số các đơn vị trong lớp ẩn khác nhau và dự báo lỗi tổng quát hóa của từng lựa chọn. Cách tốt nhất là sử dụng phương pháp thử-sai số (trial-and-error).

## Mạng Perceptron và qui tắc học Perceptron

### Mô hình của mạng Perceptron

Mạng Perceptron có cấu trúc như hình vẽ



##### Mô hình của mạng Perceptron

Trong đó ma trận trọng số của mạng:



Chúng ta sẽ định nghĩa một vector bao gồm các phần tử trong hàng thứ i của ma trận trọng số W:



Kết quả vector ra của mạng được xác định là :



Với hàm hardlim hay còn gọi là hàm nhảy là một loại hàm kích hoạt đã được nêu trong phần 1.2.4 của đồ án: 

Như vậy, nếu tích vô hướng của vector dòng thứ i của ma trận trọng số với vector vào là lớn hơn hoặc bằng -b thì giá trị ra là 1, trái lại giá trị ra là 0.

Điều này có ý nghĩa rằng mỗi Nơron trong mạng chia không gian vào thành hai miền.

### Thuật toán luyện mạng Perceptron

Thuật toán học cho mạng Perceptron còn được gọi là thuật toán Perceptron là thuật toán học cho mạng Nơron ra đời sớm. Mặc dù khá đơn giản nhưng không vì thế mà thuật toán này chỉ mang ý nghĩa lịch sử.

**Thuật toán 1.2 - thuật toán luyện mạng Perceptron**

***Đầu vào****: bộ dữ liệu học gồm các cặp vector vào và đầu ra mong muốn*

*, trong đó *

***Đầu ra****: bộ các trọng số thích hợp của mạng Nơron sao cho giá trị đầu ra gần với đầu ra mong muốn (trong trường hợp hội tụ)*

***Các bước***

*Bước 1: khởi tạo ma trận trọng số W và ngưỡng b ban đầu cho mạng Perceptron (có kiến trúc phù hợp với bài toán, nghĩa là có R đầu vào, S Nơron)*

*Bước 2: chọn một mẫu p trong tập dữ liệu học L, lan truyền qua mạng và xác định vector giá trị đầu ra a.*

*Bước 3: xác định vector sai số*  *(vector t tương ứng với vector mẫu p đang xét) và cập nhật lại ma trận trọng số W và ngưỡng b theo công thức sau:*



và



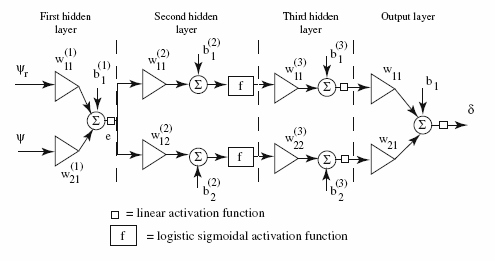
*Thuật toán kết thúc khi không còn sự cập nhật lại ma trận trọng số với mọi mẫu trong tập dữ liệu học L*

## Mạng Perceptron nhiều lớp

Trong phần 1.6 chúng ta đã cùng nhau bàn luận về mạng Perceptron, từ đó đã thấy được rằng có rất nhiều bài toán mà mạng Nơron đó không thể giải quyết được. Tiếp theo chúng ta sẽ bàn về một lớp mạng cho đến nay vẫn đang còn được nghiên cứu và đã có một số thành tựu rất nổi bật.

### Cấu trúc của mạng Perceptron nhiều lớp

Mạng Perceptron nhiều lớp là mạng Nơron có cấu trúc nhiều lớp, nghĩa là các Nơron được phân thành nhiều lớp (số Nơron ở mỗi lớp có thể khác nhau), các lớp được sắp xếp theo một thứ tự nào đó, các Nơron ở lớp trước được liên kết đầy đủ với cái Nơron ở ngay lớp phía sau.



##### Mô hình mạng Perceptron nhiều lớp

### Thuật toán lan truyền ngược

Thuật toán lan truyền ngược là một thuật toán rất quan trọng và được rất nhiều tài liệu bàn luận. Thuật toán lan truyền ngược là một phương pháp cho phép xác định tập trọng số tốt nhất cho mạng Nơron từ các dữ liệu cho trước. Việc áp dụng phương pháp lan truyền ngược là một quá trình lặp đi lặp lại nhiều lần hai tiến trình chính: lan truyền mẫu và lan truyền ngược sai số. Hai tiến trình này được áp dụng trên một tập mẫu xác định. Quá trình luyện mạng được bắt đầu với các giá trị trọng số tùy ý và tiến hành lặp đi lặp lại. Trong mỗi lần lặp mạng hiệu chỉnh các trọng số sao cho sai số giảm dần (sai số là độ lệch giữa các giá trị ra thực và giá trị ra mong muốn). Tiến trình điều chỉnh nhiều lần giúp cho trọng số dần đạt đến bộ giá trị tối ưu.

Để cập nhật trọng số trong mỗi lần lặp, mạng phải xử lý tất cả các mẫu trong tập mẫu. Đối với từng mẫu, mạng phải thực hiện các quá trình sau đây:

* Trước tiên mạng thực hiện các quá trình lan truyền mẫu, nghĩa là đưa mẫu vào mạng và tính toán giá trị ở đầu ra, với bộ trọng số hiện tại.
* Kế tiếp sai số được tính dựa trên giá trị đầu ra thực tế và giá trị đầu ra mong muốn. Trên cơ sở sai số tính toán được, mạng sẽ cập nhật lại các trọng số theo nguyên tắc lan truyền ngược sai số - giai đoạn lan truyền ngược sai số.
* Kỹ thuật sử dụng trong quá trình lan truyền ngược sai số là cập nhật trọng số theo hướng giảm Gradient (thông tin hướng dẫn là đạo hàm bậc nhất của hàm lỗi), sai số (lỗi) của mô hình thường được sử dụng là sai số trung bình bình phương.

Thuật toán lan truyền ngược được áp dụng đối với những mạng Nơron lan truyền thẳng cấu thành bởi các đơn vị xử lý chứa hàm kích hoạt khả vi liên tục. Những mạng lan truyền thẳng như vậy cùng với thuật toán học lan truyền ngược được gọi là mạng lan truyền ngược. Cho trước một tập mẫu học gồm các cặp , như đã nói ở trên, thuật toán lan truyền ngược có hai luồng dữ liệu, trước hết là mẫu vào p(k) được lan truyền từ lớp vào đến lớp ra và kết quả của luồng dữ liệu này cho ta vector đầu ra thực tế o(k). Sau đó lớp tín hiệu lỗi do sự sai khác giữa t(k) và o(k) được lan truyền ngược trở lại từ lớp ra đến các lớp trước để cập nhật lại các trọng số liên kết.

Thuật toán lan truyền ngược lỗi (còn gọi là thuật toán BP) được trình bày như sau:

**Thuật toán 1.3 - thuật toán lan truyền ngược lỗi**

|  |
| --- |
| *Xét một mạng lan truyền thẳng có Q lớp, q = 1,2,…,Q và gọi qneti và qoi là tổng tín hiệu vào và tín hiệu ra của Nơron thứ i trong lớp q. Mạng Nơron này có m đầu vào và n đầu ra. Gọi qwi,j là trọng số liên kết từ q-1oi đến qoi*  ***Đầu vào****: tập thể mẫu để huấn luyện mạng gồm các cặp*  ***Đầu ra****: mạng Nơron với các trọng số đã được điều chỉnh.*  ***Các bước****:*  ***Bước 0*** *(khởi tạo) : Chọn  (sai số lớn nhất chấp nhận được)*  *Khởi tạo các trọng số liên kết với các giá trị ngẫu nhiên bé. Đặt E=0 và k=1*  ***Bước 1*** *(lan truyền tiến): Lan truyền tín hiệu vào qua mạng Nơron theo công thức:*    *Với mọi i,q cho đến khi các đầu ra ở lớp ra Qoi đều xác định được.*  ***Bước 3*** *(xác định sai số):Tính giá trị sai số và tín hiệu lỗi  tại lớp ra*    ***Bước 4*** *(lan truyền ngược sai số): Lan truyền ngược sai số để cập nhật các trọng số và các tín hiệu lỗi  của các lớp trước:*    ***Bước 5*** *(kiểm tra xem toàn bộ mẫu đã được sử dụng hay chưa): Nếu k<L, thì k=k+1 và quay về bước 1, trái lại thì chuyển sang bước 6.*  ***Bước 6*** *(kiểm tra sai số): nếu E<Emax , thì quá trình luyện mạng dừng (thuật toán kết thúc) và đưa ra mạng với bộ trọng số cuối cùng; trái lại E=0,k=1 và thực hiện một vòng huấn luyện mới, nghĩa là trở lại bước 1.* |