

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



HOÀNG THU THỦY

**ỨNG DỤNG KHAI PHÁ DỮ LIỆU ĐỀ TƯ VẤN HỌC TẬP TẠI
TRƯỜNG ĐẠI HỌC SƯ PHẠM THỂ DỤC THỂ THAO HÀ NỘI**

CHUYÊN NGÀNH: HỆ THỐNG THÔNG TIN

MÃ SỐ: 60.48.01.04

TÓM TẮT LUẬN VĂN THẠC SĨ KỸ THUẬT

HÀ NỘI - 2016

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: PGS.TS TRẦN ĐÌNH QUẾ

Phản biện 1:

Phản biện 2:

Luận văn sẽ được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại
Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: ... giờ ngày tháng năm

Có thể tìm hiểu luận văn tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông

MỞ ĐẦU

Tính cấp thiết của đề tài

Giáo dục và đào tạo giữ vai trò hết sức quan trọng đối với sự phát triển của mỗi quốc gia, mỗi dân tộc. Một trong những vấn đề chính mà các sinh viên phải đối mặt khi ngồi trên ghế nhà trường là có một quyết định đúng đắn liên quan tới quá trình học tập của bản thân để có thể đạt được thành tích học tập tốt nhất.

Vì vậy, việc tư vấn học tập để chọn chương trình học phù hợp nhằm đạt được kết quả cao luôn được quan tâm đặc biệt. Khai phá dữ liệu đã và đang được ứng dụng thành công trong giáo dục, có thể giúp sinh viên có thể đưa ra lựa chọn tốt hơn cho quá trình học tập của bản thân.

Trường Đại học Sư phạm Thể dục Thể thao Hà Nội, nơi đào tạo ra đội ngũ giáo viên giáo dục thể chất tương lai cho đất nước luôn cố gắng để hoàn thành tốt công việc của mình. Để giúp các em sinh viên chính quy có thể đưa ra quyết định lựa chọn đúng đắn theo học một chuyên sâu phù hợp với năng lực, mong muốn của bản thân trong quá trình học tập tại trường, tác giả đã lựa chọn đề tài luận văn “*Ứng dụng khai phá dữ liệu để tư vấn học tập tại trường Đại học Sư phạm Thể dục Thể thao Hà Nội*”.

Tổng quan về vấn đề nghiên cứu

Trong những thập kỷ gần đây sự phát triển nhanh chóng của mạng Internet và công nghệ đa phương tiện đã được áp dụng nhiều hơn trong giáo dục. Lợi ích của EDM ngày càng tăng nên các nhà nghiên cứu EDM đã thành lập một tạp chí khoa học vào năm 2009, “Tạp chí khai thác dữ liệu giáo dục”, để chia sẻ và phổ biến kết quả nghiên cứu.

Khai phá dữ liệu giáo dục đề cập đến các kỹ thuật, công cụ, và nghiên cứu thiết kế để tự động trích xuất thông tin có ích từ các kho dữ liệu lớn được tạo bởi người học, liên quan đến người học hoặc các hoạt động trong môi trường giáo dục.

Các kỹ thuật khai phá dữ liệu đã được xem xét và sử dụng trong xây dựng hệ thống tư vấn môn học cho sinh viên, giúp sinh viên đang theo học tại các trường đào tạo theo tín chỉ có thể định hướng trong lựa chọn môn học hay chuyên ngành. Hay xây dựng mô hình khai phá dữ liệu dựa vào thông tin tuyển sinh đầu vào và kết quả thu thập được của sinh viên, nhằm dự đoán kết quả học tập, từ đó giúp sinh viên có thể chọn lựa một lộ trình học đạt kết quả tối ưu nhất phù hợp với điều kiện và năng lực của mình.

Luận văn của tác giả tập trung vào nghiên cứu một số kỹ thuật phân cụm dữ liệu, từ đó chọn kỹ thuật phù hợp để xây dựng hệ thống tư vấn học tập giúp sinh viên trường Đại học

Sư phạm Thể dục Thể thao Hà Nội đánh giá đúng về kỹ năng và năng lực của bản thân trước khi đăng ký theo học một chuyên sâu phù hợp nhất với bản thân.

Mục đích nghiên cứu

- Nghiên cứu, tìm hiểu các vấn đề cơ bản về khai phá dữ liệu, một số kỹ thuật phân cụm dữ liệu để đưa ra một bản tổng hợp có thể giúp cho những nghiên cứu sau này.

- Ứng dụng để xây dựng được hệ thống tư vấn học tập giúp sinh viên chính quy lựa chọn theo học một chuyên sâu phù hợp với bản thân, dựa vào kết quả học tập của sinh viên và dữ liệu thu thập được từ giảng viên trường Đại học Sư phạm Thể dục Thể thao Hà Nội.

Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu:

- Một số kỹ thuật phân cụm dữ liệu.
- Dữ liệu đào tạo chuyên ngành giáo dục thể chất.

Phạm vi nghiên cứu:

- Giới hạn trong một số kỹ thuật phân cụm dữ liệu.
- Dữ liệu thu thập được tại trường Đại học Sư phạm TDTT Hà Nội.

Cấu trúc luận văn:

Ngoài phần mở đầu và kết luận, luận văn được cấu trúc thành 3 chương như sau:

Chương 1: Tổng quan về khám phá tri thức và khai phá dữ liệu

Trình bày tổng quan về khám phá tri thức, khai phá dữ liệu và một số ứng dụng của khai phá dữ liệu trong giáo dục.

Chương 2: Một số kỹ thuật phân cụm dữ liệu

Chương này trình bày khái quát về một số kỹ thuật phân cụm dữ liệu. Phân tích, đánh giá các kỹ thuật để quyết định lựa chọn được thuật toán phù hợp cho việc xây dựng hệ thống tư vấn mà luận văn đưa ra.

Chương 3: Tư vấn học tập cho sinh viên trường Đại học Sư phạm Thể dục Thể thao Hà Nội dựa trên khai phá dữ liệu.

Giới thiệu về bài toán thực tế trong chương trình đào tạo cho sinh viên tại trường Đại học Sư phạm Thể dục Thể thao Hà Nội. Khó khăn cho các sinh viên khi quyết định lựa chọn cho mình một chuyên sâu phù hợp tại trường. Dựa trên khai phá dữ liệu và thuật toán lựa chọn được để xây dựng hệ thống tư vấn học tập cho sinh viên, giúp sinh viên có thể đưa ra quyết định đúng đắn để kết quả học tập đạt tối ưu.

CHƯƠNG 1: TỔNG QUAN VỀ KHÁM PHÁ TRI THỨC VÀ KHAI PHÁ DỮ LIỆU

1.1. Giới thiệu chung về khám phá tri thức và khai phá dữ liệu

1.1.1. Khái niệm về khám phá tri thức và khai phá dữ liệu

Khám phá tri thức (KPTT) là quá trình tìm ra những tri thức, đó là những mẫu tiềm ẩn, trước đó chưa biết và là thông tin hữu ích đáng tin cậy.

Khai phá dữ liệu (KPDL) là một giai đoạn quan trọng trong quá trình khám phá tri thức. Về bản chất nó là giai đoạn duy nhất tìm ra được thông tin mới. KPDL được định nghĩa là quá trình trích lọc các thông tin có giá trị ẩn trong lượng lớn dữ liệu được lưu trữ trong các CSDL hoặc các kho dữ liệu.

Có thể nói rằng hai thuật ngữ khám phá tri thức và khai phá dữ liệu là tương đương nhau nếu ở khía cạnh tổng quan, còn nếu xét ở một góc độ chi tiết thì khai phá dữ liệu là một giai đoạn có vai trò quan trọng trong khám phá tri thức.

1.1.2. Các hướng tiếp cận cơ bản trong khai phá dữ liệu

Khai phá dữ liệu được chia nhỏ thành một số hướng chính như sau:

- Mô tả khái niệm (Concept description)
- Luật kết hợp (Association rules)
- Phân lớp và dự đoán (Classification and prediction)
- Phân cụm (Clustering)
- Khai phá chuỗi (Sequential/Temporal patterns)

1.1.3. Những vấn đề khó khăn trong khai phá dữ liệu

- Các cơ sở dữ liệu lớn, các tập dữ liệu cần xử lý có kích thước rất lớn.
- Mức độ nhiễu cao hoặc dữ liệu bị thiếu.
- Số chiều lớn.
- Thay đổi dữ liệu và tri thức có thể làm cho các mẫu đã phát hiện không còn phù hợp.
- Quan hệ giữa các trường phức tạp.

1.2. Quá trình khám phá tri thức và khai phá dữ liệu

1.2.1. Quá trình khám phá tri thức

Quá trình khám phá tri thức là một chuỗi lặp gồm các bước sau:

Data Cleaning (Làm sạch dữ liệu)

Data Intergration (Tích hợp dữ liệu)

Data Selection (Lựa chọn dữ liệu)

Data Transformation (Biến đổi dữ liệu)

Data Mining (Khai phá dữ liệu)

Pattern Evaluation (Đánh giá mẫu)

Knowledge Presentation (Biểu diễn tri thức)

1.2.2. Quá trình khai phá dữ liệu

Quá trình khai phá dữ liệu bao gồm:

Xác định nhiệm vụ: Xác định chính xác các vấn đề cần giải quyết.

Xác định dữ liệu liên quan: Dùng để xây dựng giải pháp.

Thu thập và tiền xử lý dữ liệu: Thu thập các dữ liệu liên quan và tiền xử lý chính xác cho thuật toán KPDL có thể hiểu được.

Thuật toán KPDL: Lựa chọn thuật toán KPDL và thực hiện việc KPDL để tìm được các mẫu có ý nghĩa.

1.2.3. Các phương pháp khai phá dữ liệu

1.3. Ứng dụng khai phá dữ liệu trong giáo dục

1.3.1. Khai phá dữ liệu giáo dục

Khai phá dữ liệu giáo dục (EDM) mô tả một lĩnh vực nghiên cứu liên quan đến việc áp dụng khai thác dữ liệu, máy học và thống kê các thông tin được tạo ra từ các thiết lập giáo dục (ví dụ, các trường đại học và các hệ thống thông minh).

Khai phá dữ liệu giáo dục đề cập đến các kỹ thuật, công cụ, và nghiên cứu thiết kế để tự động trích xuất thông tin có ích từ các kho dữ liệu lớn được tạo bởi người học, liên quan đến người học hoặc các hoạt động trong môi trường giáo dục.

Ứng dụng khai phá dữ liệu trong giáo dục cung cấp những thông tin hữu ích để thiết kế môi trường học tập, cho phép học sinh, sinh viên, giáo viên, các nhà quản lý và hoạch định chính sách giáo dục đưa ra các quyết định phù hợp.

1.3.2. Mục tiêu của khai phá dữ liệu giáo dục

Baker và Yacef xác định bốn mục tiêu sau đây của EDM:

Dự đoán hành vi học tập trong tương lai của sinh viên.

Khám phá hoặc cải thiện các mô hình miền: thông qua các phương pháp khác nhau và các ứng dụng của EDM, phát hiện mới và cải tiến mô hình hiện tại là có thể.

Nghiên cứu ảnh hưởng của hỗ trợ giáo dục có thể được thực hiện thông qua hệ thống học tập.

Thúc đẩy sự hiểu biết khoa học về việc học tập bằng cách xây dựng và kết hợp mô hình sinh viên, các lĩnh vực nghiên cứu EDM và các công nghệ và phần mềm sử dụng.

1.3.3. Các giai đoạn của khai phá dữ liệu giáo dục

1.3.4. Một số lĩnh vực ứng dụng của EDM

Một số lĩnh vực ứng dụng của EDM là:

- Phân tích và trực quan dữ liệu.
- Cung cấp thông tin phản hồi để hỗ trợ giáo viên.
- Dự đoán kết quả học tập.
- Kiến nghị cho sinh viên.
- Phát hiện hành vi sinh viên không mong muốn.
- Xây dựng chương trình học.
- Kế hoạch và lập kế hoạch.

1.4. Kết luận chương

Nội dung chương đã tìm hiểu quá trình phát hiện tri thức và các vấn đề khai phá dữ liệu. Phát hiện tri thức là một quá trình rút ra tri thức từ dữ liệu mà trong đó khai phá dữ liệu là giai đoạn chủ yếu. Khai phá dữ liệu là nhiệm vụ khám phá các mẫu có ích từ số lượng lớn dữ liệu, ở đó dữ liệu có thể được lưu trữ trong các CSDL, kho dữ liệu hoặc kho lưu trữ thông tin khác. Chương này đã tóm tắt một số phương pháp phổ biến dùng để khai phá dữ liệu và phân tích việc khai phá dữ liệu, ứng dụng khai phá dữ liệu trong giáo dục.

CHƯƠNG 2: MỘT SỐ KỸ THUẬT PHÂN CỤM DỮ LIỆU

2.1. Một số kỹ thuật phân cụm

2.1.1. Phương pháp phân hoạch (*Partitioning Methods*)

2.1.1.1. Thuật toán k-means

Mục đích của thuật toán là sinh ra k cụm dữ liệu $\{C_1, C_2, \dots, C_k\}$ từ một tập dữ liệu ban đầu gồm n đối tượng trong không gian d chiều $X_i = (x_{i1}, x_{i2}, \dots, x_{id})$ ($i = \overline{1, n}$), sao cho hàm tiêu chuẩn $E = \sum_{i=1}^k \sum_{x \in C_i} D^2(x - m_i)$ đạt giá trị cực tiểu. Trong đó m_i là trọng tâm của cụm C_i . D là khoảng cách giữa hai đối tượng. Thuật toán k-means gồm các bước cơ bản sau

Input: Số các cụm k , cơ sở dữ liệu gồm n đối tượng.

Output: Các cụm C_i ($i=1, \dots, k$) sao cho hàm tiêu chuẩn E đạt giá trị tối thiểu.

Bước 1: Khởi tạo k điểm trọng tâm cụm bằng cách chọn k đối tượng tùy ý.

Bước 2: Lặp các bước

- Với mỗi đối tượng X_i ($1 \leq i \leq n$), tính khoảng cách từ nó tới mỗi trọng tâm m_j với $j=1, \dots, k$. Sau đó tìm trọng tâm gần nhất đối với mỗi đối tượng.

- Với mỗi $j=1, \dots, k$, cập nhật trọng tâm cụm m_j bằng cách xác định trung bình cộng của các vector đối tượng dữ liệu.

Bước 3: Thuật toán dừng khi giá trị E không thay đổi.

2.1.1.2. Thuật toán PAM (Partitioning Around Medoids)

Thuật toán PAM là thuật toán mở rộng của thuật toán k-means, có khả năng xử lý hiệu quả đối với dữ liệu nhiễu hoặc các phần tử ngoại lai. PAM sử dụng các đối tượng medoid (lấy một đối tượng đại diện trong cụm gọi là medoid, nó là điểm đại diện được định vị trung tâm nhất trong cụm) để biểu diễn cho các cụm dữ liệu.

Để xác định các medoid, PAM bắt đầu bằng cách lựa chọn k đối tượng medoid bất kỳ. Sau mỗi bước thực hiện, PAM cố gắng hoán chuyển giữa đối tượng medoid O_m và đối tượng O_p không phải medoid, miễn là sự hoán chuyển này nhằm cải thiện chất lượng của phân cụm, quá trình này kết thúc khi chất lượng của phân cụm không thay đổi. Chất lượng phân cụm được đánh giá thông qua hàm tiêu chuẩn, chất lượng phân cụm tốt nhất khi hàm tiêu chuẩn đạt giá trị tối thiểu.

2.1.2. Phương pháp phân cấp (Hierarchical Methods)

2.1.2.1. Thuật toán BIRCH

Input: CSDL gồm n đối tượng, ngưỡng T .

Output: k cụm dữ liệu.

Bước 1: Duyệt tất cả các đối tượng trong CSDL và xây dựng một cây CF khởi tạo. Mỗi đối tượng được chèn vào nút lá gần nhất tạo thành cụm con. Nếu đường kính của cụm con này lớn hơn T thì nút lá được tách. Khi một đối tượng thích hợp được chèn vào nút lá, tất cả các nút trở tới gốc của cây được cập nhật với các thông tin cần thiết.

Bước 2: Nếu cây CF hiện thời không có đủ bộ nhớ trong thì tiến hành xây dựng một cây CF nhỏ hơn bằng cách điều khiển bởi tham số T (vì tăng T sẽ làm hòa nhập một số các cụm con thành một cụm, điều này làm cho cây CF nhỏ hơn). Bước này không cần yêu cầu bắt đầu đọc dữ liệu lại từ đầu nhưng vẫn đảm bảo hiệu chỉnh cây dữ liệu nhỏ hơn.

Bước 3: Thực hiện phân cụm: các nút lá của cây CF lưu giữ các đại lượng thống kê của các cụm con. Trong bước này, BIRCH sử dụng các đại lượng thống kê này để áp dụng một số kỹ thuật phân cụm ví dụ như k-means và tạo ra một khởi tạo cho phân cụm.

Bước 4: Phân phối lại các đối tượng dữ liệu bằng cách dùng các đối tượng trọng tâm cho các cụm đã được khám phá từ bước 3. Đây là một bước tùy chọn để duyệt lại tập dữ liệu và gán nhãn lại cho các đối tượng dữ liệu tới các trọng tâm gần nhất. Bước này nhằm để gán nhãn cho các dữ liệu khởi tạo và loại bỏ các đối tượng ngoại lai.

Khi hòa nhập hai cụm ta có: $CF=CF1+CF2=(n1+n2, LS1+LS2, SS1+SS2)$

Khoảng cách giữa các cụm có thể đo bằng khoảng cách Euclidean, Manhattan,...

2.1.2.2. Thuật toán CURE

Thuật toán CURE sử dụng chiến lược Bottom – Up của kỹ thuật phân cụm phân cấp. CURE sử dụng nhiều đối tượng để diễn tả cho mỗi cụm dữ liệu.

Bước 1: Chọn một mẫu ngẫu nhiên từ tập dữ liệu ban đầu.

Bước 2: Phân hoạch mẫu này thành nhiều nhóm dữ liệu có kích thước bằng nhau, ý tưởng ở đây là phân hoạch mẫu thành p nhóm dữ liệu bằng nhau, kích thước của mỗi phân hoạch là n'/p (với n' là kích thước của mẫu).

Bước 3: Phân cụm các điểm của mỗi nhóm: ta thực hiện phân cụm dữ liệu cho các nhóm cho đến khi mỗi nhóm được phân thành $n'/(pq)$ cụm (với $q>1$).

Bước 4: Loại bỏ các phần tử ngoại lai: trước hết, khi các cụm được hình thành cho đến khi số các cụm giảm xuống một phần so với số các cụm ban đầu. Sau đó, trong trường

hợp các phần tử ngoại lai được lấy mẫu cùng với quá trình pha khởi tạo mẫu dữ liệu, thuật toán sẽ tự động loại bỏ các nhóm nhỏ.

Bước 5: Phân cụm các cụm không gian: các đối tượng đại diện cho các cụm di chuyển về hướng trung tâm cụm, nghĩa là chúng được thay thế bởi các đối tượng gần trung tâm hơn.

Bước 6: Đánh dấu dữ liệu với các nhãn tương ứng.

2.1.3. Phương pháp dựa trên mật độ (*Density-Based Methods*)

2.1.3.1. Thuật toán DBSCAN

Ý tưởng chính để phát hiện ra các cụm của thuật toán DBSCAN là bên trong mỗi cụm luôn tồn tại một mật độ cao hơn bên ngoài cụm. Hơn nữa, mật độ ở những vùng nhiều thì thấp hơn mật độ bên trong của bất kỳ cụm nào. Trong mỗi cụm phải xác định bán kính vùng lân cận (Eps) và số lượng điểm tối thiểu trong vùng lân cận của một điểm trong cụm (MinPts).

Bước 1: Chọn một đối tượng p tùy ý.

Bước 2: Lấy tất cả các đối tượng mật độ - đến được từ p với Eps và MinPts.

Bước 3: Nếu p là điểm nhân thì tạo ra một cụm theo Eps và MinPts.

Bước 4: Nếu p là một điểm biên, không có điểm nào là mật độ - đến được mật độ từ p và DBSCAN sẽ đi thăm điểm tiếp theo của tập dữ liệu.

Bước 5: Quá trình tiếp tục cho đến khi tất cả các đối tượng được xử lý.

2.1.3.2. Thuật toán OPTICS (Ordering Points To Identify the Clustering Structure)

Thuật toán OPTICS do Ankerst, Breunig Kriegel và Sander đề xuất năm 1999, là thuật toán mở rộng cho thuật toán DBSCAN, bằng cách giảm bớt các tham số đầu vào. Thuật toán thực hiện tính toán và sắp xếp các đối tượng theo thứ tự tăng dần nhằm tự động phân cụm và phân tích cụm tương tác hơn là đưa ra phân cụm một tập dữ liệu rõ ràng. Cấu trúc dữ liệu diễn tả theo thứ tự này dựa trên mật độ chứa thông tin tương đương với phân cụm dựa trên mật độ với một dãy các tham số đầu vào. OPTICS xem xét bán kính tối thiểu nhằm xác định các láng giềng phù hợp với thuật toán.

2.1.3.3. Thuật toán DENCLUDE (DENSity – Base CLUstEring)

Thuật toán DENCLUDE được xây dựng ý tưởng chính như sau:

- Ảnh hưởng của một đối tượng tới láng giềng của nó được xác định bởi hàm ảnh hưởng.

- Mật độ toàn cục của không gian dữ liệu được mô hình phân tích như là tổng tất cả các hàm ảnh hưởng của các đối tượng.

- Các cụm được xác định bởi các đối tượng mật độ cao trong đó mật độ cao là các điểm cực đại của hàm mật độ toàn cục.

Định nghĩa hàm ảnh hưởng: Cho x, y là hai đối tượng trong không gian d , chiều ký hiệu là F^d , hàm ảnh hưởng của y lên x được xác định: $f_B^y: F^d \rightarrow R_o^+$, được định nghĩa dưới dạng một hàm ảnh hưởng cơ bản $f_b: f_B^y(x) = f_b(x, y)$.

Hàm ảnh hưởng là hàm tùy chọn, miễn là nó được xác định bởi khoảng cách $d(x, y)$ của các đối tượng, ví dụ như khoảng cách Euclide. Ví dụ về hàm ảnh hưởng như sau:

Hàm ảnh hưởng sóng ngang: $f_{square}(x, y) = \begin{cases} 0 & \text{if } d(x, y) > \delta \\ 1 & \text{if } d(x, y) \leq \delta \end{cases}$ trong đó δ là một ngưỡng.

Hàm ảnh hưởng Gaussian: $f_{Gauss}(x, y) = e^{-\frac{d(x, y)^2}{2\delta^2}}$

Hàm mật độ của một đối tượng $x \in F^d$ được tính bằng tổng tất cả các hàm ảnh hưởng tác động lên x . Giả sử ta có một tập dữ liệu $D = \{x_1, x_2, \dots, x_n\}$.

Hàm mật độ của x được xác định: $f_B^D(x) = \sum_{i=1}^n f_B^{x_i}(x)$

Hàm mật độ dựa trên hàm ảnh hưởng Gauss được xác định như sau:

$$f_{Gauss}^D(d) = \sum_{i=1}^n e^{-\frac{d(x, x_i)^2}{2\delta^2}}$$

2.1.4. Phương pháp dựa trên lưới (Grid-Based Methods)

Thuật toán STING

Thuật toán STING được đề xuất năm 1997 bởi Wang, Yang và Muntz, trong đó vùng không gian dữ liệu được phân rã thành hữu hạn các ô chữ nhật ở nhiều mức khác nhau. Các ô này hình thành cấu trúc phân cấp như sau: mỗi ô ở mức cao được phân hoạch thành các ô mức thấp hơn trong cấu trúc phân cấp. Giá trị các tham số thống kê cho các đối tượng dữ liệu được tính toán và lưu trữ thông qua các tham số thống kê ở các ô mức thấp hơn (điều này giống với cây CF). Các tham số này gồm có: tham số đếm (count), tham số tối đa (max),...

Các đối tượng dữ liệu lần lượt được chèn vào lưới và các tham số thống kê trên được tính thông qua các đối tượng dữ liệu này. STING có khả năng mở rộng cao, nhưng vì sử dụng phương pháp đa phân giải nên nó phụ thuộc chặt chẽ vào trọng tâm của mức thấp nhất.

2.2. Tổng hợp các thuật toán

Từ các thuật toán đã tìm hiểu được ở trên, ta có bảng tổng hợp đặc tính của các thuật toán như sau:

Bảng 2.1: Đặc tính của các thuật toán

Phương pháp phân hoạch					
Thuật toán	Thông số đầu vào	Tối ưu	Cấu trúc cụm	Xử lý nhiễu	Độ phức tạp
k-means	Số lượng cụm	Cụm riêng biệt	Hình cầu	Không	$O(Ikn)$
PAM	Số lượng cụm	Cụm riêng biệt, bộ dữ liệu nhỏ	Hình cầu	Không	$O(Ik(n-k)^2)$
Phương pháp phân cấp					
BIRCH	Yếu tố nhánh, ngưỡng đường kính	Bộ dữ liệu lớn	Hình cầu	Có	$O(n)$
CURE	Số lượng cụm, số lượng cụm đại diện	Cụm hình dạng bất kỳ, bộ dữ liệu tương đối lớn	Hình dạng bất kỳ	Có	$O(n^2 \log n)$
Phương pháp dựa trên mật độ					
DBSCAN	Bán kính của cụm, số lượng tối thiểu của các điểm trong cụm	Cụm hình dạng bất kỳ, bộ dữ liệu lớn	Hình dạng bất kỳ	Có	$O(n \log n)$
DENCLUE	Bán kính của cụm, số lượng tối thiểu của các đối tượng	Cụm hình dạng bất kỳ	Hình dạng bất kỳ	Có	$O(n \log n)$
OPTICS	Bán kính của cụm (min, max), số lượng tối thiểu của các đối tượng	Cụm hình dạng bất kỳ	Hình dạng bất kỳ	Có	$O(n \log n)$
Phương pháp dựa trên lưới					
STING	Số lượng ô ở mức thấp nhất, số lượng đối tượng trong ô	Dữ liệu không gian lớn	Hình dạng dọc và biên ngang	Có	$O(n)$

n : số lượng đối tượng, k : số lượng cụm, I : số lần lặp

2.3. Kết luận chương

Nội dung chương đã đề cập đến một số kỹ thuật phân cụm dữ liệu: phương pháp phân hoạch, phương pháp phân cấp, phương pháp dựa trên mật độ, phương pháp dựa trên lưới và một số thuật toán tiêu biểu của từng phương pháp. Đánh giá từng thuật toán để từ đó có thể đưa ra quyết định lựa chọn thuật toán phù hợp cho bài toán mà luận văn đưa ra.

CHƯƠNG 3: TƯ VẤN HỌC TẬP CHO SINH VIÊN TRƯỜNG ĐẠI HỌC SƯ PHẠM THỂ DỤC THỂ THAO HÀ NỘI DỰA TRÊN KHAI PHÁ DỮ LIỆU

3.1. Giới thiệu bài toán

Trường Đại học Sư phạm Thể dục Thể thao Hà Nội là nơi đào tạo ra đội ngũ giáo viên giáo dục thể chất tương lai cho đất nước. Sinh viên trong trường sau khi hoàn thành năm đầu tiên cơ sở phải đứng trước việc lựa chọn theo học một chuyên sâu tại nhà trường. Một số yếu tố ảnh hưởng tới việc lựa chọn này là:

- Thứ nhất, là sở thích của sinh viên. Sinh viên sẽ đăng ký vào chuyên sâu mà mình thích.
- Yếu tố tiếp theo là căn cứ vào chính năng lực học tập của sinh viên. Năng lực học tập phần lớn được phản ánh qua thành tích và điểm số.

Với mong muốn giúp các sinh viên có thể đưa ra một quyết định đúng đắn trong việc lựa chọn theo học một chuyên sâu phù hợp với năng lực mà vẫn đúng sở thích của bản thân, tác giả đã có ý tưởng xây dựng một hệ thống tư vấn học tập cho sinh viên trường Đại học Sư phạm TDTT Hà Nội.

Đặc điểm tuyển chọn sinh viên chuyên sâu:

Tuyển chọn sinh viên chuyên sâu do các Bộ môn chuyên sâu tổ chức dưới sự lãnh đạo của Ban Giám hiệu cùng Phòng Đào tạo và các Phòng ban chức năng của nhà trường. Các bộ môn sẽ đồng thời tổ chức thi tuyển chọn đầu vào chuyên sâu cho sinh viên trước khi kỳ học thứ 3 bắt đầu. Sinh viên được đăng ký theo thứ tự chuyên sâu 1, chuyên sâu 2, chuyên sâu 3. Chuyên sâu 1 là nguyện vọng được ưu tiên hàng đầu, tiếp theo là chuyên sâu 2 và 3.

Số lượng sinh viên trúng tuyển được xét điểm từ tên cao xuống cho đến khi hết chỉ tiêu tuyển chọn và phải đạt tiêu chuẩn đối với từng chuyên sâu. Sinh viên cũng có thể bị loại ngay từ bước đầu nếu không đáp ứng được tiêu chí về chiều cao, cân nặng khi đăng ký vào một Bộ môn chuyên sâu nào đó. (Ví dụ tiêu chí tuyển chọn chuyên sâu Bóng rổ, xem phụ lục).

Hiện nay, trường bao gồm 5 bộ môn lý thuyết và 9 bộ môn thực hành (với 13 chuyên sâu). Sinh viên cùng với việc học các môn học kiến thức cơ sở, xã hội,... thì phần lớn thời gian dành cho việc học tập và rèn luyện chuyên sâu thể dục thể thao.

Tác giả lựa chọn 4 chuyên sâu của trường: CS Thể dục, CS Điền kinh, CS Bơi lội và CS Bóng rổ để xây dựng hệ thống tư vấn học tập cho sinh viên vì số lượng sinh viên của CS Thể dục và CS Điền kinh chiếm lượng lớn trong tổng số SV toàn trường. CS Bơi lội và CS Bóng rổ đang ngày càng được quan tâm hơn do nhu cầu xã hội.

3.2. Lựa chọn thuật toán

Từ tìm hiểu và đánh giá một số thuật toán phân cụm dữ liệu trong chương 2, tác giả đã quyết định lựa chọn thuật toán k-means để áp dụng vào bài toán mà luận văn đưa ra. Vì tác giả sử dụng dữ liệu bài toán là các dữ liệu số và các dữ liệu điểm số của sinh viên (giá trị chỉ trải từ 0 - 10 và không có phần tử nhiễu) đáp ứng tốt yêu cầu của thuật toán k-means.

3.3. Xây dựng hệ thống tư vấn học tập

3.3.1. Mục đích của hệ thống

- Hệ thống cho phép sinh viên xem danh sách các sinh viên và thành tích của các sinh viên đăng ký thi tuyển vào các chuyên sâu.
- Phân cụm điểm của các sinh viên đăng ký trong các chuyên sâu để từ đó sinh viên có thể xác định xem thành tích của bản thân hiện đang nằm trong khoảng điểm nào và so sánh với chỉ tiêu của chuyên sâu đưa ra, từ đó có thái độ học tập, rèn luyện đúng đắn.
- Đưa ra thành tích của sinh viên, xếp hạng của sinh viên trong chuyên sâu mà bản thân đăng ký, từ đó đưa ra đánh giá sinh viên có khả năng đỗ vào chuyên sâu mà mình đăng ký hay không hay phải cố gắng.

3.3.2. Yêu cầu hệ thống

- + Dữ liệu được tổ chức trên hệ quản trị cơ sở dữ liệu Microsoft SQL Server 2008.
- + Công cụ lập trình sử dụng Microsoft Visual Studio 2008.

3.3.3. Phân tích xây dựng hệ thống

3.3.3.1. Cơ sở dữ liệu

Dữ liệu được thu thập được từ phòng Đào tạo gồm các đơn đăng ký tuyển chọn CS của sinh viên thuộc hệ đại học chính quy tại trường Đại học Sư phạm TDTT Hà Nội cùng với thông tin về các nội dung thi tuyển và cách đánh giá của các bộ môn chuyên sâu. (Xem phụ lục)

Cơ sở dữ liệu được xây dựng khi tác giả thu thập và trích lọc các thông tin có ích, gồm các bảng sau:

- DanhSachSV (MaSV, TenSV, Gioitinh) lưu trữ Mã sinh viên, Tên sinh viên và giới tính của các sinh viên đăng ký vào 4 chuyên sâu trên.

- CSTheDuc (NV, TTCoTay, TTCoBung, TTChongDay, TTBatBuc) lưu trữ nguyện vọng và thành tích các môn: co tay xà đơn, ke bụng thang gióng, chống đẩy, bật bực 2 phút của sinh viên đăng ký chuyên sâu Thể dục.

- CSDienKinh (NV, TTBatXa, TTChayXPC, TTChayCuLyTB, TTDayTa) lưu trữ nguyện vọng và thành tích các môn: bật xa, chạy 100m xuất phát cao, chạy cự ly trung bình, đẩy tạ của sinh viên đăng ký chuyên sâu Điền kinh.

- CSBoiLoi (NV, TTChongDay, TTBatXa, TTGapCui, TTLatVai) lưu trữ nguyện vọng và thành tích các môn: chống đẩy, bật xa, gập cúi, lật vai của sinh viên đăng ký chuyên sâu Bơi lội.

- CSBongRo (NV, ChieuCao, CanNang, TTBatCao, TTChayConThoi, TTPhoiHop) lưu trữ nguyện vọng, chiều cao, cân nặng và thành tích các môn: bật cao với, chạy con thoi 5x28m, khả năng phối hợp vận động của sinh viên đăng ký chuyên sâu Bóng rổ.

Hình 3.1: Các bảng CSDL

Ví dụ một bảng CSDL của sinh viên đăng ký chuyên sâu Thể dục bao gồm: Mã SV, nguyện vọng SV đăng ký (ở đây CS Thể dục là nguyện vọng 1), thành tích các môn thi mà SV đăng ký.

Table - dbo.CSTheDuc		Summary				
	MaSV	NV	TTCoTay	TTCoBung	TTChongDay	TTBatBuc
▶	SV001	NV1	7	26	27	135
	SV002	NV1	2	20	28	150
	SV003	NV1	12	25	17	155
	SV004	NV1	14	27	20	160
	SV005	NV1	12	23	15	160
	SV006	NV1	13	22	17	165
	SV007	NV1	9	24	13	130
	SV008	NV1	6	24	25	135
	SV009	NV1	5	25	25	150
	SV010	NV1	12	22	15	155
	SV011	NV1	6	27	22	145
	SV012	NV1	9	21	8	160
	SV013	NV1	11	21	11	125
	SV014	NV1	14	24	18	168
	SV015	NV1	13	22	17	168

Hình 3.3: Bảng CSDL sinh viên đăng ký chuyên sâu Thẻ đục

3.3.3.2. Các chức năng chính của hệ thống

- Giao diện chính của hệ thống:

Giao diện chính của hệ thống gồm 2 phần: phần bên trái bao gồm danh sách các SV đăng ký vào các CS cùng các thông tin của SV. Thông tin gồm có: Mã SV, Tên SV, Giới tính, nguyện vọng, và thành tích các nội dung thi cùng số điểm tương ứng, tổng điểm SV có thể đạt được. Phần bên phải là thống kê, hiển thị thông số các cụm: tâm cụm, số SV của từng cụm, điểm cao nhất và thấp nhất của từng cụm (sử dụng thuật toán k-means để phân cụm điểm SV); thông tin của sinh viên: tên SV, điểm từng môn, tổng điểm và thứ tự SV trong tổng số SV đăng ký cùng chuyên sâu; cuối cùng là đánh giá: Sinh viên có khả năng đỗ vào chuyên sâu mà mình đăng ký hay không hay cần phải cố gắng hơn. (Hình 3.4)

The screenshot displays the main interface of the system, titled "Tư vấn học tập". It is divided into two main sections: "Danh sách sinh viên" (Student List) on the left and "Thống kê" (Statistics) on the right.

Danh sách sinh viên: This section includes a search bar with "Chuyên sâu: Thẻ đục" and input fields for "Mã sinh viên:" and "Tên sinh viên:". Below is a table listing students with columns: Mã SV, Tên sinh viên, Giới tính, Nguyên vọng, Co tay xã đơn(Lần), Điểm, Ke bụng thang giòng(Lần), Điểm, Chồng đày(Lần), Điểm, and Bậ buc(L). The table lists 20 students, with SV001 highlighted.

Thống kê: This section displays "Thông số các cụm:" with a table showing statistics for different clusters. The table has columns: Tâm cụm, Số sinh viên, Điểm cao nhất, and Điểm thấp nhất. The first cluster (23.04) is highlighted, showing 27 students, a high score of 25, and a low score of 17.

Below the statistics table, there is a section for "Đánh giá:" (Evaluation) showing the student's name (Trịnh Thị Hiền), scores for each subject (Điểm môn 1: 10, Điểm môn 3: 8, Điểm môn 2: 8, Điểm môn 4: 6), total score (Tổng điểm: 32), and rank (Thứ tự: 80). A red message at the bottom states: "Bạn có khả năng đỗ vào chuyên sâu Thẻ đục".

Hình 3.4: Giao diện chính của hệ thống

- Truy xuất thông tin các sinh viên đăng ký cùng một chuyên sâu: kích chọn tên chuyên sâu cần truy xuất thông tin. Ví dụ, muốn xem toàn bộ thông tin sinh viên chuyên sâu Điền kinh, ta chọn chuyên sâu Điền kinh trong mục Chuyên sâu. Màn hình sẽ hiển thị toàn bộ danh sách sinh viên đã đăng ký CS Điền kinh. (Hình 3.5)

The screenshot shows a window titled 'Tư vấn học tập' (Academic Advisor). It has two main panes. The left pane, 'Danh sách sinh viên' (Student List), contains a table with columns: Mã SV, Tên sinh viên, Giới tính, Nguyên vọng, Bật xa(m), Điểm, Chạy 100m, Điểm, Chạy cự ly TB, Điểm, and Điểm. The right pane, 'Thống kê' (Statistics), contains a table with columns: Tâm cụm, Số sinh viên, Điểm cao nhất, and Điểm thấp nhất. Below this table, it displays the name of the selected student, their scores for each subject, total score, and ranking. At the bottom, there is a message: 'Bạn có khả năng đỗ vào chuyên sâu Điền kinh' (You have the potential to pass into the Track and Field specialization).

Mã SV	Tên sinh viên	Giới tính	Nguyên vọng	Bật xa(m)	Điểm	Chạy 100m	Điểm	Chạy cự ly TB	Điểm	Điểm
SV170	Phạm Hồng Thắm	Nữ	NV1	2,18	6	14s16	9	2p34s00	10	94
SV171	Mã Văn Chùng	Nam	NV1	2,56	6	12s35	8	4p39s01	10	101
SV172	Vũ Diệu Mai	Nữ	NV1	2,2	7	15s52	4	3p53s00	1	75
SV173	Nguyễn Thị Tinh	Nữ	NV1	2,14	5	15s37	5	2p16s00	10	86
SV174	Triệu Văn Chiến	Nam	NV1	2,4	3	13s18	6	5p53s01	2	76
SV175	Nguyễn Minh Hải	Nam	NV1	2,48	4	13s31	5	4p45s01	10	96
SV176	Phan Văn Huỳnh	Nam	NV1	2,65	8	12s30	9	4p33s01	10	77
SV177	Mạc Văn Huy	Nam	NV1	2,45	4	14s23	2	4p25s01	10	72
SV178	Phạm Thị Miên	Nữ	NV1	2,24	7	16s39	2	2p25s00	10	94
SV179	Trần Văn Chiêu	Nam	NV1	2,62	7	13s36	5	5p45s01	3	81
SV180	Trương Việt Hà	Nam	NV1	2,68	8	14s30	2	4p56s01	8	104
SV181	Nguyễn Thanh Huy...	Nữ	NV1	2,23	7	15s88	3	3p30s00	3	98
SV182	Vũ Văn Quý	Nam	NV1	2,68	8	14s22	2	5p58s01	2	98
SV183	Bùi Văn Quỳnh	Nam	NV1	2,66	8	13s37	5	4p47s01	10	71
SV184	Ngô Quang Thanh	Nam	NV1	2,58	6	14s24	2	5p45s01	3	95
SV185	Trần Tiến Đạt	Nam	NV1	2,35	2	14s38	2	4p15s01	10	91
SV186	Đinh Văn Nhuận	Nam	NV1	2,39	2	13s31	5	5p55s01	2	98
SV187	Phạm Thủy Trang	Nữ	NV1	2,27	8	15s22	5	3p55s00	1	61
SV188	Trần Văn Chuyên	Nam	NV1	2,48	4	14s32	2	4p48s01	10	89
SV189	Hà Văn Hoan	Nam	NV1	2,67	8	13s36	5	5p14s01	6	88
SV190	Phạm Trầm Kiên	Nam	NV1	2,37	7	13s15	6	5p47s01	3	106

Tâm cụm	Số sinh viên	Điểm cao nhất	Điểm thấp nhất
18,4	5	20	16
21,86	7	23	21
24,25	4	25	24
26,75	8	28	26
32,42	26	38	29

Tên sinh viên: Phạm Hồng Thắm
Điểm môn 1: 6 Điểm môn 3: 10
Điểm môn 2: 9 Điểm môn 4: 10
Tổng điểm: 35 Thứ tự: 6
Đánh giá:
Bạn có khả năng đỗ vào chuyên sâu Điền kinh

Hình 3.5: Thông tin sinh viên đăng ký chuyên sâu Điền kinh

- Truy xuất thông tin của một sinh viên:

+ Bước 1: chọn chuyên sâu,

+ Bước 2: nhập Mã SV, Tên SV.

Màn hình bên trái hiển thị các thông tin về sinh viên cần tìm kiếm, phần bên phải là thống kê bao gồm: thông số các cụm, tên sinh viên, điểm thi từng môn, tổng điểm, và thứ tự của sinh viên đó trong chuyên sâu mình đăng ký và đánh giá sinh viên có khả năng đỗ vào chuyên sâu mình đã đăng ký hay không.

Ta có màn hình truy xuất thông tin của một SV. (Hình 3.6)

Danh sách sinh viên

Chuyên sâu: **Thế dục** Mã sinh viên: **SV001** Tên sinh viên: **Trịnh Thị Hiền**

Mã SV	Tên sinh viên	Giới tính	Nguyên vọng	Cơ tay và đơn(Lần)	Điểm	Kẻ bụng thang gióng(Lần)	Điểm	Chống dấy(Lần)	Điểm	Bật bực(Lần)
SV001	Trịnh Thị Hiền	Nữ	NV1	7	10	26	8	27	8	135

Thông kê

Thông số các cụm:

Tâm cụm	Số sinh viên	Điểm cao nhất	Điểm thấp nhất
23.04	27	25	17
27.23	13	28	26
29.29	14	30	29
31	15	31	31
34.18	80	38	32

Tên sinh viên: **Trịnh Thị Hiền**

Điểm môn 1: **10** Điểm môn 3: **8**

Điểm môn 2: **8** Điểm môn 4: **6**

Tổng điểm: **32** Thứ tự: **80**

Đánh giá:

Bạn có khả năng đỗ vào chuyên sâu **Thế dục**

Hình 3.6: Thông tin của một sinh viên

- **Phân cụm dữ liệu:** thực hiện phân cụm điểm của sinh viên bằng cách áp dụng thuật toán k-means.

Input: Điểm số của các sinh viên trong từng chuyên sâu, số cụm mặc định $k=5$ (tương ứng với các mức đánh giá sinh viên: không đạt, trung bình, trung bình -khá, khá, giỏi).

Output: Các cụm với thông tin về tâm cụm, số phần tử trong cụm, điểm số cao nhất và thấp nhất trong cụm.

Thực hiện phân cụm k-means: PhanCum(float[] _data, int _socum)

- Đầu vào: _data: điểm của sinh viên trong từng môn.

_socum: số lượng cụm.

- Đầu ra: một mảng trong đó phần tử thứ i lưu giá trị của cụm mà phần tử đó thuộc về. Ví dụ: điểm của sinh viên thứ 3 nằm ở cụm 2 thì phanbocum[2]=2.

Các bước của thuật toán:

Bước 1: Khởi tạo tâm cụm với hàm KhoiTaoTamCum(_data, _socum);

- Đầu vào: _data: điểm của sinh viên trong từng môn.

_socum: số lượng cụm.

- Đầu ra: Một mảng lưu các tâm cụm khởi tạo ban đầu. Ví dụ cụm thứ i là phần tử thứ j thì ta có $_tamcum[i]=j$.

Thực hiện bằng cách chọn ngẫu nhiên các phần tử trong cụm sao cho các phần tử này không trùng nhau làm tâm cụm.

public float[] KhoiTaoTamCum(float[] _data, int _socum)

```

{
    Random rd = new Random();
    float[] _tamcum = new float[_socum];
    _tamcum[0] = _data[rd.Next(0, _data.Length - 1)];
    for (int i = 1; i < _socum; i++)
    {
        bool dung = true;
        while (dung)
        {
            _tamcum[i] = _data[rd.Next(0, _data.Length - 1)];
            int k = 0;
            for (int j = 0; j < i; j++)
            {
                if (_tamcum[i] == _tamcum[j])
                    k++;
            }
            if (k == 0)
                dung = false;
        }
    }

    return _tamcum;
}

```

Bước 2: Phân bố các phần tử vào các cụm với hàm PhanBoCum(float[][] _khoangcach).

- Đầu vào: _khoangcach: mảng hai chiều lưu khoảng cách từ phần tử thứ i tới cụm thứ j. (có nghĩa là tính khoảng cách từ mỗi phần tử tới tất cả các cụm).

- Đầu ra: là mảng thể hiện sự phân bố của các phần tử trong các cụm. Ví dụ phần tử thứ i nằm trong cụm j thì ta có _phanbocum[i]=j

Thực hiện bằng cách: tính khoảng cách của mỗi phần tử tới các tâm cụm bằng hàm TinhKhoangCach(float[] _data, float[] _tamcum). Thực hiện so sánh các khoảng cách tới các tâm của mỗi phần tử. Sau đó đưa phần tử đó vào cụm mà có khoảng cách nhỏ nhất.

```

public int[] PhanBoCum(float[][] _khoangcach)
{
    int[] _phanbocum = new int[_khoangcach.Length];
    for (int i = 0; i < _phanbocum.Length; i++)
    {
        _phanbocum[i] = XacDinhCum(_khoangcach[i]);
    }
    return _phanbocum;
}
public int XacDinhCum(float[] _khoangcachcumi)
{

```

```

float min = Math.Abs(_khoangcachcumi[0]);
int cum = 0;
for (int i = 1; i < _khoangcachcumi.Length; i++)
{
    if (Math.Abs(_khoangcachcumi[i]) < min)
    {
        min = _khoangcachcumi[i];
        cum = i;
    }
}
return cum;
}

```

Xác định lại tâm cụm bằng hàm `XacDinhLaiTamCum(float[] _data, int[] _phanbocum, float[] _tamcum)`.

- Đầu vào: `_data` là điểm của sinh viên ở mỗi môn.

`_phanbocum`: phân bố cụm hiện tại.

`_tamcum`: tâm cụm.

- Đầu ra: tâm cụm mới.

Thực hiện bằng cách: tính tâm cụm mới bằng cách tính trung bình giá trị của các phần tử trong cụm.

```

public float[] XacDinhLaiTamCum(float[] _data, int[] _phanbocum, float[]
_tamcum)
{
    for (int i = 0; i < _tamcum.Length; i++)
        _tamcum[i] = 0;
    int[] _sophantucum = new int[_tamcum.Length];
    for (int i = 0; i < _tamcum.Length; i++)
        _sophantucum[i] = 0;
    for (int i = 0; i < _data.Length; i++)
    {
        for (int j = 0; j < _tamcum.Length; j++)
        {
            if (_phanbocum[i] == j)
            {
                _tamcum[j] += _data[i];
                _sophantucum[j]++;
            }
        }
    }
    for (int i = 0; i < _tamcum.Length; i++)
        _tamcum[i] = _tamcum[i] / _sophantucum[i];
    return _tamcum;
}

```

Phân bố lại cụm với hàm `PhanBoCum(float[][] _khoangcach)` như ở trên.

Bước 3: Kiểm tra điều kiện dừng với hàm KiemTraDieuKienDung(int[] _phanbocumcu, int[] _phanbocummoi).

Thực hiện: kiểm tra xem có sự thay đổi về các cụm hay không. Bằng cách kiểm tra xem sau khi thực hiện bước 2 thì sự phân bố của các cụm có thay đổi hay không?

```
public bool KiemTraDieuKienDung(int[] _phanbocumcu, int[] _phanbocummoi)
{
    int dem = 0;
    for (int i = 0; i < _phanbocumcu.Length; i++)
    {
        if (_phanbocumcu[i] != _phanbocummoi[i])
            dem++;
    }
    if (dem == 0)
        return false;
    else
        return true;
}
```

Kết quả sau khi thực hiện phân cụm điểm của SV bằng cách sử dụng thuật toán k-means. (Hình 3.8)

	Tâm cụm	Số sinh viên	Điểm cao nhất	Điểm thấp nhất
►	23,04	27	25	17
	27,23	13	28	26
	29,29	14	30	29
	31	15	31	31
	34,18	80	38	32
*				

Hình 3.8: Phân cụm điểm của sinh viên

- Tư vấn cho SV thông qua điểm từng môn, tổng điểm và thứ hạng của sinh viên trong tổng số sinh viên đăng ký cùng chuyên sâu, đánh giá SV có khả năng đỗ vào chuyên sâu mình đăng ký hay không. Sinh viên có khả năng đỗ vào chuyên sâu mình đăng ký khi không có điểm môn nào dưới 5 và điểm trung bình các môn đạt điểm khá trở lên. (7 điểm)

<u>Tên sinh viên:</u>	Trịnh Thị Hiền		
Điểm môn 1:	10	Điểm môn 3:	8
Điểm môn 2:	8	Điểm môn 4:	6
Tổng điểm:	32	Thứ tự:	80
<u>Đánh giá:</u>			
Bạn có khả năng đỗ vào chuyên sâu Thể dục			

Hình 3.9: Kết quả điểm các môn thi và đánh giá cho một sinh viên

3.4. Kết luận chương

Nội dung chương đã giới thiệu bài toán thực tế về việc đào tạo và đặc điểm tuyển chọn sinh viên chuyên sâu của trường Đại học Sư phạm TDTT Hà Nội. Dựa trên khai phá dữ liệu và ứng dụng thuật toán k-means tác giả đã xây dựng được hệ thống tư vấn học tập giúp các sinh viên có thể định hướng và đánh giá được năng lực của bản thân, từ đó có kế hoạch học tập và rèn luyện đúng đắn để đạt được kết quả học tập tối ưu.

KẾT LUẬN

Kết quả đạt được

Luận văn “*Ứng dụng khai phá dữ liệu để tư vấn học tập tại trường Đại học Sư phạm Thể dục Thể thao Hà Nội*” đã trình bày được một số vấn đề sau:

Tổng quan về khám phá tri thức và ứng dụng khai phá các dữ liệu được lưu trữ trong các hệ thống thông tin. Khai phá dữ liệu được ứng dụng nhiều trong các lĩnh vực khác nhau của cuộc sống, đặc biệt là ứng dụng khai phá dữ liệu trong giáo dục.

Một số kỹ thuật phân cụm dữ liệu: phương pháp phân hoạch, phương pháp phân cấp, phương pháp dựa trên mật độ, phương pháp dựa trên lưới. Các thuật toán điển hình trong từng phương pháp và đánh giá các thuật toán để lựa chọn được thuật toán k-means áp dụng trong bài toán mà luận văn đưa ra.

Dựa trên khai phá dữ liệu, tác giả đã xây dựng được hệ thống tư vấn học tập cho SV trường Đại học Sư phạm TĐTT Hà Nội. Áp dụng thuật toán k-means để phân cụm điểm của sinh viên đăng ký chuyên sâu, giúp sinh viên có thể xác định thành tích của bản thân. Từ đó có kế hoạch học tập và rèn luyện đúng đắn để đạt được kết quả học tập tốt nhất.

Hướng phát triển

- Để quá trình tư vấn học tập có hiệu quả, cần xây dựng một hệ thống hoàn chỉnh hỗ trợ cả quá trình đào tạo (hỗ trợ thêm chức năng: dự báo kết quả học tập của sinh viên,...)
- Xem xét và nghiên cứu thêm một số ứng dụng khác của khai phá dữ liệu vào các bài toán thực tế trong giáo dục.