Journal of educational equipment: Applied research, Volume 2, Issue 291 (June 2023)
ISSN 1859 - 0810

Ứng dụng kĩ thuật phân cụm dữ liệu hỗ trợ sinh viên lựa chọn chuyên ngành

Nguyễn Thị Tâm, Nguyễn Thị Quỳnh Như*, Nguyễn Thị Thúy Lan**
**Khoa Công nghệ thông tin, Trường Đại học Mở Hà Nội*
Received: 26/4/2023; Accepted: 6/5/2023; Published: 15/5/2023

***Abstract**: Data mining with clustering techniques is applied in many different areas of life. In choosing a training major of students of the Faculty of Information Technology, Hanoi Open University, most students choose based on feelings and personal preferences, but there is no information channel to support and advise students to choose according to their ability. Therefore, in this paper, the authors propose to apply the technique of clustering students by major based on the score groups of specific subjects of each major. The authors research and experiment with data clustering algorithms as K-Means. The grouping results are a suggestion to help students of the Faculty of Information Technology, Hanoi Open University have a basis to choose the right major for themselves.*

***Keywords**: K-Means, clustering, select maijor, centroid, Euclid distance.*

1. Đặt vấn đề

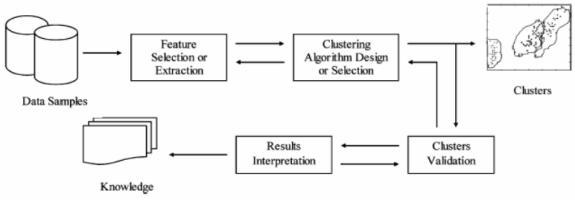
Lựa chọn đúng chuyên ngành đào tạo là rất quan trọng đối với các sinh viên (SV) trong quá trình học tập tại các trường đại học. Chuyên ngành học phù hợp với năng lực sẽ giúp cho SV có niềm đam mê trong học tập để đạt được kết quả tốt nhất. Tuy nhiên hầu hết với các bạn SV việc lựa chọn chuyên ngành thường là cảm tính, theo sở thích của SV mà chưa có căn cứ cụ thể dẫn đến việc chọn chuyên ngành không phù hợp, ảnh hưởng đến kết quả học tập của SV. Khoa Công nghệ thông tin Trường Đại học Mở Hà Nội triển khai đăng ký chuyên ngành đào tạo cho SV từ năm thứ 3, khi SV tích lũy được tối thiểu 100 tín chỉ trên tổng số 140 tín chỉ với 3 chuyên ngành bao gồm: Công nghệ phần mềm, Công nghệ Đa phương tiện, Mạng và an toàn hệ thống. Mỗi chuyên ngành sẽ có những học phần chuyên sâu thể hiện khối kiến thức đặc thù của chuyên ngành đó. Với mỗi đợt đăng kí chuyên ngành SV thường lúng túng trong lựa chọn nên cần sự trợ giúp của cố vấn học tập. Khi đó, cố vấn học tập cần kết hợp kinh nghiệm của bản thân và dữ liệu thu được từ việc tra cứu kết quả học tập của từng SV để tư vấn chuyên ngành phù hợp với năng lực của SV. Công việc này tiêu tốn khá nhiều thời gian và công sức của cố vấn học tập. Vấn đề đặt ra là làm thế nào để sử dụng nguồn dữ liệu kết quả học tập sẵn có nhằm khai thác, phân tích và đưa ra đánh giá, từ đó có thể gợi ý cho SV lựa chọn chuyên ngành học phù hợp nhất một cách hiệu quả. Trong bài báo này, chúng tôi sẽ triển khai thử nghiệm phân cụm dữ liệu với thuật toán K-Means nhằm mục đích hỗ trợ tư vấn, gợi ý lựa chọn chuyên ngành cho SV dựa vào kết quả học tập.

2. Nội dung nghiên cứu

2.1. Phân cụm dữ liệu

Phân cụm dữ liệu là bài toán gom nhóm các đối tượng dữ liệu thành từng cụm sao cho các đối tượng trong cùng một cụm có sự tương đồng theo một tiêu chí nào đó. Tất cả các dữ liệu được biểu diễn bởi các đặc trưng đó là vector n chiều. Các bước cơ bản để phân cụm dữ liệu:

- Trích lọc và tiền xử lý dữ liệu: là bước xác định các đặc trưng dữ liệu liên quan đến yêu cầu bài toán, làm sạch dữ liệu bao gồm xử lý các dữ liệu thiếu, nhiễu,..
- Lựa chọn hoặc thiết kế thuật toán phân cụm: trong rất nhiều thuật toán phân cụm đã có cần lựa chọn thuật toán phù hợp với bài toán để đạt hiệu quả phân cụm.
- Công nhận kết quả phân cụm: khi có kết quả phân cụm chúng ta cần kiểm tra tính đúng đắn của kết quả.
- Giải thích kết quả: Dựa trên kết quả thực nghiệm ta cần phân tích đưa ra được kết luận.



Hình 2.1: Quá trình gom cụm dữ liệu [1]

Kỹ thuật phân cụm đã được áp dụng trong nhiều lĩnh vực [2] [3]. Một số phương pháp phân cụm điển

hình: phân cụm phân hoạch, phân cụm phân cấp, phân cụm dựa trên mật độ, phân cụm dựa trên mô hình, phân cụm có dữ liệu ràng buộc. Trong bài báo này nhóm tác giả lựa chọn phương pháp phân cụm phân hoạch và thử nghiệm với thuật toán phân cụm K-Means nhằm phân cụm SV theo chuyên ngành dựa trên điểm tích lũy của một số học phần.

2.2. Thuật toán K-Means

Thuật toán phân cụm phân hoạch K-Means do MacQueen đề xuất trong lĩnh vực thống kê năm 1967. Thuật toán này thực hiện phân cụm các đối tượng đã cho vào *k* cụm - *k* là số cụm được xác định trước và nguyên dương, sao cho tổng bình phương khoảng cách giữa các đối tượng đến trọng tâm cụm là nhỏ nhất [4].

Từ tập dữ liệu chứa *n* đối tượng, mỗi đối tượng có *d* thuộc tính, ta sẽ phân chia các đối tượng thành *k* cụm $\{C_1, C_2,...C_k\}$ dựa trên các thuộc tính của đối tượng bằng thuật toán này. Ta coi mỗi thuộc tính của đối tượng là một tọa độ trong không gian *d* chiều và biểu diễn đối tượng như một điểm trong không gian *d* chiều $X_i = \{x_{i1}, x_{i2}, ..., x_{id}\}$, $i = 1,2,\ldots n$ sao cho hàm tiêu chuẩn *E* đạt giá trị tối thiểu. :

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} D^2(x, m_i)$$

Trong đó: m_i *là trọng tâm cụm* C_i, *D là khoảng cách giữa hai đối tượng.*

Thuật toán K-Means.

Đầu vào : *n* đối tượng và số cụm *k*

Đầu ra : Các cụm C_i *(i=1,..,k)* sao cho hàm tiêu chuẩn *E* là cực tiểu.

Bước 1. Khởi tạo

● Chọn *k* đối tượng *Cj (j=1,..,k)* là tâm ban đầu của *k* cụm dữ liệu đầu vào (lựa chọn ngẫu nhiên)

Bước 2. Gán tâm cụm theo khoảng cách

● Với mỗi đối tượng x_i *(i=1,..n)* tính khoảng cách của nó tới mỗi tâm C_j với *j=1,..,k.*

● Đối tượng thuộc về cụm C_s khi khoảng cách từ tâm C_s tương ứng đến đối tượng đó là nhỏ nhất $D(x,C_s) = \min D(x, C_j)$

Bước 3. Cập nhật tâm cụm

● Đối với mỗi *j=1,..k*, cập nhật lại tâm cụm C_j bằng cách xác định trung bình cộng của các vector đối tượng dữ liệu đã được gán về cụm.

Bước 4. Lặp và kiểm tra điều kiện dừng

● Lặp lại các bước 2 và 3 cho đến khi các tâm cụm không thay đổi.

2.3. Giải pháp phân cụm chuyên ngành

Khoa Công nghệ thông tin Đại học Mở Hà Nội thực hiện phân chuyên ngành đào tạo khi SV tích lũy được tối thiểu 100/140 tín chỉ. Chương trình đào tạo theo hệ thống tín chỉ hệ đại học chính quy ngành Công nghệ thông tin được chia thành các chuyên ngành bao gồm: Công nghệ phần mềm, Công nghệ Đa phương tiện, Mạng và an toàn hệ thống. Mỗi chuyên ngành sẽ có những học phần chuyên sâu thể hiện khối kiến thức đặc thù của chuyên ngành đó [5]. Hàng kì, Khoa Công nghệ thông tin sẽ tổ chức các buổi giới thiệu chuyên ngành dành cho SV từ đó SV sẽ lựa chọn và đăng kí chuyên ngành phù hợp. Tuy nhiên việc lựa chọn chuyên ngành của SV phần lớn theo sở thích, cảm tính chứ chưa có căn cứ cụ thể nên có thể dẫn đến việc chọn chuyên ngành chưa phù hợp. Nhóm tác giả đưa ra giải pháp và thực nghiệm giải quyết việc gợi ý lựa chọn chuyên ngành cho SV dựa vào kết quả học tập những học phần có kiến thức bổ trợ cho từng chuyên ngành. Cụ thể như sau:

Chuyên ngành Công nghệ phần mềm: Kỹ thuật lập trình cơ sở (CSLT), Cơ sở dữ liệu (CSDL), Lập trình hướng đối tượng (LTHDT)

Chuyên ngành Công nghệ đa phương tiện: Kỹ thuật lập trình cơ sở, Thiết kế Web (NTKW), Lập trình Web cơ bản (LTWE).

Chuyên ngành Mạng và An toàn hệ thống: Kỹ thuật lập trình cơ sở, Mạng và truyền thông (MMT), Quản trị Mạng (QTM).

2.3.1. Thu thập và tiền xử lý dữ liệu

Dữ liệu thu thập ban đầu chứa thông tin điểm học tập của SV. Các tập tin chứa thông tin điểm của các môn trong từng học kì tổ chức dưới dạng bảng có nhiều cột và dòng, trong đó mỗi cột là một môn học, mỗi dòng là kết quả học tập của một SV trong học kì đó. Điểm được lưu dạng số (hệ 10) và chữ (A, B, C, D).

Do dữ liệu được lấy từ nhiều học kì, mỗi học kì xét một số môn nên nhóm tác giả phải tổng hợp dữ liệu từ nhiều tập tin, sau đó loại bỏ các môn học chung chỉ xét các môn học cơ sở ngành có kiến thức bổ trợ cho từng chuyên ngành.

Tiếp đến dữ liệu điểm hiện được lưu cả ở dạng số theo điểm hệ 10 và dạng chữ theo hệ 4 (tín chỉ) nên nhóm cũng cần xử lý để đưa dữ liệu về thuần dạng số. Và theo phương thức đào tạo tín chỉ, SV có thể cải thiện điểm nên nhóm thu thập xét điểm cao nhất trong các lần thi, loại bỏ các SV có điểm không đạt (dưới 4) và các SV chưa có đủ các đầu điểm.

2.3.2. Thực nghiệm

Thuật toán K-Means được áp dụng phân cụm cho từng chuyên ngành, mỗi chuyên ngành có kết quả là 2 cụm, một cụm gồm các SV có khả năng theo học

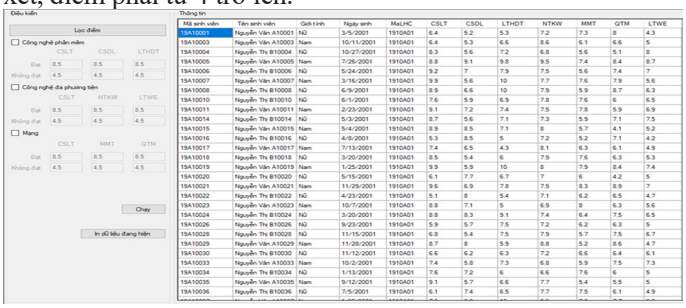
chuyên ngành đó và cụm còn lại là SV không có khả năng học. Trọng tâm ban đầu của mỗi cụm trong từng chuyên ngành được chỉ định với 2 mức gọi là ngưỡng trên và ngưỡng dưới. Ngưỡng trên là những SV có khả năng học chuyên ngành (Đạt), ngưỡng dưới là những SV không có khả năng học chuyên ngành (Không đạt).

Thuật toán K-Means áp dụng vào bài toán như sau:

Đầu vào: Bảng điểm các môn học của SV được tổng hợp đã qua bước làm sạch dữ liệu. Trọng tâm của 2 cụm ứng với ngưỡng trên và ngưỡng dưới.

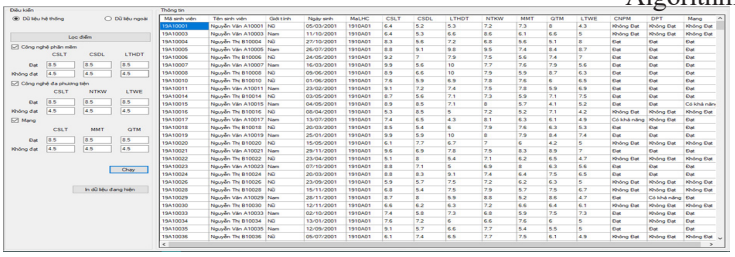
Đầu ra: Danh sách SV được phân cụm theo từng chuyên ngành.

Kết quả triển khai: Tổng hợp điểm SV sau bước tiền xử lý dữ liệu, SV phải có đủ điểm các môn cần xét, điểm phải từ 4 trở lên.



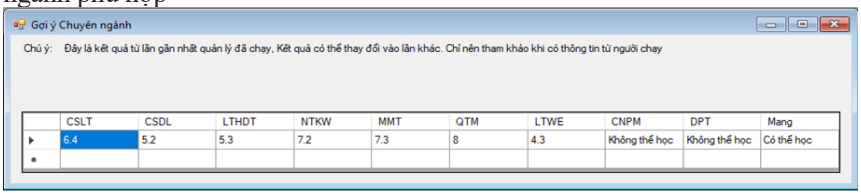
Hình 2.2: Dữ liệu điểm sau tiền xử lý

Khởi tạo trọng tâm ban đầu của 2 cụm ứng với ngưỡng trên (Đạt - điểm xét đạt 8.5) và ngưỡng dưới (Không đạt - điểm xét đạt 4.5) cho mỗi chuyên ngành.



Hình 2.3: Kết quả sau phân cụm

Sinh viên có thể xem điểm và xem gợi ý chuyên ngành phù hợp



Hình 2.4: Kết quả gợi ý chuyên ngành dựa trên điểm

Như vậy sau phân cụm, với mỗi chuyên ngành sẽ được chia thành 2 cụm, cụm 1 là "Đạt" nếu điểm các môn được xét gần ngưỡng trên, ngược lại xét vào cụm 2 "Không đạt". Tuy nhiên do thuật toán K-Means dựa trên khoảng cách để so sánh nên kết quả phân cụm gợi ý vẫn có những hạn chế. Cụ thể với những SV có một đầu điểm xét bị thấp nhưng các điểm khác rất cao thì thuật toán vẫn đánh giá là phù hợp với chuyên ngành. Nhóm giải quyết việc này bằng cách đánh dấu những SV này thuộc diện "Có khả năng" tham gia học chuyên ngành đó.

3. Kết luận

Lựa chọn chuyên ngành đào tạo phù hợp với năng lực SV trong quá trình học tập tại trường đại học là rất quan trọng. Vì thế việc ứng dụng kỹ thuật phân cụm dữ liệu điểm để hỗ trợ tư vấn lựa chọn chuyên ngành học cho SV sẽ là một kênh tham khảo giúp SV Khoa Công nghệ thông tin Trường Đại học Mở Hà Nội có định hướng học tập đúng đắn để sau khi ra trường SV có thể có được công việc đúng khả năng và sở thích. Bài viết này trình bày các nghiên cứu liên quan đến vấn đề hỗ trợ SV lựa chọn chuyên ngành đào tạo sử dụng thuật toán phân cụm dữ liệu K-Means. Thực nghiệm từ các nghiên cứu này cho những gợi ý lựa chọn chuyên ngành khả thi với SV và có thể triển khai trong thực tế.

(* Nghiên cứu này được tài trợ bởi đề tài cấp cơ sở thuộc Trường Đại học Mở Hà Nội, mã số MHN2021-02.03).

Tài liệu tham khảo

1. D. W. I. Rui Xu, "Survey of Clustering Algorithms," *IEEE Transactions on Neural Networks,* vol. 16(3), pp. 645-678, May 2005.

2. S. K. I. M. a. H. Y. S. Winiarti, "Determining the nutrition of patient based on food packaging product using fuzzy C means algorithm," *International Conference on Electrical Engineering, Computer Science and Informatics (EECSI),* pp. 1-6, 2017.

3. P. P. a. S. P. M. Phanich, "Food Recommendation System Using Clustering Analysis for Diabetic Patients," *Conf. Inf. Sci. Appl,* pp. 1-8, 2010.

4. J.MacQueen, "Some methods for Classification and Analysis of Multivariate Observations," *Berkeley Symposium on Mathematical Statistics and Probability,* pp. 281-297, 1967.

5. T. Đ. h. M. H. N. Khoa Công nghệ thông tin, 04 10 2022. [Online]. Available: http://fithou.edu.vn/Article.aspx?aid=2112&cid=3. [Accessed 04 10 2022].