Báo cáo Bài tập thực hành Xử lý đa chiều 03

Nguyễn Phú Thành - MSSV: 18110014

21/6/2021

PROBLEM: Import the CSV file *College.csv* and do the following analytics:
a) Find all private schools with a graduation rate of higher than 90%.
b) Using all attributes from the initial dataset except for the Private attribute, do the K-mean clustering method with two clusters. After this step, we will split the original data into two groups (private or not).
c) Illustrate the predicted cluster and Private attribute (ground truth) to compare them. Please write your comment about the results
Note: Trong bài report này, chủ yếu là về câu c), đưa ra suy nghĩ về kết quả của K-Means clustering trong trường hợp này
Từ kết quả K-Means clustering, ta chia dữ liệu thành hai nhóm, một nhóm là các trường tư, nhóm thứ hai là các trường công
Xem xét, thì ta thấy:

- Trong số 565 trường tư trong bộ dữ liệu, thì có 531 trường thuộc vào nhóm 1 và 34 trường thuộc vào nhóm 2
- Trong số 212 trường công trong bộ dữ liệu, thì có 138 trường thuộc vào nhóm 1 và 74 trường thuộc vào nhóm 2

Như vậy, kết quả phân cụm K-Means dựa vào các feature của bộ dữ liệu không phân tách bộ dữ liệu thành hai phần trường tư và trường công như ta mong muốn
Nhìn vào các thống kê ở 2 nhóm do K-Means phân tách thì dễ dàng nhận thấy các chỉ số trung bình ở nhóm 1 hầu như nhỏ hơn các chỉ số trung bình ở nhóm 2.
Điều này cũng phù hợp với mô tả của K-Means, tức kết quả phân cụm của K-Means chính là phân cụm dựa trên sự tương đồng của các trường trong cùng 1 nhóm. Xem xét các trường nhóm 1 và nhóm 2 ta thấy rõ sự khác nhau về chất lượng dạy và học ở 2 nhóm:

- **Chất lượng giảng viên:**
  - Các trường thuộc nhóm 2 thường có số lượng giảng viên có bằng cấp từ PhD trở lên cao
  - Các trường thuộc nhóm 2 thường có số lượng giảng viên có bằng cấp cuối cùng (terminal degree, bằng cấp cao nhất có thể nhận được trong một ngành) cao
- **Chất lượng sinh viên:**
  - Các trường thuộc nhóm 2 thường có nhiều sinh viên có điểm cao nằm trong top 10/top 25 trong các lớp tích lũy tín chỉ đại học (H.S Class)
  - Tỉ lệ đậu tốt nghiệp cao ở các trưởng thuộc nhóm 2
- Tuy nhiên kèm theo đó là **Chi phí cao**: Học phí từng kì, chi phí phòng, chi phí sách,$\cdots$ ở các trường thuộc nhóm 2 cao hơn hẳn

Nhìn các tên trường có trong nhóm 2, các trường đại học thuộc nhóm Ivy League (một vài thành viên tiêu biểu là Đại học Princeton và Đại học Harvard) đều nằm trong nhóm này, ngoại trừ trường đại học Cornell (thuộc vào nhóm 1). Điều này giải thích khá nhiều về những nhận xét ở trên
Để nhấn mạnh hơn về sự khác biệt này, ta sử dụng Bảng xếp hạng các trường đại học tại Mỹ năm 2021, trong các trưởng top 10 thì có 8 trường thuộc vào nhóm 2, 2 trường còn lại thì không có trong bộ dữ liệu
Như vậy, K-Means ***không*** phân chia các trường thành trường tư hay trường công, mà chia các trường thành 2 nhóm trong đó nhóm 2 ưu thế hơn về chất lượng dạy và học so với nhóm 1