p-ISSN 3030 - 4628
e-ISSN 3030 - 4636

CẢI TIẾN HÀM MỤC TIÊU CHO BÀI TOÁN PHÂN CỤM

Trịnh Thị Anh Loan[1], Phạm Thế Anh[1], Nguyễn Văn Cường[1], Bùi Lương Vũ Ngọc[2]

TÓM TẮT

Phân cụm dữ liệu (Data Clustering) là một phương pháp học máy không giám sát có nhiều ứng dụng trong thực tiễn, đặc biệt là trong kỷ nguyên bùng nổ của dữ liệu. Bài báo này nghiên cứu các mô hình phân cụm dựa trên mạng nơ ron học sâu, tập trung chủ yếu vào các mô hình AutoEncoder như DEC, IDEC. Cụ thể, bài báo nghiên cứu cải tiến hàm mục tiêu của DEC để nâng cao hiệu quả phân cụm dữ liệu. Kết quả thử nghiệm trên tập dữ liệu phức tạp và khó (FMNIST) cho thấy tính hiệu quả của hàm mục tiêu đề xuất so với các mô hình phân cụm hiện đại khác.

Từ khóa: *Phân cụm dữ liệu, mạng học sâu, AutoEncoder, Deep Embedding Clustering.*

DOI: *https://doi.org/10.70117/hdujs.2.2024.742*

1. ĐẶT VẤN ĐỀ

Cho trước một tập gồm n điểm dữ liệu $\{x_i \in X\}_{i=1}^{n}$, bài toán phân cụm được phát biểu như sau: hãy chia tập dữ liệu thành k cụm, mỗi cụm biểu diễn bởi một tâm cụm μ_j với $j = 1, \dots, k$ sao cho các điểm trong mỗi cụm gần nhau nhất có thể; các điểm thuộc các cụm khác nhau sẽ xa nhau nhất có thể. Thông thường, các điểm dữ liệu đầu vào là vector đặc trưng trong không gian số thực nhiều chiều. Do vậy, khoảng cách giữa các điểm dữ liệu được tính toán bởi một độ đo khoảng cách nào đó, tiêu biểu nhất là độ đo Euclidean.

Bài toán phân cụm dữ liệu thu hút được sự quan tâm nghiên cứu rất lớn của các nhà khoa học bởi tiềm năng ứng dụng trong thực tiễn ở nhiều lĩnh vực khác nhau. Trong thời đại phát triển bùng nổ của dữ liệu, các doanh nghiệp và tổ chức đang cạnh tranh trong lĩnh vực chiếm lĩnh và làm chủ thông tin từ dữ liệu. Dữ liệu góp phần quan trọng vào doanh thu, làm thay đổi chiến lược hoạt động và kinh doanh của các cơ quan tổ chức. Phân tích dữ liệu tự động luôn là một trong những ưu tiên hàng đầu để giành lợi thế về năng lực canh tranh. Tuy nhiên, các thuật toán phân cụm dữ liệu truyền thống (tiêu biểu là K-means [1]) thường không đáp ứng được yêu cầu về độ chính xác, thời gian xử lý.

Gần đây, sự phát triển mạnh mẽ của các công nghệ mạng học sâu đã tạo cảm hứng để các nhà nghiên cứu khai thác ưu điểm và thế mạnh của các mô hình nơ ron nhân tạo cho bài toán phân cụm dữ liệu. Nhiều mô hình phân cụm học sâu đã được nghiên cứu đề xuất, tiêu biểu như DEC [4], IDEC [6], Deep spectral clustering [6]. Chi tiết về các mô hình nơ ron học sâu cho bài toán phân cụm dữ liệu được trình bày và phân tích trong [3] [5].

[1]*Khoa Công nghệ thông tin và Truyền thông, Trường Đại học Hồng Đức; Email: trinhthianhloan@hdu.edu.vn*
[2]*Phân hiệu Trường Đại học Y Hà Nội tại Thanh Hóa*

p-ISSN 3030 - 4628
e-ISSN 3030 - 4636

Các mô hình phân cụm dựa trên mạng học sâu ở trên hoạt động khá hiệu quả và vượt trội so với các phương pháp phân cụm truyền thống. Tuy nhiên, hiệu năng phân cụm vẫn khá thấp với các tập dữ liệu khó, chẳng hạn theo kết quả thử nghiệm của [7] thì độ chính xác (ACC) phân cụm của DEC, IDEC trên tập dữ liệu FMNIST [2] tương ứng chỉ đạt 0.518 và 0.529. Nguyên nhân chính nhiều khả năng do tính phân biệt của hàm mục tiêu chưa đủ mạnh, do vậy mô hình chưa được ép buộc để học các đặc trưng đủ tốt. Trong bài bào này, chúng tôi nghiên cứu cải tiến hàm mục tiêu phân cụm của DEC để định hướng mô hình học các đặc trưng có độ tương quan cao với mục tiêu phân cụm. Kết quả nghiên cứu được thử nghiệm trên tập dữ liệu khó FMNIST cho độ chính xác là 0.647, qua đó khẳng định tính ưu việt của giải pháp đề xuất.

2. PHƯƠNG PHÁP NGHIÊN CỨU

Trong quá trình nghiên cứu, nhóm tác giả đã áp dụng và triển khai quy trình nghiên cứu như sau:

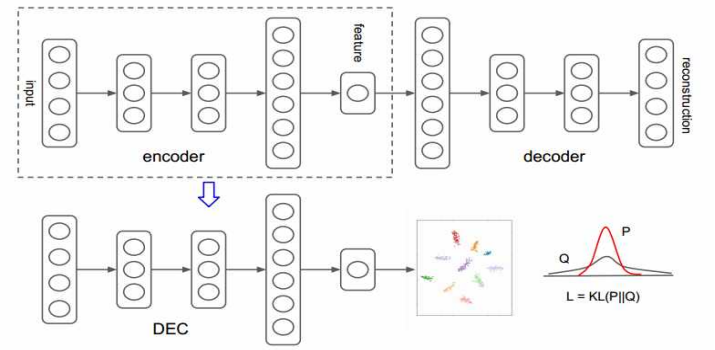
Phương pháp nghiên cứu lý thuyết: Tiến hành thu thập, tổng hợp và nghiên cứu kỹ thuật và thuật toán có liên quan đến lĩnh vực thị giác máy, máy học, mạng nơ ron học sâu, các mô hình phân cụm AutoEncoder, các kiến trúc mạng nơ ron hiện đại. Thiết kế các kiến trúc mạng tích chập học sâu, phân tích và đánh giá về mặt lý thuyết các ưu nhược điểm của các thành phần mạng về phương diện độ phức tạp tính toán cũng như tính năng dự đoán của mạng.

Phương pháp nghiên cứu thực nghiệm: Tổ chức chuẩn bị dữ liệu huấn luyện; Cài đặt đánh giá mô hình mạng xây dựng trên tập dữ liệu chuẩn và so sánh kết quả thực nghiệm với các phương pháp khác.

3. KẾT QUẢ NGHIÊN CỨU VÀ THẢO LUẬN

3.1. Tổng quan về tình hình nghiên cứu

DEC (Deep Embedded Clustering) [4] được xem là mô hình phân cụm đầu tiên dựa trên mạng AutoEncoder. Về cơ bản, một mạng AutoEcnoder gồm hai thành phần là Encoder và Decoder. Cả hai thành phần đều được xây dựng từ các tầng nơ ron (nhân chập hoặc kết nối đầy đủ). Nhiệm vụ của Encoder là mã hóa tín hiệu đầu vào thành các đặc trưng tiềm ẩn trong không gian mới. Ngược lại, thành phần Decoder có vai trò tái tạo lại tín hiệu gốc từ các đặc trưng tiềm ẩn.



Hình 1. Kiến trúc mạng phân cụm DEC [4]

p-ISSN 3030 - 4628
e-ISSN 3030 - 4636

DEC được thiết kế để hoạt động như một AutoEncoder và đạt được những kết quả ấn tượng về hiệu năng phân cụm so với các phương pháp truyền thống. Thay vì thực hiện phân cụm trực tiếp trên tập dữ liệu X, mô hình DEC đầu tiên sẽ biến đổi X sang không gian tiềm ẩn mới bằng thành phần Encoder $f_\theta: X \to Z$ trong đó θ là các tham số có thể huấn luyện được và Z là không gian đặc trưng tiềm năng (hay còn gọi Embebded Feature Space). Thông thường, f_θ là phi tuyến và không gian Z có số chiều nhỏ hơn nhiều so với X. Mô hình DEC cho phép học đồng thời các tâm cụm $\{\mu_i \in Z\}_{i=1}^{k}$ trong không gian Z và các tham số θ của mạng nơ ron. Cụ thể, DEC gồm hai pha xử lý (Hình 1):

Khởi tạo: Khởi tạo các tham số bằng cách huấn luyện một mô hình AutoEncoder. Để điều khiển quá trình học, hàm mục tiêu L_R được sử dụng. L_R đo khoảng cách giữa tín hiệu tái tạo và tín hiệu gốc, được tính như sau:

$$L_R = d_{\mathrm{AE}}\big(x_i, f(x_i)\big) = \sum_i \|x_i - f(x_i)\|^2 \quad (1)$$

Làm mịn: mục tiêu bước này nhằm cập nhật các trọng số của nhánh Encoder để học các đặc trưng phù hợp cho bài toán phân cụm. Để làm việc này, DEC đề xuất mục tiêu L_C được tính toán như sau:

Tính xác suất q_{ij} để điểm dữ liệu $z_i = f(x_i)$ thuộc về tâm cụm μ_j:

$$q_{ij} = \frac{\left(1 + \|z_i - \mu_j\|^2/\alpha\right)^{-\frac{\alpha+1}{2}}}{\sum_{j'}\left(1 + \|z_i - \mu_{j'}\|^2/\alpha\right)^{-\frac{\alpha+1}{2}}} \quad (2)$$

Tính phân bố xác suất đích p_{ij} để giám sát kết quả học:

$$p_{ij} = \frac{q_{ij}^2/f_j}{\sum_{j'} q_{ij'}^2/f_{j'}} \quad (3)$$

Tính giá trị hàm mục tiêu L_C:

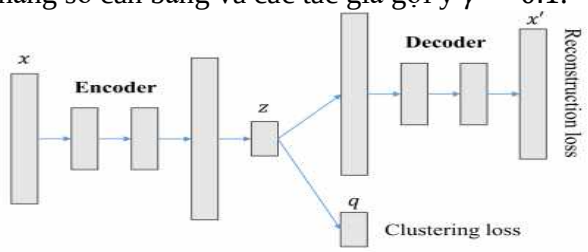
$$L_C = \mathrm{KL}(P \parallel Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \quad (4)$$

Mô hình DEC cho kết quả phân cụm khá tốt trên một số tập dữ liệu tiêu chuẩn. Tuy nhiên, trong pha làm mịn DEC chỉ tối ưu mình nhánh Encoder mà bỏ qua Decoder. Việc này có thể làm biến đổi không nhỏ theo hướng bất lợi các trọng số đã học được của Encoder. Hiệu năng phân cụm do vậy có thể bị ảnh hưởng không nhỏ.

IDEC (Improved DEC) [6] cải tiến DEC ở chỗ trong pha làm mịn cả hai nhánh Encoder và Decoder đều được tối ưu đồng thời. Hình 2 mô tả ý tưởng của IDEC. Cụ thể, IDEC cũng gồm 2 bước là khởi tạo và làm mịn trong đó bước khởi tạo được thực hiện giống như DEC. Tuy nhiên, bước làm mịn sẽ sử dụng hàm mục tiêu mới được tạo thành từ L_C và L_R như sau:

$$L_{IDEC} = L_R + \gamma L_C \quad (5)$$

Trong đó γ là hằng số cân bằng và các tác giả gợi ý $\gamma = 0.1$.



Hình 2. Kiến trúc mô hình IDEC [6]

p-ISSN 3030 - 4628
e-ISSN 3030 - 4636

Kết quả thử nghiệm cho thấy hàm mục tiêu mới giúp IDEC cho kết quả phân cụm vượt trội so với DEC trên nhiều tập dữ liệu. Tuy nhiên, hiệu năng của IDEC, DEC trên tập dữ liệu FMINST vẫn rất hạn chế. Trong phần tiếp theo, chúng tôi sẽ đề xuất một hàm mục tiêu mới giúp nâng cao hiệu năng phân cụm của DEC và IDEC.

3.2. Cải tiến hàm mục tiêu phân cụm

3.2.1. Xây dựng hàm mục tiêu

Trước hết về kiến trúc mạng AutoEncoder, chúng tôi giữ nguyên kiến trúc mạng của DEC gồm hai thành phần là Encoder và Decoder. Cụ thể, mạng Encoder được xây dựng từ các tầng kết nối đầy đủ với số lượng nơ ron ở mỗi tầng lần lượt là: D-500-500-2000-10 trong đó D là chiều của dữ liệu đầu vào, 10 là chiều của không gian tiềm ẩn. Thành phần Decoder sẽ có các tầng đối xứng với Encoder nên sẽ có các nơ ron tại mỗi tầng lần lượt là: 10-2000-500-500-D. Việc giữ nguyên kiến trúc mạng AutoEncoder của DEC giúp việc so sánh hiệu năng được công bằng và khách quan.

Chúng tôi đề xuất hàm mục tiêu mới có tên L_S và sử dụng hàm này trong pha làm mịn (fine-tuning) cùng với hàm L_C ở phương trình (4). Về cơ bản, chúng tôi thiết kế hàm mục tiêu L_S dựa trên quan sát sau:

Nếu hai điểm dữ liệu z_i và z_j được phân về cùng một cụm thì chúng có khả năng cao sẽ nằm gần nhau, và do vậy khoảng cách giữa chúng sẽ nhỏ;

Ngược lại, nếu z_i và z_j được nhóm về hai cụm khác nhau thì chúng có khả năng cao sẽ nằm xa nhau, tương ứng với khoảng cách giữa chúng là lớn.

Để xác định hai điểm z_i và z_j có thuộc về một cụm hay không, chúng tôi sử dụng biến gán mềm nếu q_{ij} ở phương trình (2). Cụ thể:

Gọi $c_i =$ arg max (q_{it}) và $c_j =$ arg max (q_{jt}) trong đó t=1,2,…,K.

Nếu $c_i = c_j$ thì z_i và z_j thuộc về cùng một cụm và ngược lại.

Do vậy, hàm mục tiêu L_s được định nghĩa như sau:

$$L_s = \frac{1}{n_s} \sum_{\substack{i,j \\ c_i = c_j}} \|z_i - z_j\|_2^2 - \frac{1}{n_d} \sum_{\substack{i,j \\ c_i \neq c_j}} \|z_i - z_j\|_2^2 \quad (6)$$

Trong đó: n_s số lượng các cặp điểm z_i và z_j mà $c_i = c_j$, n_d số lượng các cặp điểm z_i và z_j mà $c_i \neq c_j$.

Trên cơ sở hàm mục tiêu mới xây dựng, chúng tôi huấn luyện mạng AutoEncoder như sau:

Khởi tạo: Chúng tôi thực hiện huấn luyện theo kỹ thuật end-to-end (tổng thể) dùng hàm mục tiêu L_R để khởi tạo các tham số cho mạng AutoEncoder. Sau khi huấn luyện xong, K-means được áp dụng trên không gian tiềm ẩn Z để sinh ra K cụm khác nhau. Tâm của K cụm sẽ được gán vào tầng phân cụm như cách mà DEC đã thực hiện. Chúng tôi huấn luyện mạng trên tập dữ liệu FMNIST sẽ công bố kết quả huấn luyện trong phần kết quả thử nghiệm.

Làm mịn: Chúng tôi tiếp tục tích hợp hàm mục tiêu L_C của DEC với các hàm L_S đề xuất để tạo thành hàm mục tiêu tổng thể như sau:

$$L_{new} = L_C + \beta L_s \quad (7)$$

p-ISSN 3030 - 4628
e-ISSN 3030 - 4636

Trong đó β là hằng số cân bằng. Trong thực nghiệm của chúng tôi, giá trị này được gán như sau: $\beta = 0.02$ thông qua kỹ thuật tìm kiếm lưới (grid search) với $\beta \in \{0.001, 0.01, 0.02, 0.1, 0.2, 0.5, 0.75, 1.0\}$.

3.2.2. Kết quả thử nghiệm

Chúng tôi so sánh mô hình trong bài báo này với các mô hình DEC và IDEC trên tập dữ liệu FMNIST gồm có 70K ảnh. Cấu hình mạng và tham số mạng được giữ nguyên như trong các bài báo gốc của DEC, IDEC. Kết quả phân cụm được đo bằng độ chính xác ACC được tính như sau:

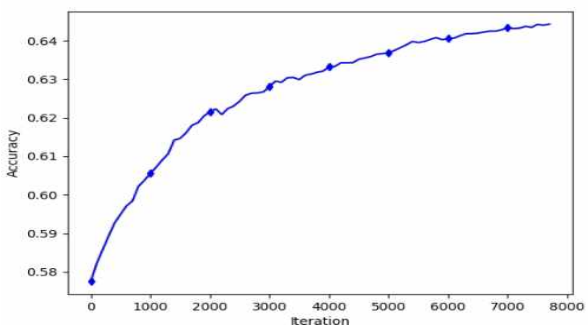
$$ACC = \max_{m} \frac{\sum_{i=1}^{n} 1\{l_i = m(c_i)\}}{n} \quad (8)$$

Trong đó l_i là nhãn đúng của điểm dữ liệu thứ i, c_i là nhãn được gán của điểm dữ liệu thứ i cho một cụm bởi thuật toán phân cụm, m đại diện cho tất cả các phép gán giữa các cụm và các nhãn đúng của dữ liệu. Mỗi phép gán sẽ gán một nhãn đúng cho một cụm tương ứng. Kết quả thử nghiệm được trình bày trên Bảng 1. Như chúng ta có thể quan sát thấy, hệ thống DEC với hàm mục tiêu của chúng tôi cho độ chính xác khá cao so với DEC và IDEC. Hệ thống đề xuất có độ chính xác tăng khoảng 12% so với các hệ DEC và IDEC. Đây là một sự tăng trưởng mạnh về ACC trên tập dữ liệu khó và phức tạp FMNIST.

Bảng 1. Kết quả thử nghiệm (ACC) trên tập dữ liệu FMINST

Hệ thống	DEC	IDEC	Ours
ACC	0.518	0.529	**0.647**

Hình 3 mô tả đồ thị độ chính xác của hệ thống của chúng tôi trong quá trình huấn luyện.



Hình 3. Đồ thị độ chính xác của hệ thống trong quá trình huấn luyện

4. KẾT LUẬN

Trong bài báo này chúng tôi nghiên cứu các mô hình phân cụm dựa trên mạng AutoEncoder. Hai mô hình cụ thể là DEC và IDEC được tìm hiểu và phân tích đánh giá. Trên cơ sở đó, chúng tôi đề xuất cải tiến hàm mục tiêu cho DEC để hoạt động hiệu quả và cho kết quả phân cụm chính xác hơn. Kết quả thử nghiệm trên tập dữ liệu FMNIST cho thấy mô hình của chúng tôi hoạt động khá hiệu quả, cho độ chính xác tăng khoảng 12% so với DEC và IDEC. Trong các nghiên cứu tiếp theo, chúng tôi dự kiến khai thác các ý tưởng các mạng tăng cường để giúp mô hình học các đặc trưng tốt hơn. Các hàm mục tiêu mới cũng được nghiên cứu tích hợp vào hệ thống.

p-ISSN 3030 - 4628
e-ISSN 3030 - 4636

TÀI LIỆU THAM KHẢO

[1] J. B. MacQueen (1967), *Some Methods for classification and Analysis of Multivariate Observations*, Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281-297.

[2] Han Xiao, Kashif Rasul, and Roland Vollgraf (2017), *Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms*, arXiv preprint arXiv:1708.07747.

[3] Elie Aljalbout, Vladimir Golkov, Yawar Siddiqui, Maximilian Strobel, Daniel Cremers (2018), Clustering with Deep Learning: Taxonomy and New Methods, arXiv:1801.07648v2 [cs.LG].

[4] J. Xie, R. Girshick, and A. Farhad (2016), *Unsupervised deep embedding for clustering analysis*, International Conference on Machine Learning (ICML).

[5] Bo Yang, Xiao Fu, Nicholas D. Sidiropoulos, Mingyi Hong (2016), Towards K-means-friendly Spaces: Simultaneous Deep Learning and Clustering, arXiv:1610.04794v2 [cs.LG].

[6] Xifeng Guo, Long Gao, Xinwang Liu, and Jianping Yin (2017), *Improved deep embedded clustering with local structure preservation*, In Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI), pages 1753–1759.

[7] Yang, X., Deng, C., Zheng, F., Yan, J., Liu, W. (2019), *Deep spectral clustering using dual autoencoder network*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

A NEW LOSS FUNCTION FOR DEEP EMBEDED CLUSTERING

Trinh Thi Anh Loan, Pham The Anh, Nguyen Van Cuong, Bui Luong Vu Ngoc

ABSTRACT

Data clustering is an unsupervised machine learning method that has many practical applications, especially in the era of data explosion. This article studies clustering models based on deep learning neural networks, focusing mainly on AutoEncoder models such as DEC, IDEC. Specifically, the study concentrates on improving the loss function of DEC so as to force the model to learn discriminated features sharing high correlation to clustering task. Experimental results on a hard and complicated dataset (i.e., FMNIST) show the effectiveness of the proposed loss function compared to other state-of-the-art clustering models.

Keywords: *Data clustering, deep neural network, AutoEncoder, deep embedding clustering.*

** Ngày nộp bài: 10/3/2024; Ngày gửi phản biện: 20/3/2024; Ngày duyệt đăng: 15/11/2024*