Kỷ yếu Hội nghị Quốc gia lần thứ VIII về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR); Hà Nội, ngày 9-10/7/2015
DOI: 10.15625/vap.2015.000205

TĂNG CHẤT LƯỢNG THUẬT TOÁN PHÂN CỤM NỬA GIÁM SÁT BẰNG PHƯƠNG PHÁP HỌC TÍCH CỰC

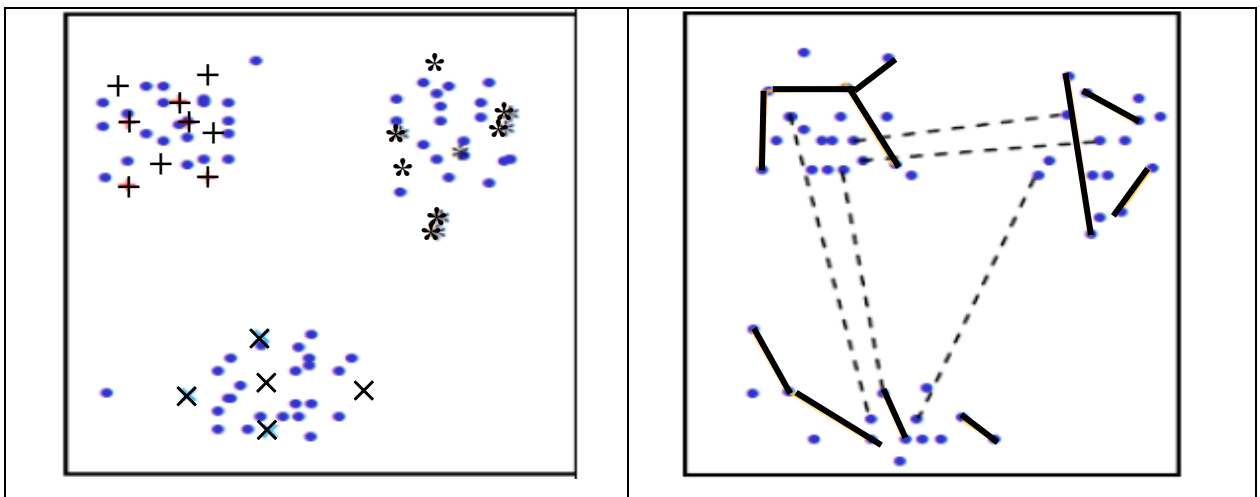
Vũ Việt Vũ
Khoa Điện tử, Trường Đại học Kỹ thuật Công nghiệp – Đại học Thái Nguyên
vuvietvu@tnut.edu.vn

TÓM TẮT *- Vấn đề học tích cực cho bài toán phân cụm nửa giám sát là một trong những chủ đề được quan tâm nghiên cứu trong những năm gần đây. Mục đích của bài báo này là đề xuất một phương pháp học tích cực để thu thập các nhãn cho các điểm dữ liệu nhằm làm tăng chất lượng của các thuật toán phân cụm nửa giám sát. Để thực hiện mục tiêu này, chúng tôi sử dụng một đồ thị k-láng giềng gần nhất để biểu diễn dữ liệu đầu vào đồng thời áp dụng một hàm đánh giá mật độ địa phương trên các đỉnh của đồ thị; dựa vào đánh giá mật độ, các điểm nằm trong vùng đậm đặc của dữ liệu sẽ được lựa chọn để đánh nhãn bởi các chuyên gia. Kết quả thực nghiệm cho thấy, thuật toán mới của chúng tôi cho kết quả tốt hơn các thuật toán cùng loại.*

Từ khóa *- Phân cụm, học tích cực, phân cụm nửa giám sát, seed.*

I. GIỚI THIỆU

Trong khoảng mười năm trở lại đây, thuật toán phân cụm nửa giám sát đã thu hút sự quan tâm của nhiều nhà nghiên cứu trên thế giới. Các thuật toán phân cụm nửa giám sát dạng này hoặc (1) sử dụng các dữ liệu đã gán nhãn (labeled data hay còn gọi là seed) (2) hoặc sử dụng các ràng buộc giữa các điểm dữ liệu (must-link, cannot-link) nhằm mục đích tăng chất lượng của quá trình phân cụm dữ liệu [8,9]. Hình 1 mô tả hai dạng của bài toán học nửa giám sát.

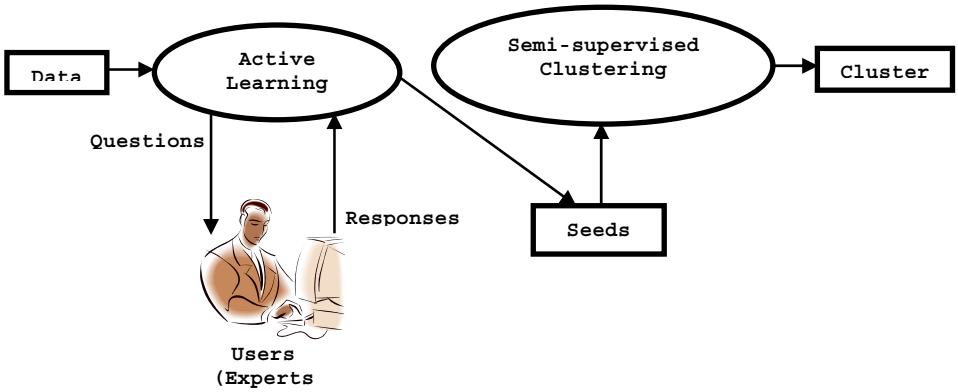


(a) Seed-based clustering | *(b) Constraint-based clustering*

Hình 1. (a) Các điểm tương ứng với tập dữ liệu đầu vào, các seed (các dữ liệu đã được gán nhãn) tương ứng là các điểm ký hiệu bởi các dấu cộng, dấu nhân và dấu sao. (b) các ràng buộc (constraint) must-link (ML) và cannot-link (CL) được biểu diễn tương ứng bằng các đoạn thẳng nét liền và nét đứt: ML(u,v) cho biết u và v thuộc cùng một cụm và CL(u,v) biểu thị u và v sẽ thuộc về hai cụm khác nhau [8].

Một số thuật toán phân cụm nửa giám sát đã được phát triển bao gồm Seed K-Means [2], Seed Fuzzy C-Means [10], SSDBSCAN [11]. Các thuật toán này sử dụng các seed nhằm trợ giúp quá trình khởi tạo dữ liệu và tìm kiếm lời giải (thuật toán Seed K-Means), dùng trong khởi tạo dữ liệu và điều khiển kích thước của các cụm (Seed Fuzzy C-Means) hay dùng để ước lượng mật độ và tính toán tự động tham số (SSDBSCAN). Tuy nhiên các nghiên cứu trên mới chỉ sử dụng các *seed* được lựa chọn ngẫu nhiên và rõ ràng kết quả phân cụm sẽ phụ thuộc vào mỗi lần thực hiện của thuật toán.

Trong bài báo này chúng tôi tập trung phát triển thuật toán nhằm chọn ra các dữ liệu được gán nhãn bởi người sử dụng sao cho tăng chất lượng phân cụm và đồng thời giảm thiểu số câu hỏi đưa ra bởi hệ thống. Hình 2 mô tả chức năng của thuật toán học tích cực. Xuất phát từ dữ liệu đầu vào, thuật toán học tích cực sẽ chọn các dữ liệu để gán nhãn, kết quả sẽ thu được tập các seed, các seed này sẽ sử dụng cho các thuật toán phân cụm nửa giám sát sử dụng các seed (seed-based clustering). Mục đích của thuật toán học tích cực là chọn các seed tốt làm cải thiện chất lượng của các thuật toán phân cụm nửa giám sát.



Hình 2. Sơ đồ học tích cực ứng dụng cho bài toán phân cụm nửa giám sát

Ý tưởng cơ bản của thuật toán đề xuất là dựa trên đồ thị người láng giềng gần nhất. Dữ liệu đầu vào sẽ được biểu diễn bởi đồ thị và trọng số của các cạnh trên đồ thị sẽ được tính toán dựa vào các điểm láng giềng của mỗi đỉnh. Chúng tôi cũng sử dụng một hàm đánh giá mật độ địa phương của các điểm dữ liệu và từ đó có thể chọn ra các ứng cử viên tốt nhất đưa ra gán nhãn bởi các chuyên gia. Chúng tôi cũng lưu ý rằng các nghiên cứu về vấn đề học tích cực ứng dụng cho bài toán phân lớp nửa giám sát (semi-supervised classification) được nghiên cứu mạnh mẽ từ những năm 90 của thế kỷ không nằm trong phạm vi nghiên cứu của bài báo này [12].

Phần tiếp theo của bài báo này được trình bày như sau: Mục II trình bày một số nghiên cứu đã công bố liên quan đến bài báo; mục III trình bày phương pháp thu thập các seed; mục IV trình bày các kết quả thực nghiệm và cuối cùng mục V trình bày kết luận và hướng phát triển của đề tài.

II. CÁC NGHIÊN CỨU ĐÃ CÔNG BỐ

Vấn đề lựa chọn các seed tốt trong bài toán phân cụm đã được nghiên cứu cho thuật toán K-Means, tuy nhiên chỉ dừng lại việc xây dựng các thuật toán nhằm tìm các chiến lược tốt cho việc lựa chọn các điểm trọng tâm trong pha khởi tạo chẳng hạn như trong các nghiên cứu [2 - 6].

Phần tiếp theo chúng tôi trình bày một thuật toán học tích cực nhằm mục đích thu thập các seed cho thuật toán Seed K-Means được giới thiệu trong [7]. Phương pháp này dựa trên thuật toán Min-Max (chúng tôi sẽ gọi là thuật toán SMM). Ý tưởng cơ bản của thuật toán Min-Max là đi xây dựng tập các seed sao cho phủ đều tập dữ liệu đầu vào.

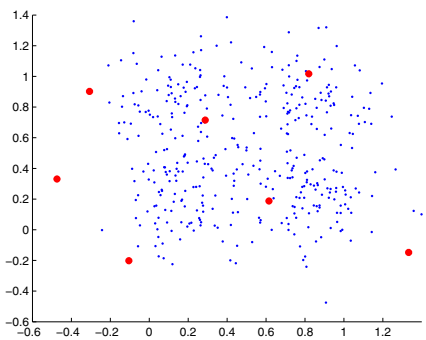
Với phương pháp Min-Max, đầu tiên sẽ chọn ngẫu nhiên một điểm trong tập dữ liệu X đem đi đánh nhãn. Các bước tiếp theo sẽ chọn điểm (ynew) nhằm cực đại hóa các khoảng cách nhỏ nhất từ các điểm chưa có nhãn đến các điểm đã gán nhãn trong tập Y. Điểm y_{new} được xác định theo công thức sau:

$$y_{new} = \text{argmax}_{x_\in X}(\min_{y_\in Y} d(x,y))$$

trong đó y_{new} biểu thị điểm mới sẽ cập nhật vào tập Y và d(.) là hàm khoảng cách (có thể là hàm tính khoảng cách theo Ơ cơ lit hay Mahananobis,…)

Thuật toán SMM bao gồm một quá trình lặp và tại mỗi bước sẽ xác định được một seed mới được đánh nhãn bởi người sử dụng. Chúng tôi cũng lưu ý rằng trong tất cả các hệ thống học tích cực luôn có giả thiết rằng có các chuyên gia trong lĩnh vực bài toán cần giải quyết để đánh nhãn cho các dữ liệu mà hệ thống yêu cầu.

Cuối cùng, hạn chế của phương pháp SMM là nó chỉ phù hợp cho các thuật toán phân cụm phân hoạch chẳng hạn như K-Means hay Fuzzy C-Means. Hơn nữa, các seed thu thập được đôi khi phụ thuộc vào việc lựa chọn ngẫu nhiên seed đầu tiên. Hình 3 minh họa một ví dụ về các seed thu thập được bởi thuật toán SMM.



Hình 3. Ví dụ về các điểm dữ liệu sẽ được đem đi gán nhãn (hình tròn đặc) của thuật toán SMM

III. PHƯƠNG PHÁP HỌC TÍCH CỰC DỰA TRÊN ĐỒ THỊ

A. Đồ thị k-láng giềng gần nhất

Để phát triển phương pháp học tích cực mới, chúng tôi sử dụng đồ thị k-láng giềng gần nhất được giới thiệu trong [13]. Một đồ thị k-láng giềng gần nhất là một đồ thị vô hướng có trọng số và tại mỗi đỉnh sẽ có nhiều nhất k cạnh liên thuộc với nó. Trọng số của hai đỉnh x_i và x_j (kí hiệu là $\omega(x_i, x_j)$) được định nghĩa là số điểm chung trong k-láng giềng gần nhất của u và v và được định nghĩa như sau:

$$\omega(x_i, x_j) = |NN(x_i) \cap NN(x_j)|$$

trong đó NN(v) kí hiệu cho tập k điểm láng giềng gần nhất của v. Một tính chất rất quan trọng của việc tính toán trọng số này là nó sẽ không phụ thuộc vào mật độ địa phương giữa các vùng dữ liệu (khác với việc tính trọng số dựa trên khoảng cách chẳng hạn như khoảng cách Ơ cơ lit). Hình 5-10 minh họa dữ liệu và đồ thị k-láng giềng gần nhất tương ứng với các giá trị khác nhau của k.

B. Thuật toán LOF

Thuật toán LOF [14] được đưa ra nhằm đánh giá các điểm trong tập dữ liệu X ở hai khía cạnh: điểm ngoại lai và điểm bình thường (điểm thuộc cụm).

Để xác định và phân loại các điểm theo tính chất trên, các tác giả đã sử dụng khái niệm độ đo mật độ, dựa trên công thức đánh giá một điểm có phải là điểm ngoại lai hay không.

Cho một điểm $x \in X$ (X là tập dữ liệu trong không gian n chiều), chúng ta định nghĩa các điểm hàng xóm của x như sau:

$$N(x, mp) = \{y \in X \mid d(x, y) \le d(x, x_{mp})\}$$

Trong đó x_{mp} là điểm gần thứ mp trong số các hàng xóm của x; mp là một tham số ngưỡng cho trước. Như vậy N(x,mp) chứa ít nhất mp điểm. Mật độ của x được tính như sau:

$$density(x, mp) = \left(\frac{\sum_{y \in N(x,mp)} d(x, y)}{|N(x, mp)|} \right)^{-1}$$

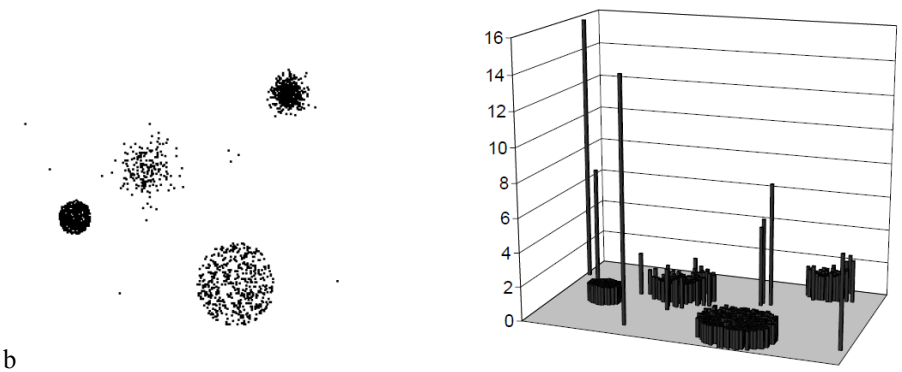
Công thức trên có tính chất sau: giá trị mật độ của x nhỏ thì mật độ của x với các điểm hàng xóm của nó là lớn. Mật độ liên kết trung bình (kí hiệu là *ard*) của x được tính bằng tỷ lệ giữa mật độ của x và mật độ trung bình của các hàng xóm của x như sau:

$$ard(x, mp) = \frac{density(x, mp)}{\left(\frac{\sum_{y \in N(x,mp)} density(y, mp)}{|N(x, mp)|} \right)}$$

Cuối cùng, giá trị LOF được tính như sau :

$$LOF(x, mp) = ard(x, mp)^{-1}$$

Theo [14], một điểm thuộc về một cụm nào đó sẽ có giá trị LOF xấp xỉ 1, nghĩa là mật độ của nó và mật độ của các hàng xóm của nó là tương tự nhau. Hình 4 minh họa dữ liệu đầu vào và giá trị của LOF cho các điểm tương ứng của nó. Hàm tính LOF có thể coi như hàm đánh giá mật độ địa phương của các điểm sẽ được dùng để phát triển thuật toán học tích cực ở phần tiếp theo.



b

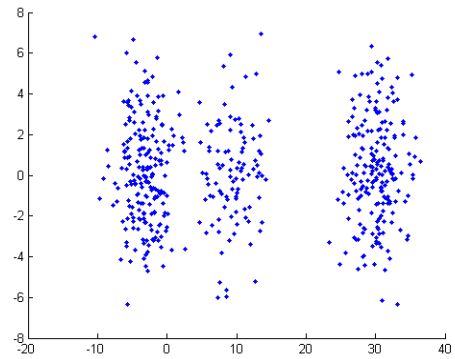
Hình 4. Minh họa thuật toán LOF, giá trị LOF của các điểm dữ liệu bên trái được tính và minh họa bởi đồ thị bên phải [14]

C. Thuật toán học tích cực SkNN-LOF

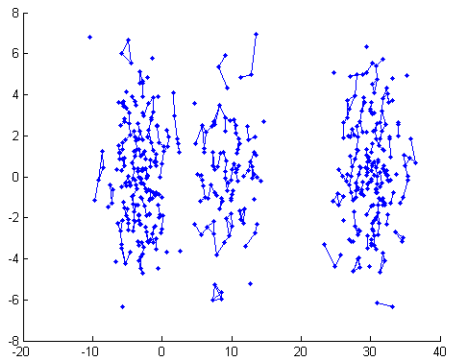
Thuật toán mới của chúng tôi kết hợp giữa đồ thị k-láng giềng gần nhất và thuật toán LOF (kí hiệu là SkNN-LOF) sẽ sử dụng các dữ liệu ứng viên cho việc gán nhãn từ tập *Candidate_set* sau:

$$Candidate\_set = \{p \in X: LOF(x) \approx 1\}$$

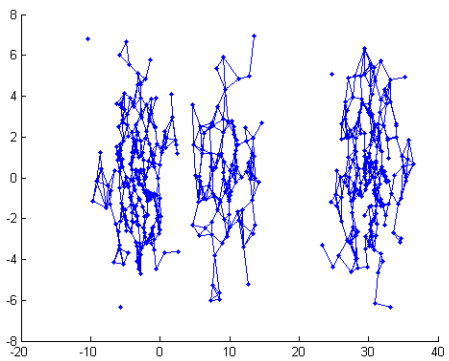
Sử dụng Candidate_set, chúng ta sẽ xây dựng tập các thành phần liên thông và sau đó sắp xếp các thành phần liên thông này theo thứ tự giảm dần về số lượng các đỉnh. Tại mỗi bước lặp của thuật toán, miền liên thông chưa được gán nhãn sẽ được lựa chọn một đỉnh bất kỳ để gán nhãn bởi các chuyên gia. Thuật toán sẽ dừng lại khi hết các ứng viên hoặc dừng bởi người sử dụng. Thuật toán SkNN-LOF được trình bày trong *Thuật toán 1*.



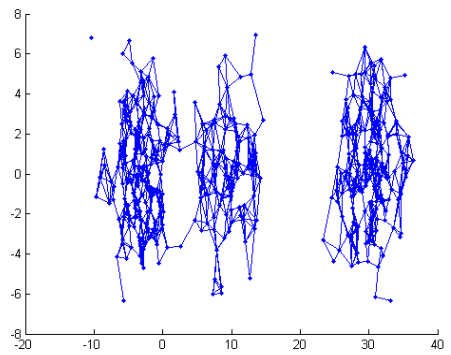
Hình 5. Dữ liệu



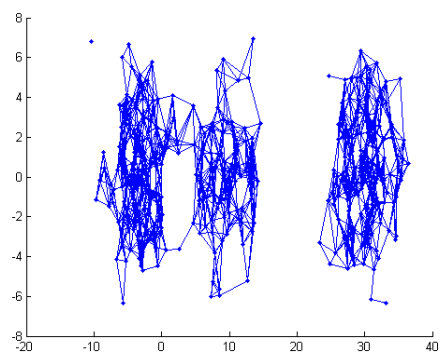
Hình 6. Đồ thị 3-láng giềng gần nhất



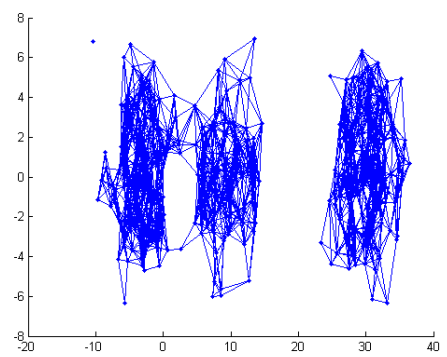
Hình 7. Đồ thị 7-láng giềng gần nhất



Hình 8. Đồ thị 10-láng giềng gần nhất



Hình 9. Đồ thị 15-láng giềng gần nhất



Hình 10. Đồ thị 22-láng giềng gần nhất

Thuật toán 1: SkNN-LOF
Input: Tập dữ liệu X, số lượng láng giềng k
Output: Tập các *seed* Y
Begin
$Y = \Phi$;
Xây dựng đồ thị k-láng giềng gần nhất kNN
$C = \{u \in X: LOF(u) \approx 1\}$
Xây dựng tập các thành phần liên thông từ tập C:
$CC = \{C_1, C_2, \ldots, C_m\}$
Repeat
Chọn ngẫu nhiên $u \in C_v$ sao cho $|C_v|=\max_{c_e CC}|c|$
Đặt câu hỏi với chuyên gia về nhãn của u
$Y = Y \cup \{C_v\}$
$CC = CC - \{C_v\}$
Until (CC = ∅) or (User_stop = true)
End

Độ phức tạp của thuật toán SkNN-LOF phụ thuộc vào việc tính LOF và việc xây dựng đồ thị k-láng giềng gần nhất. Tuy nhiên bản chất việc tính giá trị các LOF cũng chính là dựa trên đồ thị k-láng giềng gần nhất vì vậy độ phức tạp của thuật toán tổng thể phụ thuộc vào việc xây dựng đồ thị k-láng giềng gần nhất. Theo tác giả của LOF [14], với dữ liệu có số chiều nhỏ, chúng ta có thể dùng phương pháp lưới để xác định các láng giềng gần nhất và độ phức tạp sẽ là O(n); với dữ liệu có số chiều trung bình chúng ta có thể sử dụng các phương pháp biểu diễn dữ liệu như R-Tree và độ phức tạp tổng thể sẽ là O(n.log(n)); với dữ liệu có số chiều lớn thì độ phức tạp tổng thể của thuật toán là $O(n^2)$.

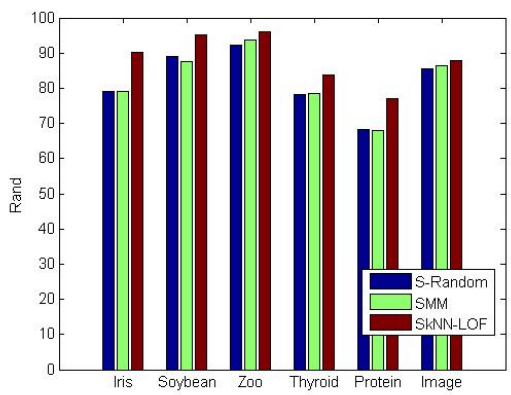
IV.KẾT QUẢ THỰC NGHIỆM

Để đánh giá hiệu quả của thuật toán, chúng tôi sử dụng 6 tập dữ liệu lấy từ trang Machine Learning Repository [15]. Chi tiết về các tập dữ liệu được trình bày trong bảng 1. Chúng tôi sử dụng độ đo Rand Index (Rand) [13], để tính toán chất lượng phân cụm cho các thuật toán. Để so sánh hiệu quả của phương pháp lựa chọn các seed, chúng tôi so sánh 3 thuật toán gồm SkNN-LOF, SMM và S-Random (phương pháp lựa chọn ngẫu nhiên các điểm để đánh nhãn). Thuật toán phân cụm nửa giám sát chúng tôi sử dụng là SSDBSCAN, một phương pháp phân cụm dựa trên mật độ. Kết quả thực nghiệm được cho bởi hình 11.

Bảng 1. Dữ liệu dùng trong thực nghiệm

ID	Tên bộ dữ liệu	N	M	K
1	Protein	115	20	6
2	Iris	150	4	3
3	Glass	214	9	6
4	Thyroid	215	5	3
5	LetterIJL	227	16	3
6	Image	1200	19	7

Từ hình 11, chúng ta thấy rằng phương pháp SkNN-LOF đã thu thập được các nhãn tốt hơn các phương pháp còn lại là SMM và S-Random. Điều này được giải thích rằng việc sử dụng đồ thị k-láng giềng gần nhất là phù hợp để biểu diễn dữ liệu, hơn nữa sử dụng hàm đánh giá mật độ (hàm LOF) để xác định mật độ các điểm nhằm tìm ra các điểm tốt cho việc đánh nhãn và vì vậy đã tăng được chất lượng của thuật toán phân cụm.



Hình 11. Kết quả phân cụm của SSDBSCAN với các seed được lựa chọn bởi 3 phương pháp S-Random SMM, và SkNN-LOF

V. KẾT LUẬN

Trong bài báo này, chúng tôi đề xuất một thuật toán học tích cực nhằm thu thập các seed cho bài toán phân cụm nửa giám sát. Thuật toán được xây dựng dựa trên đồ thị k-láng giềng gần nhất và một hàm đánh giá mật độ các đỉnh của đồ thị. Kết quả thực nghiệm cho thấy, thuật toán của chúng tôi cho kết quả tốt hơn các thuật toán đã có. Trong thời gian tới chúng tôi tiếp tục nghiên cứu đề xuất các thuật toán mới và áp dụng vào các ứng dụng thực tế trong các lĩnh vực như xử lý ảnh, xử lý tiếng nói.

VI.TÀI LIỆU THAM KHẢO

[1] J.M. Penna, J. A. Lozano, and P. Larranaga. An empirical comparison of four initialization methods for the k-means algorithm. Pattern Recognition Letters, 20:1027–1040, 1999.

[2] Amine Ben said, Lawrence O. Hall, James C. Bezdek, Laurence P. Clarke: Partially supervised clustering for image segmentation. Pattern Recognition 29(5): 859-871, 1996.

[3] M. R. Anderberg. Cluster Analysis for Applications. Academic Press, New York, 1973.

[4] L. Kaufman and P. J. Rousseeuw. Finding groups in data: An introduction to cluster analysis. In John Wiley and Sons, 1990.

[5] M. Snarey, N. K. Terrett, P. Willet, and D.J. Wilton. Comparison of algorithms for dissimilarity-based compound selection. J.Mol. Graphics and Modelling, 15: 372-385, 1997.

[6] J. Heer and E. Chi. Identification of web user traffic composition using multi-modal clustering and information scent. In proc. Of the workshop on web mining, SIAM conference on Data mining, 2001.

[7] Viet-Vu Vu, Nicolas Labroche, and Bernadette Bouchon-Meunier. Active Learning for Semi-Supervised K-Means Clustering. In Proceedings of the 22nd IEEE International Conference on Tools with Artificial Intelligence (ICTAI-2010), Arras, France, October, 2010.

[8] Anil K. Jain: Data clustering: 50 years beyond K-means. Pattern Recognition Letters (PRL) 31(8):651-666, 2010.

[9] S. Basu, I. Davidson, and K. L. Wagstaff, Constrained Clustering: Advances in Algorithms, Theory and Applications, Chapman and Hall/CRC Data Mining and Knowledge Discovery Series, 1st edn., 2008.

[10] Sugato Basu, Arindam Banerjee, Raymond J. Mooney: Semi-supervised Clustering by Seeding. ICML 2002: 27-34, 2002.

[11] Levi Lelis, Jörg Sander: Semi-supervised Density-Based Clustering. ICDM : 842-847, 2009.

[12] B. Settles, Active Learning Literature Survey, Technical Report 1648, University of Wisconsin-Madison, 2010.

[13] R.A. Jarvis and E.~A. Patrick, Clustering using a similarity measure based on shared near neighbors, IEEE Transactions on Computer, 22(11), pp: 1025-2034, 1973.

[14] M.M. Breunig, H.P. Kriegel, R.T. Ng, J. Sander. LOF: Identifying Density-based Local Outliers. Proceedings of the ACM SIGMOD International Conference on Management of Data, pp: 93–104, 2000.

[15] M. Lichman, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2013.

[16] W. M. Rand. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, 66 (336): 846–850, 197, 1971.

BOOSTING SEMI-SUPERVISED CLUSTERING BY ACTIVE LEARNING

Vu Viet Vu

ABSTRACT - *The active learning problem for semi-supervised clustering is one of interesting topics in recent years. The purpose of this paper is to develop a method that can collect the labeled data (called seed) to boost the quality of seed based clustering algorithms and reduce the questions to experts. To do that, we use the k-nearest neighbor graph to present input data and apply a local density function to evaluate the density of each data point. Then, the points that are in the dense regions will be chosen to get label by experts. Experimental results show the benefits of our method when compared with Min-Max based method.*