THUẬT TOÁN PHÂN CỤM K –MEANS*

Cao Ngọc Ánh[1], Vũ Việt Vũ[2], Phùng Thị Thu Hiền[3]
[3]Trường Đại học Kinh tế - Kỹ thuật Công nghiệp
[2]Viện Công nghệ Thông tin – Đại học Quốc gia Hà nội
[3]Trường Đại học Kỹ thuật Công nghiệp - Đại học Thái Nguyên

TÓM TẮT

Thuật toán phân cụm nửa giám sát thu hút được nhiều nghiên cứu trong thời gian gần đây. Dựa trên một số thông tin cung cấp bởi người sử dụng như một số điểm dữ liệu đã được gán nhãn sẵn hoặc một số ràng buộc giữa các cặp dữ liệu (must-link, cannot-link) được tích hợp vào các thuật toán phân cụm nửa giám sát sẽ làm cho chất lượng của quá trình phân cụm được cải tiến. Bài báo này đề xuất một phương pháp phân cụm nửa giám sát trong trường hợp dữ liệu trợ giúp là không đầy đủ cho thuật toán phân cụm K-Means, thuật toán mới được đặt tên là K-Means*. Thuật toán K-Means* được lai giữa hai phương pháp gồm phương pháp phân cụm nửa giám sát và phương pháp ước lượng các trọng tâm của các cụm cho K-Means. Kết quả thực nghiệm từ UCI chỉ ra hiệu quả của phương pháp đề xuất.

Từ khóa: *thuật toán phân cụm K-Means, phân cụm nửa giám sát, dữ liệu gán nhãn, phương pháp min-max, dữ liệu UCI*

GIỚI THIỆU

Phân cụm dữ liệu (clustering) là một trong những vấn đề cơ bản của lĩnh vực học máy và khai phá dữ liệu. Bài toán phân cụm nhằm phân tách một tập dữ liệu gồm n phần tử thành k cụm sao cho các phần tử trong mỗi cụm là tương tự nhau và các phần tử nằm khác cụm thì sẽ không tương tự nhau theo một nghĩa nào đó. Phân cụm dữ liệu thuộc lớp bài toán học không giám sát khác loại với các thuật toán học có giám sát chẳng hạn như mạng Nơron hay phương pháp Support Vector Machine [14]. Kết quả của quá trình phân cụm sẽ cho chúng ta biết cấu trúc ẩn chứa bên trong của toàn bộ tập dữ liệu và mối liên hệ giữa các phần tử trong dữ liệu. Pha phân cụm có thể đóng vai trò làm công đoạn chính trong một ứng dụng chẳng hạn như phân cụm người dùng website, phân cụm các trạng thái cảm xúc khuôn mặt,… hoặc trong công đoạn tiền xử lý dữ liệu như phân đoạn ảnh, làm sạch dữ liệu.

Bài toán phân cụm đã được nghiên cứu và đề xuất từ cách đây hơn nửa thế kỷ [14]: Thuật toán K-Means 1957, thuật toán phân cụm kiểu phân tầng năm 1967, thuật toán phân cụm dựa trên đồ thị năm 1973, thuật toán phân cụm mờ 1981, thuật toán phân cụm dựa trên mô hình mật độ (DBSCAN) năm 1996 [6]. Với mỗi loại thuật toán đã đề xuất luôn tồn tại những ưu điểm và nhược điểm riêng. Với K-Means đó là kết quả phân cụm phụ thuộc vào các điểm trọng tâm khởi tạo ban đầu, với thuật toán DBSCAN và phương pháp phân cụm thứ bậc là độ phức tạp của thuật toán cao, đối với phương pháp sử dụng đồ thị là khó xác định tham số,… Trong số những thuật toán cơ bản nêu trên, thuật toán K-Means là thuật toán được sử dụng nhiều nhất [12].

Một trong những nghiên cứu thu hút các nhà nghiên cứu trong những năm gần đây là hướng nghiên cứu về thuật toán phân cụm nửa giám sát [1], [2], [3], [4], [8], [10], [11], [13]. Xuất phát từ những thuật toán cơ bản kể trên, giả thiết rằng chúng ta có thêm một số thông tin chẳng hạn như có một số ràng buộc giữa các phần tử trong tập dữ liệu hay đã có trước một số phần tử được gán nhãn sẵn xác định chúng ở cụm nào. Ý tưởng cơ bản là phát triển các thuật toán có thể tích hợp các thông tin bổ trợ nhằm tăng chất lượng của quá trình phân cụm. Hàng loạt các công trình nghiên cứu đã được công bố như đã nhắc đến ở trên.

* *Email: pthientng@gmail.com*

Mục tiêu của bài báo này tập trung vào phương pháp K-Means. Để cải tiến chất lượng của K-Means, năm 2002, Basu và các cộng sự đề xuất phương pháp Seed-Kmeans với giả thiết nếu được cung cấp đầy đủ mỗi cụm một vài các phần tử đã được gán nhãn, các trọng tâm ban đầu ở bước khởi tạo sẽ được sử dụng bởi chính các thông tin đó và sẽ làm cho chất lượng của thuật toán tăng lên [1]. Năm 2009, M. A. Hasan và các cộng sự đề xuất thuật toán ROBIN bằng cách đề xuất sử dụng phương pháp *min-max* để ước lượng các trọng tâm ở bước khởi tạo và cũng làm cho chất lượng phân cụm của K-Means được tăng lên [7]. Trong [15], Yoder và Priebe giới thiệu phương pháp phân cụm nửa giám sát cho K-Means++, trong đó tác giả đã sử dụng phương pháp tìm các trọng tâm theo thuật toán Lloyd và kết hợp với một số dữ liệu đã được gán nhãn trước. Dựa trên ý tưởng của thuật toán này, trong bài báo này chúng tôi đề xuất phương pháp phân cụm nửa giám sát cho K-Means bằng cách sử dụng kết hợp phương pháp min-max với một số dữ liệu đã gán nhãn sẵn ban đầu; kết quả thực nghiệm cho kết quả tốt khi so sánh với các phương pháp cùng loại.

CÁC NGHIÊN CỨU LIÊN QUAN

Trong phần này chúng tôi trình bày ba nghiên cứu liên quan đó là thuật toán K-Means, thuật toán phân cụm nửa giám sát Seed K-Means và thuật toán xác định các điểm khởi tạo cho các trọng tâm ban đầu (thuật toán ROBIN).

Thuật toán K-Means

Thuật toán K-Means là một trong những thuật toán kinh điển của lĩnh vực học máy, thuật toán này được giới thiệu vào những năm 50 của thế kỷ 20 và là một trong 10 thuật toán được sử dụng nhiều nhất trong lĩnh vực khai phá dữ liệu và phát hiện tri thức [12]. Ý tưởng của thuật toán K-Means là dựa trên k trọng tâm tại bước đầu tiên, thuật toán tiến hành gán các điểm vào các trọng tâm gần nhất và lặp lại quá trình trên cho đến khi các trọng tâm không thay đổi được nữa.

Thuật toán K-Means thực hiện đơn giản và có độ phức tạp tính toán là O(k × n). Hình 1 trình bày các bước cơ bản của thuật toán K-Means.

Thuật toán K-Means;

Input: Tập dữ liệu X, số lượng cụm cần phân tách *k*

Output: *k* cụm của X

+ Khởi tạo các trọng tâm cho các cụm ngẫu nhiên.

Repeat

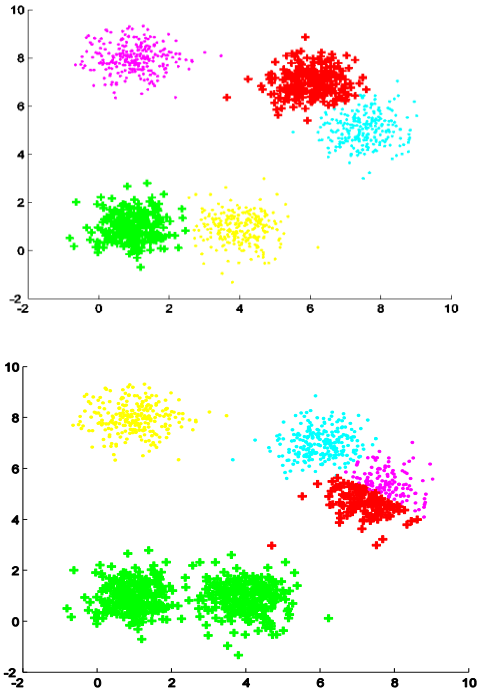
+ Gán các điểm của X vào cụm gần nó nhất

+ Tính toán lại trọng tâm bằng trung bình cộng các điểm trong mỗi cụm tương ứng

Until (Các trọng tâm không thay đổi được nữa)

Hình 1. *Thuật toán K-Means*

Hạn chế cơ bản nhất của thuật toán K-Means là nó phụ thuộc vào sự lựa chọn các trọng tâm nên mỗi lần chạy có thể cho ra kết quả khác nhau (xem Hình 2). Ngoài ra K-Means chỉ phát hiện được các cụm có dạng hình cầu do dựa trên tính chất về khoảng cách. Một số cải tiến của K-Means đã được nghiên cứu như sử dụng hàm hạt nhân, hoặc chia nhỏ tập dữ liệu thành các cụm nhỏ và rồi lại tiến hành hòa nhập chúng lại để được các cụm có hình dạng bất kỳ.



Hình 2. *Minh họa thuật toán K-Means: Hai lần thực hiện chương trình có thể cho ra hai kết quả khác nhau với tập dữ liệu như trên*

Thuật toán Seed K-Means

Trong thuật toán Seed K-Means [1], tác giả đã sử dụng một số phần tử dữ liệu đã được gán nhãn sẵn (gọi là các seed) để xây dựng thuật toán Seed K-Means nhằm vượt qua hạn chế của thuật toán K-Means trong vấn đề lựa chọn các trọng tâm của các cụm – thuật toán K-Means thường có các kết quả khác nhau sau mỗi lần thực hiện vì chúng phụ thuộc nhiều vào vấn đề lựa chọn các trọng tâm ngẫu nhiên đầu tiên. Ý tưởng cơ bản của thuật toán Seed K-Means là sử dụng các seed vào ngay bước khởi tạo các trọng tâm và vì vậy thuật toán sẽ cho ra kết quả tốt hơn chỉ sau duy nhất một lần thực hiện. Thuật toán Seed K-Means yêu cầu mỗi cụm phải có ít nhất một điểm dữ liệu gán nhãn sẵn.

Thuật toán ROBIN

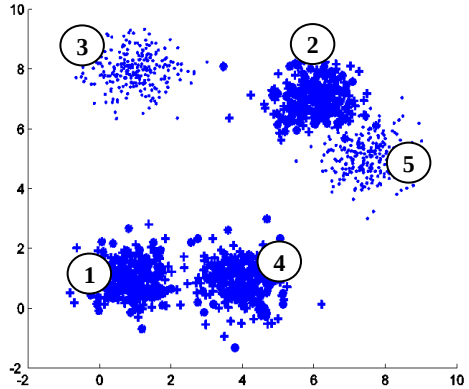
Từ khi K-Means ra đời, có rất nhiều phương pháp tìm kiếm k trọng tâm đầu tiên. Chúng tôi trình bày sau đây thuật toán ROBIN được giới thiệu năm 2009 [7]. Ý tưởng chính của việc lựa chọn các trọng tâm cho K-Means của thuật toán ROBIN là sử dụng phương pháp *min-max*.

Cho tập dữ liệu X, phương pháp min-max cho phép xác định một tập con Y của X sao cho các điểm trong Y là cách nhau xa nhất có thể. Để thực hiện điều này quá trình xây dựng tập Y như sau: trước tiên chọn một điểm *y* bất kỳ làm điểm đầu tiên cho Y, tại mỗi bước tiếp theo chúng ta chọn điểm y_{new} thỏa mãn công thức sau:

$$y_{new} = \arg\max_{x \in X} \left(\min_{y \in Y} d(x, y) \right)$$

trong đó *d(.)* là khoảng cách *Euclidean* trong không gian *n* chiều. Nói một cách tổng quát, điểm y_{new} được chọn sẽ là điểm có khoảng cách xa nhất có thể đối với các điểm đã có trong Y.

Hình 3 minh họa thuật toán min-max với tập dự liệu có 5 cụm. Chúng ta có thể thấy rằng các trọng tâm sẽ được chọn bằng 5 điểm xa nhau nhất có thể. Để đảm bảo các điểm chọn là không phải điểm ngoại lai, trong thuật toán ROBIN có sử dụng thêm hàm LOF [5], để kiểm tra xem một điểm dữ liệu bất kỳ có là điểm ngoại lai hay không. Thuật toán ROBIN đã chỉ ra đạt chất lượng tốt so sánh với các phương pháp cùng loại



Hình 3. *Minh họa phương pháp min-max với 5 điểm khởi tạo*

Thuật toán K-Means*

Trong phần này chúng tôi đề xuất việc kết hợp hai thuật toán Seed K-Means và thuật toán min-max thành thuật toán Seed K-Means*. Thuật toán K-means* đơn giản chỉ là giải quyết trường hợp rất hay gặp trong thực tế đó là các dữ liệu gán nhãn sẵn cho các cụm đôi khi không đầy đủ. Trong trường hợp đó, thuật toán K-Means* sẽ sử dụng phương pháp min-max để tiếp tục tìm kiếm các trọng tâm còn lại. Thuật toán K-means* được trình bày trong hình 4.

Thuật toán K-Means*;

Input: Tập dữ liệu X, số lượng cụm cần phân tách *k, tập các nhãn S={s_1, s_2,…,s_v}.*

Output: *k* cụm của X

+ Khởi tạo các trọng tâm cho các cụm sử dụng tập S.

+ Nếu số lượng các trọng tâm nhỏ hơn k thì sử dụng phương pháp *min-max* để tìm đủ k trọng tâm.

Repeat

+ Gán các điểm của X vào cụm gần nó nhất

+ Tính toán lại trọng tâm bằng trung bình cộng các điểm trong mỗi cụm tương ứng

Until (Các trọng tâm không thay đổi được nữa)

Hình 4. *Thuật toán K-Means**

Kết quả thực nghiệm

Để đánh giá hiệu quả của thuật toán chúng tôi sử dụng 7 tập dữ liệu được lấy từ website UCI Machine Learning như trong bảng 1, trong đó n, m, và k tương ứng là số phần tử trong mỗi tập dữ liệu, số chiều của mỗi tập dữ liệu, và số cụm của mỗi tập dữ liệu [16].

Chúng tôi sử dụng độ đo Rand Index (RI) để đánh giá độ chính xác của các thuật toán phân cụm [9]. Thuật toán RI so sánh kết quả phân cụm với giá trị thật của các nhãn trên thực tế. Với hai cách phân cụm p_1 và p_2 của tập dữ liệu X có n phần tử trong đó p_1 và p_2 chứa các nhãn của các phần tử tương ứng của các cụm, công thức tính giá trị RI như sau:

$$RI = \frac{2(a+b)}{n(n-1)}$$

trong đó a là tổng các cặp xuất hiện cùng một cụm trong p_1 và cũng cùng một cụm trong p_2; b là tổng các cặp xuất hiện trong hai cụm khác nhau trong p_1 và trong p_2. RI có giá trị nằm trong đoạn [0,1]. Giá trị của RI càng lớn thì chất lượng phân cụm càng cao.

Bảng 1. *Các tập dữ liệu cho thực nghiệm*

STT	Tập dữ liệu	n	m	k
1	Iris	150	4	3
2	Protein	116	6	6
3	Soybean	47	34	4
4	Thyroid	215	5	3
5	Haberman	306	3	2
6	Zoo	101	16	7
7	Image	2100	19	7

Để đánh giá kết quả, chúng tôi sử dụng ba thuật toán là min-max K-Means (MMKM) trong đó các trọng tâm cụm được xác định bằng thuật toán ROBIN, thuật toán Seed K-Means (SKM) và thuật toán K-Means* (KM*). Kết quả phân cụm được trình bày trong bảng 2, giá trị RI được chuyển sang dạng %. Với mỗi bộ dữ liệu, chúng tôi đã thực nghiệm với một vài bộ seed khác nhau. Kết quả phân cụm của mỗi tập dữ liệu được lấy trung bình sau 50 lần thực hiện cho mỗi thuật toán.

Từ bảng 2 chúng ta có các nhận xét như sau:

Chất lượng phân cụm của thuật toán MMKM là kém hơn so với thuật toán KM* và SKM. Điều này có thể giải thích rằng với việc xác định các trọng tâm bằng thuật toán min-max, trong một số trường hợp chúng ta vẫn không xác định được các trọng tâm tốt cho thuật toán KM (xem Hình 5).

Thuật toán KM* cho kết quả tương đối tốt so sánh với thuật toán SKM. Thậm chí KM* còn cho kết quả tốt hơn đối với hai tập dữ liệu là Iris và Haberman. Tập dữ liệu Haberman gồm hai lớp là một tập dữ liệu rất khó đối với các thuật toán phân cụm (xem Hình 6). Chúng ta thấy rằng với SKM chất lượng chỉ đạt 52,8%. Đối với MMKM và KM* đạt kết quả cao hơn có thể giải thích rằng khi sử dụng phương pháp min-max hai trọng tâm ban đầu được lựa chọn xa nhau nhất có thể.

Tóm lại việc kết hợp giữa phương pháp min-max và phương pháp sử dụng seed cho kết quả tương đối tốt và quan trọng là có thể áp dụng trong thực tế đối với một số dạng bài toán chúng ta không thể tìm đủ các dữ liệu đã gán nhãn cho các lớp ban đầu.

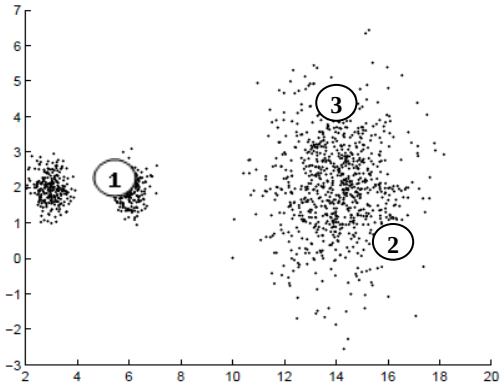
Bảng 2. *Kết quả thực nghiệm so sánh giữa các phương pháp MMKM, KM*, và SKM*

Dữ liệu	MMKM	KM*[%Seed]	SKM
Iris	87.5	**87.9**[33%] **87.9**[66%]	87.6
Protein	75.9	76.8[48%] 77.7[64%] 78.1[80%]	**79.9**
Soybean	83.4	83.3[25%] 83.5[50%] 83.5[75%]	**85.3**
Thyroid	69.9	69.9[33%] 70.1[66%]	**71.5**
Haberman	60.8	**61.6**[50%]	52.8
Zoo	91.3	91.6[28%] 90.4[42%] 91.8[56%] 92.7[70%]	**93.7**
Image	82.1	82.5[33%] 82.8[66%]	**83.2**

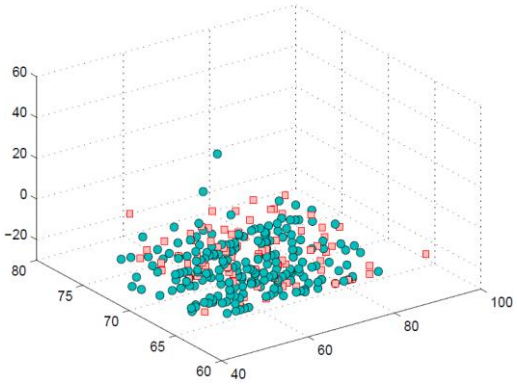
KẾT LUẬN

Trong bài báo này, chúng tôi đã trình bày thuật toán K-Means*, một phương pháp lai giữa hai thuật toán Seed-Kmean và phương pháp xác định các trọng tâm bằng phương pháp min-max. Kết quả thực nghiệm cho thấy

tình hiệu quả trong việc kết hợp hai phương pháp đồng thời K-Means* hoàn toàn có thể ứng dụng trong thực tế trong trường hợp dữ liệu trợ giúp là không đầy đủ. Trong thời gian tới các thuật toán về phân cụm nửa giám sát sẽ tiếp tục được nghiên cứu và triển khai cùng với các ứng dụng thực tế.



Hình 5. *Ví dụ về dữ liệu với 3 cụm mà thuật toán min-max cho kết quả không tốt về việc xác định các trọng tâm: chúng ta thấy rằng cụm thứ 3 có 2 seed trong khi cụm 1 không có seed nào*



Hình 6. *Tập dữ liệu Haberman được hiển thị trong không gian 3 chiều; chúng ta có thể thấy dữ liệu gồm 2 cụm với phần tử của cụm này xuất hiện đan xen vào cụm kia nên rất khó để phân tách các cụm trong trường hợp này*

TÀI LIỆU THAM KHẢO

1. Sugato Basu, Arindam Banerjee, Raymond J. Mooney (2002), "Semi-supervised Clustering by Seeding", *ICML 2002*, pp. 27-34.

2. S. Basu, I. Davidson, and K. L. Wagstaff (2008), *Constrained Clustering: Advances in Algorithms, Theory, and Applications*, Chapman and Hall/CRC Publisher, USA.

3. A. Bar-Hillel, T. Hertz, N. Shental, D. Weinshall (2003), "Learning distance functions using equivalence relations", *Proceedings of 20th International Conference on Machine Learning ICML*, pp.11-18.

4. Amine Bensaid, Lawrence O. Hall, James C. Bezdek, Laurence P. Clarke (1996), "Partially supervised clustering for image segmentation", *Pattern Recognition*, 29(5), pp. 859-871.

5. Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, Jörg Sander (2000), "LOF: Identifying Density-Based Local Outliers", *SIGMOD Conference*, pp. 93-104.

6. Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu (1996), "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", *KDD*, pp. 226-231.

7. Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, Mohammed J. Zaki (2009): "Robust partitional clustering by outlier and density insensitive seeding", *Pattern Recognition Letters*, 30(11), pp. 994-1002.

8. Levi Lelis, Jörg Sander (2009): "Semi-supervised Density-Based Clustering", *ICDM*, pp. 842-847.

9. W. M. Rand (1971), "Objective criteria for the evaluation of clustering methods", *Journal of the American Statistical Association*, 66 (336), pp. 846–850.

10. Kiri Wagstaff, Claire Cardie, Seth Rogers, Stefan Schrödl (2001): "Constrained K-means Clustering with Background Knowledge", *ICML 2001*, pp. 577-584.

11. Xiang Wang, Buyue Qian, Ian Davidson (2014), "On constrained spectral clustering and its applications", *Data Mining and Knowledge Discovery*, 28(1), pp. 1-30.

12. Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus F. M. Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, Dan Steinberg (2008), "Top 10 algorithms in data mining", *Knowl. Inf. Syst*, 14(1), pp. 1-37.

13. Eric P. Xing, Andrew Y. Ng, Michael I. Jordan, Stuart J. Russell (2002): "Distance Metric Learning with Application to Clustering with Side-Information", *NIPS 2002*, pp. 505-512.

14. Rui Xu, Donald C. Wunsch II (2005): "Survey of clustering algorithms", *IEEE Trans. Neural Networks*, 16(3), pp. 645-678.

15. J. Yoder, C.E. Priebe (2017), "Semi-supervised K-Means++", *Journal of Statistical Computation and Simulation*, Vol. 87(13), pp. 2597-2608.

16. http://archive.ics.uci.edu/ml/

SUMMARY

K-MEANS* CLUSTERING ALGORITHM

Cao Ngoc Anh[1], Vu Viet Vu[2], Phung Thi Thu Hien[3*]
[1]*University of Economics and Technical Industries*
[2]*The Information Technology Institute – Vietnam National University, Hanoi*
[3]*University of Technology - TNU*

Recently, semi-supervised clustering has received a great deal of attention. By using backgrond knowledge, either labeled data (seeds) or pairwise constraints (must-link or can-not link) between data objects, semi-supervised clustering algorithms can improve the quality of clustering process. This paper proposes an algorithm for semi-supervised K-Means clustering when the side information is incomplete, namely K-Means*. K-Means* is the combination of seed based K-Means clustering and min-max based initialization K-Means algorithm. Experimemt results on real data sets from UCI show the effectiveness of our algorithm.

Key words: *K-Means clustering, semi-supervised clustering, seeds, min-max method, UCI data.*

Ngày nhận bài: 02/8/2017; Ngày phản biện: 15/8/2017; Ngày duyệt đăng: 30/9/2017

[*] *Email: pthientng@gmail.com*