TỔNG QUAN CÁC CÁCH TIẾP CẬN TRONG HỌC LUẬT QUYẾT ĐỊNH

Nguyễn Đức Cường

Khoa Công nghệ thông tin, Trường Đại học Ngoại ngữ - Tin học TP. HCM
cuong.nd@huflit.edu.vn

TÓM TẮT — Luật quyết định dạng "IF điều_kiện THEN thực_thi" là một trong những loại tri thức được sử dụng phổ biến nhất của con người do tính dễ hiểu và dễ vận dụng. Học luật (Rule Induction), một hướng nghiên cứu được quan tâm trong ngành Khoa học dữ liệu, nhằm trích xuất ra tập luật quyết định từ các tập dữ liệu huấn luyện. Bài báo tập trung vào việc hệ thống hóa các cách tiếp cận chính trong Học luật, đồng thời nêu ra điểm mạnh và yếu của các phương án trên. Bài báo cũng chỉ ra thách thức và các hướng nghiên cứu đang được quan tâm trong Học luật.

Từ khoá— Data Science, Data Mining, Rule Induction, Covering method.

I. GIỚI THIỆU

Luật quyết định dạng "IF điều_kiện THEN thực_thi" là một trong những loại kiến thức mà con người sử dụng thường xuyên nhất, dễ hiểu nhất và sử dụng sớm nhất. Con người thu thập kiến thức về luật quyết định thông qua việc dạy dỗ của cha mẹ hay thầy cô, hoặc tự học hỏi thông qua các trải nghiệm của bản thân. Trải qua suốt cuộc đời của mình, con người tự học rất nhiều luật quyết định cũng như điều chỉnh các luật đã biết.

Bài toán phân lớp (Classification) là một bài toán phổ biến nhất trong khai phá dữ liệu (Data Mining) và học máy (Machine Learning), trong đó, các bộ phân loại (Classifier) sẽ phân một đối tượng dữ liệu chưa biết vào một trong những lớp đã được xác định. Chất lượng của một bộ phân loại được đánh giá qua nhiều tiêu chí, trong đó, độ chính xác (accuracy) là tiêu chuẩn được sử dụng phổ biến nhất.

Học luật quyết định (Rule Induction), hay còn gọi là giải thuật bao phủ (Covering Method), là một trong những kỹ thuật được sử dụng phổ biến trong bài toán phân lớp. Học luật tập trung vào việc trích xuất các luật quyết định dạng "IF điều_kiện THEN thực_thi" từ tập dữ liệu học. Với một tập dữ liệu học, được gọi là tập dữ liệu huấn luyện (training data set), trong đó mỗi phần tử sẽ được gán cho 1 lớp, giải thuật học luật có thể học ra các tập luật quyết định (Ruleset) có thể phân lớp cho các dữ liệu chưa biết. Tập dữ liệu dùng để đánh giá hiệu quả của tập luật đã được học ra, được gọi là tập dữ liệu thử nghiệm (test data set). Tập luật nếu có kích thước nhỏ thì rất dễ để người dùng hiểu và kiểm soát được. Tuy nhiên, khi tập luật có kích thước lớn, chất lượng của tập luật rất khó kiểm soát, người dùng rất khó chọn luật nào đúng để vận dụng.

Học luật quyết định được áp dụng thành công trong nhiều lĩnh vực khác nhau. Trong y học, học luật quyết định hỗ trợ bác sĩ xác định bệnh tự kỷ [1] ở các bệnh nhân. Trong sinh học, học luật quyết định giúp phân loại sự đa dạng sinh học [2] hay sử dụng trong hệ thống sinh học về ung thư [3]… Học luật quyết định cũng được dùng để xác định qui luật trong các hệ thống phát hiện xâm nhập (Intrusion Detection Systems) vào mạng máy tính [4] hay dự đoán các báo động nghiêm trọng trong mạng viễn thông [5].

Bài báo này nhằm đưa ra một bản đánh giá tổng quan về các giải thuật học luật quyết định. Phần còn lại của bài báo này được trình bày theo bố cục sau: phần II trình bày về các cách tiếp cận chính trong học luật; phần III trình bày về các nghiên cứu đang được quan tâm trong học luật và phần IV là kết luận.

II. CÁC CÁCH TIẾP CẬN CHÍNH TRONG HỌC LUẬT

A. HỌC TỪNG LUẬT VỚI TẬP HUẤN LUYỆN SUY GIẢM DẦN

Giải thuật tóm tắt: xem Hình 1.

Các giải thuật điển hình: AQ [6], CN2 [7], GuideR [8].

Một số nhận xét về cách tiếp cận này:

- Đặc điểm của giải thuật trong cách tiếp cận này là sử dụng một chu trình trên tập dữ liệu huấn luyện hiện thời Dcurrent để tuần tự sinh ra từng luật cho đến khi điều kiện kết thúc của giải thuật được thỏa. Điều kiện kết thúc thông dụng là kích thước tập $D_{current}$ chứa các phần tử dữ liệu chưa được bao phủ còn lớn hơn không. Để xử lý nhiễu hay học quá mức (overfitting), một số giải thuật trong cách tiếp cận này sẽ tỉa nhánh (prunning) bằng cách sử dụng thêm một số điều kiện như "luật mới sinh ra bởi *Induce_One_RuleA* có chất lượng quá thấp" hay "độ bao phủ của luật tốt nhất được sinh ra bởi *Induce_One_RuleA* quá nhỏ".
- Tập dữ liệu huấn luyện hiện thời $D_{current}$ có kích thước giảm dần qua các chu trình học luật. Ưu điểm việc kích thước giảm dần này là giải thuật sẽ học luật ngày càng nhanh hơn đối với các chu trình lặp sau trong giải thuật chính.

- Khuyết điểm của việc tập dữ liệu huấn luyện giảm dần là độ đo đánh giá của các luật học được trong các chu trình sau sẽ có giá trị đánh giá trên tập $D_{current}$ chứ không phải là tập huấn luyện đầy đủ D. Khi chỉ đánh giá luật đang học trên $D_{current}$ sẽ dẫn đến việc đánh giá chất lượng của luật đang học sẽ mang tính tức thời, hiển nhiên là có giá trị khác biệt so với việc đánh giá chất lượng của luật đó trên tập dữ liệu huấn luyện toàn cục là tập *D*.
- Giá trị đánh giá chất lượng của một luật cần phải đánh giá lại vì việc đánh giá bên trong một chu trình học luật được tính dựa trên tập dữ liệu huấn luyện hiện thời $D_{current}$ (có kích thước giảm dần theo các chu trình) nên sẽ có giá trị khác khi được tính trên toàn bộ tập dữ liệu huấn luyện *D*. Khi sử dụng tập luật trong thực tế, luôn luôn tồn tại các đối tượng dữ liệu được bao phủ bởi nhiều luật có lớp khác nhau, khi đó, việc phân lớp sẽ dựa trên chất lượng của các luật này. Nếu sử dụng giá trị đánh giá chất lượng không chính xác (do chỉ đánh giá trên tập tạm thời là $D_{current}$), hiệu quả của tập luật có thể không chính xác. Việc đánh giá lại chất lượng của một luật sau khi học xong tập luật chưa được nghiên cứu sâu.
- Khi vận dụng Học luật trên các tập dữ liệu huấn luyện có nhiều lớp, thứ tự của các lớp được học trong giải thuật con "Học một luật" (*Induce_One_RuleA*) hoặc thứ tự phần tử chưa được bao phủ được sử dụng như hạt giống để tạo tập luật tạm thời S chưa được nghiên cứu sâu.

```
Procedure Induce_RulesetA        // Học từng luật với tập huấn luyện suy giảm
Input: Training data set D
Output: Ruleset RS
Begin
  RS ← ∅;
  Current training data set D_current ← D
  While D_current is not empty Do
    BestRule ← Induce_One_RuleA(D_current , parameters) ;
    D_current ← Remove covered samples in D_current ;
    RS = Add BestRule to RS;
  End While
  Return RS
End
```

(a) Giải thuật chính

```
Procedure Induce_One_RuleA(D_current, D, parameters)
Input: current training data set D_current; other parameters
Output: the best rule BestRule
Begin
  From D_current, form the set S of rule candidates, each of that covers at least one
    sample in D_current;
  Evaluate rule candidates in S on D_current;
  Find the best rule candidate in S as the BestRule;
    Return BestRule
End
```

(b) Giải thuật Học một luật (Induce_One_RuleA)

Hình 1. *Giải thuật Học luật theo cách tiếp cận "Học từng luật với tập huấn luyện suy giảm dần"*

B. HỌC TỪNG LUẬT TRÊN TẬP HUẤN LUYỆN BAO GỒM CÁC ĐỐI TƯỢNG CHƯA ĐƯỢC BAO PHỦ

Giải thuật tóm tắt: xem Hình 2.

Các giải thuật điển hình: RULE 3+ [9], RULES 6 [10], Rules-Machine Learning [1].

Một số nhận xét về cách tiếp cận này:

- So với cách tiếp cận "Học từng luật với tập huấn luyện suy giảm", tập dữ liệu huấn luyện sẽ được giữ nguyên nhưng đánh dấu bằng thuộc tính bao phủ/không bao phủ (covered/uncovered).
- Tập dữ liệu bao gồm các đối tượng không bao phủ $D_{uncovered}$ có kích thước giảm dần trong quá trình học và chỉ được sử dụng để sinh ra tập ứng cử viên.

- Giá trị đánh giá chất lượng luật được tính trên $D_{uncovered}$ nên có giá trị bất biến trong cả quá trình học và sử dụng tập luật. Vì vậy, giá trị đánh giá chất lượng của một luật là chính xác và không cần phải tính lại, nên đã khắc phục vấn đề này của cách tiếp cận "Học từng luật với tập huấn luyện suy giảm dần".
- Khuyết điểm của cách tiếp cận này là sinh ra khá nhiều luật dư thừa, cần sử dụng thêm các giá trị phụ như số đối tượng chưa đánh dấu nằm trong vùng bao phủ của luật.
- Thứ tự lựa chọn phần tử chưa được bao phủ từ tập $D_{uncovered}$ nhằm sử dụng như một hạt giống để tạo tập luật tạm thời *S* chưa được nghiên cứu tương xứng.

```
Procedure Induce_RulesetB      // Học từng luật trên tập huấn luyện bao gồm các
                               // đối tượng chưa được bao phủ
Input: Training data set D
Output: Ruleset RS
Begin
  RS ← ∅;
  Uncovered training data set D_uncovered ← D;
  While D_uncovered is not empty Do
    BestRule ← Induce_One_RuleB(D_uncovered, D, parameters) ;
    Mark the samples in D covered by BestRule as "covered" ;
    RS = Add BestRule to RS;
  End While
  Return RS
End
```

(a) Giải thuật chính

```
Procedure Induce_RulesetB      // Học từng luật trên tập huấn luyện bao gồm các
                               // đối tượng chưa được bao phủ
Input: Training data set D
Output: Ruleset RS
Begin
  RS ← ∅;
  Uncovered training data set D_uncovered ← D;
  While D_uncovered is not empty Do
    BestRule ← Induce_One_RuleB(D_uncovered, D, parameters) ;
    Mark the samples in D covered by BestRule as "covered" ;
    RS = Add BestRule to RS;
  End While
  Return RS
End
```

(b) Giải thuật Học 1 luật (Induce_One_RuleB)

Hình 2. *Giải thuật Học luật theo cách tiếp cận "Học từng luật trên tập huấn luyện bao gồm các đối tượng chưa được bao phủ"*

C. HỌC ĐỒNG THỜI TOÀN BỘ TẬP LUẬT QUYẾT ĐỊNH

Giải thuật tóm tắt: xem Hình 3

Các giải thuật điển hình: CWS [11], RISE [12], RULES-A [13], RIAS [4], LBRI [14].

Một số nhận xét về cách tiếp cận này:

- Thay vì học tuần tự từng luật, các giải thuật trong cách tiếp cận này sẽ học toàn bộ tập luật. Ban đầu, tập luật hiện thời là tập rỗng và sẽ thay đổi thường xuyên trong quá trình học, trong đó một luật mới có chất lượng tốt sẽ được thêm vào tập luật, còn một luật đã có nhưng chất lượng chưa tốt có thể bị loại ra khỏi tập luật hiện thời.
- Trong suốt quá trình lặp để cải thiện tập luật, một luật có thể thêm vào, thay đổi hay xóa đi khỏi tập luật hiện thời. Khi thêm vào 1 luật mới hay xóa đi 1 luật đang có, giải thuật phải kiểm tra sự thay đổi của toàn bộ tập luật còn lại. Điều này làm tăng độ phức tạp của chu trình chính từ tuyến tính do học tuần tự

từng luật sang bình phương của số luật trong tập luật hiện thời. Vì vậy, độ phức tạp giải thuật học theo cách tiếp cận này khá lớn nên giải thuật chạy chậm.

- Mối tương quan giữa các luật trong cùng 1 lớp, cũng như giữa các luật thuộc các lớp khác nhau, rất phức tạp và đã không được giải quyết tương xứng.

```
Procedure Induce_RulesetB      // Học từng luật trên tập huấn luyện bao gồm các
                               // đối tượng chưa được bao phủ
Input: Training data set D
Output: Ruleset RS
Begin
  RS ← ∅;
  Uncovered training data set D_uncovered ← D;
  While D_uncovered is not empty Do
    BestRule ← Induce_One_RuleB(D_uncovered, D, parameters) ;
    Mark the samples in D covered by BestRule as "covered" ;
    RS = Add BestRule to RS;
  End While
  Return RS
End
```

Hình 3. *Giải thuật Học luật theo cách tiếp cận "Học toàn bộ tập luật quyết định"*

D. HỌC LUẬT DỰA TRÊN KẾT QUẢ CỦA GIẢI THUẬT HỌC MÁY KHÁC

1. HỌC LUẬT DỰA TRÊN KẾT QUẢ CỦA GIẢI THUẬT HỌC CÂY QUYẾT ĐỊNH (DECISION TREE)

```
Procedure Induce_RulesetD      // Học luật dựa trên kết quả của giải thuật Học cây
                               // quyết định
Input: Training data set D
Output: Ruleset RS
Begin
  RS ← ∅;
  Construct a Decision Tree DT on D;
  For each path from the root node to a leaf node of the constructed DT  Do
    Construct a new rule R by the composition of decisions in that path;
  End                                                              While
  Return RS
End
```

Giải thuật tóm tắt: Hình 4.

Hình 4. *Giải thuật Học luật theo cách tiếp cận "Học luật dựa trên kết quả của giải thuật Học cây quyết định"*

Các giải thuật điển hình: [15].

Một số nhận xét về cách tiếp cận này:

- Giải thuật học cây quyết định [16] có tiêu chí xây dựng khác biệt với giải thuật học luật.
- Chưa có thực nghiệm nhằm đánh giá đầy đủ về chất lượng của cách tiếp cận này so với việc học luật quyết định trực tiếp từ tập dữ liệu.

2. HỌC LUẬT DỰA TRÊN KẾT QUẢ CỦA GIẢI THUẬT LUẬT LIÊN KẾT (ASSOCIATION RULES)

Giải thuật tóm tắt: Hình 5

Một số nhận xét về cách tiếp cận này:

- Giải thuật học luật liên kết [17-18] là giải thuật học không giám sát (unsupervised learning) và có mục tiêu hoàn toàn khác với giải thuật học luật. Thêm vào đó, các giải thuật học luật liên kết cần có nhiều tham số đầu vào như độ hỗ trợ (support) tối thiểu hay độ tin cậy tối thiểu. Khi sử dụng thực tế, các tham số này được điều chỉnh dựa theo đặc tính phân bố của tập dữ liệu nhằm tạo ra các luật liên kết có ý nghĩa thực tế cao nhất và được thẩm định/lựa chọn dựa trên độ hữu ích cho ứng dụng thực tế bởi các

chuyên gia về kinh doanh. Việc điều chỉnh các tham số nêu trên trong bối cảnh để sinh ra tập luật quyết định theo cách tiếp cận này có tác động lớn đến hiệu quả phân lớp [19].

- Phương pháp phổ biến nhất trong cách tiếp cận này là phân lớp dựa trên liên kết (Classification Based on Associations - CBA) [19-20]. CBA duyệt qua tập dữ liệu huấn luyện nhiều lần để tìm ra tập phổ biến (frequent itemsets) nhằm sinh ra các tập phổ biến dài nhất. Tập các luật liên kết vượt qua giá trị nền về độ hỗ trợ và độ tin cậy sẽ được xem xét để thêm vào tập luật quyết định. Việc khảo sát các luật liên kết được thực hiện dựa trên heuristic như biến thiên các giá trị của độ hỗ trợ tối thiểu, độ tin cậy tối thiểu và giá trị nền của độ dài tối đa tiền đề của một luật nhằm học ra một tập luật có số luận xác định bởi người dùng.

```
Procedure Induce_RulesetD        // Học luật dựa trên kết quả của giải thuật Học cây
                                 // quyết định
Input: Training data set D
Output: Ruleset RS
Begin
  RS ← ∅;
  Construct a Decision Tree DT on D;
  For each path from the root node to a leaf node of the constructed DT  Do
    Construct a new rule R by the composition of decisions in that path;
  End                                                                  While
  Return RS
End
```

Hình 5. *Giải thuật Học luật theo cách tiếp cận "Học luật dựa trên kết quả của giải thuật Luật liên kết (Association Rules)"*

III. CÁC NGHIÊN CỨU ĐANG ĐƯỢC QUAN TÂM TRONG HỌC LUẬT

Luật quyết định được xây dựng trên một cụm dữ liệu có phân bố chặt (condensed) với tỷ lệ cao các phần tử có cùng một lớp sẽ có chất lượng tốt. Việc xác định các cụm thường được giải quyết bằng các giải thuật phân cụm dữ liệu (Data Clustering). Tuy nhiên các giải thuật phân cụm thường gom các phần tử gần nhau vào cùng 1 cụm nhưng các phần tử này lại có thể thuộc về nhiều lớp, đây là đặc tính chung của phương thức học không giám sát (unsupervised learning). Vì vậy, việc sử dụng các kỹ thuật Phân cụm bán giám sát (Semi-supervised Data Clustering) có thể tìm ra các cụm dữ liệu của cùng một lớp để khởi tạo các luật quyết định tốt cho các giải thuật học luật quyết định.

Dữ liệu huấn luyện thường bị thiếu xót, với sự xuất hiện của dữ liệu bị mất (missing value). Thông thường, trong giai đoạn tiền xử lý dữ liệu, các dữ liệu bị mất sẽ được thay thế bằng các dữ liệu được ước lượng được tính bởi các phương pháp suy đoán dữ liệu (data imputation methods). Tiêu chí của các phương pháp suy đoán dữ liệu thường không dựa vào các phương pháp khai thác dữ liệu, vì thế việc thay thế giá trị bị mất bằng một giá trị suy đoán sẽ làm giảm hiệu quả của giải thuật phân lớp khi học trên tập dữ liệu huấn luyện đã được làm sạch. Do tính đặc thù của bản thân của các giải thuật Học luật, việc xử lý trực tiếp dữ liệu bị mất ngay trong quá trình học sẽ mang lại hiệu quả cao hơn. Giải thuật CN2 đã được vận dụng cách tiếp cận này trong xử lý trực tiếp dữ liệu bị mất và độ chính xác của giải thuật CN2 đã được cải thiện [21]. Các giải thuật học luật khác nên được khảo sát và vận dụng cách tiếp cận này để nâng cao hiệu quả phân lớp.

Đối với bài toán phân lớp, việc mất cân bằng trong phân bố các lớp gây khó khăn cho việc tìm ra bộ phân lớp hiệu quả. Đã có những nghiên cứu về việc luật phân lớp kết hợp [22] trong việc xử lý bài toán dữ liệu mất cân bằng này. Một hướng khác là điều chỉnh trực tiếp trên các giải thuật học luật. Đối với cách tiếp cận "học từng luật với tập huấn luyện suy giảm dần", phần "Giải thuật học 1 luật (*Induce_One_RuleA*)" có thể xem xét thay đổi độ ưu tiên cho các ứng viên luật dựa trên phân bố của của các lớp. Đối với cách tiếp cận "Học từng luật trên tập huấn luyện bao gồm các đối tượng chưa được bao phủ", phần "giải thuật học 1 luật (*Induce_One_RuleB*)" cũng có thể xem xét thay đổi độ ưu tiên cho các ứng viên luật dựa trên phân bố của của các lớp. Tuy nhiên, cần có các nghiên cứu cụ thể để nâng cao tính hiệu quả.

IV. KẾT LUẬN

Bài báo đã đưa ra một bản đánh giá tổng quan về các giải thuật học luật quyết định dựa trên 4 cách tiếp cận chính. Các khuyết điểm của từng cách tiếp cận cũng được phân tích và nêu ra, từ đó, có thể đưa ra các phương án tăng cường chất lượng của các giải thuật đã có. Bài báo cũng nêu ra một số hướng khác nhằm nâng cao hiệu quả của các giải thuật trong học luật quyết định.

V. TÀI LIỆU THAM KHẢO

[1] Thabtah, F., & Peebles, D. (2020). A new machine learning model based on induction of rules for autism detection. Health informatics journal, 26(1), 264-286.

[2] Swe, S. M., & Sett, K. M. (2019). Approaching rules induction: CN2 algorithm in categorizing of biodiversity. International Journal of Trend in Scientific Research and Development, 3(4), 1581-1584.

[3] Scala, G., Federico, A., Fortino, V., Greco, D., & Majello, B. (2020). Knowledge generation with rule induction in cancer omics. International journal of molecular sciences, 21(1), 18.

[4] Ashaba, A. A., Mirembe, D. P., Tumusiime, R., & Ogenrwot, D. (2019). A Rule Induction Attribution Selection Algorithm for Intrusion Detection Systems. International Journal of Technology and Management, 4(1), 14-14.

[5] Wrench, C., Stahl, F., Di Fatta, G., Karthikeyan, V., & Nauck, D. (2019, November). A rule induction approach to forecasting critical alarms in a telecommunication network. In 2019 International Conference on Data Mining Workshops (ICDMW) (pp. 480-489). IEEE.

[6] Kaufman, K. A., & Michalski, R. S. (1999). "Learning in an Inconsistent World: Rule Selection in AQ18". Reports of the Machine Learning and Inference Laboratory, MLI 99-2, George Mason University, Fairfax, VA.

[7] Clark, P., & Boswell, R. (1991). Rule induction with CN2: Some recent improvements. In European Working Session on Learning (pp. 151-163). Springer, Berlin, Heidelberg.

[8] Sikora, M., Wróbel, Ł., & Gudyś, A. (2019). GuideR: A guided separate-and-conquer rule learning in classification, regression, and survival settings. Knowledge-Based Systems, 173, 1-14.

[9] AlMana, A. M., & Aksoy, M. (2014). An overview of inductive learning algorithms. International Journal of Computer Applications, 88(4), 20-28.

[10] Pham, D. T., & Afify, A. A. (2005). RULES-6: A simple rule induction algorithm for handling large data sets. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 219(10), 1119-1137.

[11] Domingos, P. (1994). The RISE system: Conquering without separating. In Proceedings of Sixth International Conference on Tools with Artificial Intelligence. TAI 94 (pp. 704-707). IEEE.

[12] Domingos, P. (1996). Unifying instance-based and rule-based induction. Machine Learning, 24(2), 141-168.

[13] Nguyen, D. C. (2004). Flexible information management strategies in machine learning and data mining. PhD thesis, Cardiff University (United Kingdom).

[14] Ibrahim, M. H., & Hacibeyoglu, M. (2020). A novel switching function approach for data mining classification problems. Soft Computing, 24(7), 4941-4957.

[15] Kutschenreiter-Praszkiewicz, I. (2019). Decision Rule Induction Based on the Graph Theory. Application of Decision Science in Business and Management, 91.

[16] Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1(1), 81-106.

[17] Agrawal, R., Imieliński, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD international conference on Management of Data (pp. 207-216).

[18] Han, J., Pei, J., & Yin, Y. (2000). Mining frequent patterns without candidate generation. ACM sigmod record, 29(2), 1-12.

[19] Kliegr, T., & Kuchar, J. (2019). Tuning Hyperparameters of Classification Based on Associations (CBA). In ITAT (pp. 9-16).

[20] Liu, B., Hsu, W., & Ma, Y. (1998, August). Integrating classification and association rule mining. In Proceedings of KDD'98, Vol. 98, pp. 80-86.

[21] Nguyen, C. D., Tran, P. T., & Thai, T. T. T. (2018). Handling Missing Values for the CN2 Algorithm. In Context-Aware Systems and Applications, and Nature of Computation and Communication (pp. 226-234). Springer, Cham.

[22] Nguyen, T.T.L, Tran, T.M.T. and Giang, H.C. (2016). Khai thác Luật Phân lớp Kết hợp trên cơ sở dữ liệu mất cân bằng về lớp. Kỷ yếu Hội nghị Khoa học Quốc gia lần thứ IX "Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR'9)"; Cần Thơ, ngày 4-5/8/2016.

REVIEW OF APPROACHES IN RULE INDUCTION

Nguyen Duc Cuong

ABSTRACT - Decision Rule as “IF condition THEN action” is a knowledge type that is popularly used by human being due to its ease to understanding and application. Rule Induction, a research direction in the Data Science, induces a rule set from a training dataset. The paper focuses on reviewing the main approaches in Rule Induction and their advantages as well as disadvantages. The paper also states current challenges and research trends in Rule Induction.

Nguyễn Đức Cường ("Cuong Duc Nguyen" hay “Duc-Cuong Nguyên”) nhận bằng Tiến sĩ Kỹ thuật tại đại học Cardiff, Anh, vào năm 2004. Từ năm 2006 đến năm 2014, Cường là Trưởng Khoa của khoa Khoa học và Kỹ thuật Máy tính của trường Đại học Quốc tế - Đại học Quốc gia TPHCM. Từ 2017 đến 2018, Cường là Trưởng Khoa CNTT của Đại học Nguyễn Tất Thành, TPHCM. Từ tháng 2018 đến nay, Cường là Trưởng Khoa CNTT của Đại học Ngoại ngữ - Tin học, TPHCM. Các lĩnh vực nghiên cứu chính của Cường: Học luật quyết định, Phân cụm dữ liệu và Liên kết thực thể.