ỨNG DỤNG LUẬT KẾT HỢP ĐỂ GIẢI QUYẾT BÀI TOÁN KHỬ NHẬP NHẰNG NGHĨA TỪ TIẾNG VIỆT

Đinh Phú Hùng
Trường Đại học Thủy lợi. Email: hungdp@wru.edu.vn

1. GIỚI THIỆU

Trong tất cả các ngôn ngữ, tồn tại những từ mà nó có thể được dịch theo nhiều nghĩa khác nhau phụ thuộc vào ngữ cảnh mà nó xuất hiện. Đặc biệt, với một từ mà từ loại của nó đã được xác định, nếu nó có nhiều hơn một nghĩa thì được gọi là từ đa nghĩa. Bài toán khử nhập nhằng nghĩa từ (WSD) chính là việc đi xác định nghĩa phù hợp nhất trong tất cả các nghĩa của từ đa nghĩa mà từ này nằm trong một ngữ cảnh xác định. Ví dụ, đối với động từ "phát" trong tiếng Việt, chúng ta hãy xem 2 ví dụ sau:

- Anh ta đang "**phát**" cỏ ngoài bờ ao.
- Học sinh được "**phát**" quyển vở mới.

Động từ "phát" trong hai câu trên rõ ràng mang hai nghĩa khác biệt. Việc giải quyết tốt bài toán này sẽ nâng cao được hiệu quả cho các bài toán khác của xử lý ngôn ngữ tự nhiên.

Cho đến nay, mặc dù các phương pháp tiếp cận dựa trên học máy thống kê áp dụng giải quyết cho bài toán WSD đã cho thấy được những ưu điểm khi so sánh với các phương pháp khác. Tuy nhiên, các mô hình học máy thống kê tạo ra các phân lớp mà theo quan sát của chúng tôi là có thể khó hiểu và không dễ chỉnh sửa ngay trên mô hình bởi tri thức con người. Chính vì vậy, chúng tôi đề xuất việc sử dụng các luật kết hợp khai phá được từ những ngữ cảnh của từ nhập nhằng để tạo ra phân lớp dựa trên luật mà có thể dễ dàng hiểu và chỉnh sửa bởi tri thức con người.

Bài báo này được tổ chức thành sáu phần bao gồm cả phần giới thiệu. Trong phần 2, chúng tôi trình bày một số kiến thức cơ bản về luật kết hợp và phân lớp dựa trên luật kết hợp. Sau đó, chi tiết về mô hình khử nhập nhằng nghĩa từ dựa trên khai phá luật kết hợp sẽ được trình bày trong phần 3. Trong phần 4, chúng tôi trình bày cách lựa chọn đặc trưng cho các mô hình sử dụng trong bài báo. Chuẩn bị dữ liệu và kết quả thực nghiệm được chúng tôi trình bày trong phần 5. Cuối cùng, kết luận về bài báo được chúng tôi trình bày trong phần 6.

2. KIẾN THỨC CƠ SỞ

Trong phần này, chúng tôi giới thiệu các kiến thức cơ sở về khai phá luật kết hợp và phân lớp dựa trên luật kết hợp.

2.1. Khai phá luật kết hợp

Khai phá luật kết hợp được đề xuất bởi Agrawal[1], như là một phương pháp mà ban đầu dùng để phân tích nhu cầu của thị trường. Khai phá luật kết hợp là khám phá mối quan hệ giữa các mục (item) trong một tập dữ liệu.

Cho $T = \{T_1, T_2, ..., T_n\}$ là cơ sở dữ liệu chứa tập các giao dịch, cho $I = \{I_1, I_2, ..., I_n\}$ là tập của các mục có trong T. Mỗi giao dịch T_i chứa một tập các mục $(T_i \subseteq I)$. Một luật kết hợp có dạng X→Y, trong đó $X, Y \in I$ và $X \cap Y = \phi$. X, Y là tập của các mục, X được gọi là tiền đề, Y được gọi là kết quả. Một luật kết hợp có hình thức như trên chỉ ra rằng, sự xuất hiện của các mục có trong X thường kéo theo sự xuất hiện của các mục có trong Y. Các luật kết hợp mạnh sẽ được đánh giá thông qua 2 tham số cơ bản là độ hỗ trợ và độ tin cậy.

2.2. Phân lớp dựa trên luật kết hợp

Phân lớp dựa trên luật kết hợp hay còn được gọi là phân lớp kết hợp chính là việc ứng dụng các luật kết hợp để giải quyết cho bài toán phân lớp. Một luật kết hợp dùng cho phân lớp có dạng: $x \rightarrow y_i$ trong đó $X \subseteq I$, và yi ∈ Y với Y là tập tất cả các nhãn. Vấn đề của khai phá các luật kết hợp dùng cho phân lớp là tạo ra các luật có dạng trên nhưng phải thỏa mãn về ràng buộc các ngưỡng tối thiểu về độ tin cậy nhỏ nhất (minConf) và độ hỗ trợ nhỏ nhất (minSup). Cho đến nay, một số thuật toán phân lớp dựa trên luật kết hợp đã được đề xuất như: CBA [2], CMAR [3], CPAR [4]... Nhìn chung, các giải thuật này có thể phân vào hai loại chính: Phân lớp dựa trên đơn luật kết hợp và phân lớp dựa trên đa luật kết hợp. Mỗi giải thuật phân lớp đó đều được chia ra làm 2 giai đoạn cơ bản: Khai phá luật kết hợp và phân lớp.

3. MÔ HÌNH KHỬ NHẬP NHẰNG

Cho một từ đa nghĩa w, gọi $S = \{s_1, s_2, \ldots, s_m\}$ là tập các nghĩa có thể có của từ w có trong từ điển. Giả sử C là tập các ngữ cảnh của từ đa nghĩa w có trong tập huấn luyện $(C= \{c_1, c_2, \ldots, c_n\})$, trong đó mỗi ngữ cảnh c_i sẽ được biểu diễn bởi 1 tập các đặc trưng $F_i = \{f_{i1}, f_{i2}, \ldots, f_{ik}\}$ tương ứng (chi tiết cách lựa chọn đặc trưng được trình bày trong mục IV). Tập $S = \{F_1, F_2, \ldots, F_m\}$ là tập chứa tất cả các đặc trưng của tập huấn luyện C.

Để có thể sử dụng được giải thuật khai phá các luật kết hợp dùng cho phân lớp từ những ngữ cảnh của từ nhập nhằng, cần có một số thay đổi sau: Thứ nhất, ngữ cảnh của các từ nhập nhằng C tương đương với cơ sở dữ liệu giao dịch T (mỗi ngữ cảnh của từ nhập nhằng c_i đóng vai trò như một giao dịch T_i). Thứ hai, những đặc trưng trích rút được từ ngữ cảnh c_i đóng vai trò như các mục (item). Tập tất cả các nghĩa của từ nhập nhằng đóng vai trò là tập lớp nhãn Y.

4. LỰA CHỌN ĐẶC TRƯNG

Ngữ cảnh là phương tiện duy nhất để xác định nghĩa phù hợp của từ đa nghĩa. Việc lựa chọn đặc trưng đóng một vai trò rất quan trọng đối với các mô hình phân lớp.

Giả sử w là từ đa nghĩa cần khử nhập nhằng và $S = \{s_1, s_2, \ldots, s_m\}$ là tập các nghĩa có thể có của từ w có trong từ điển. Giả sử ngữ cảnh W của từ đa nghĩa w được biểu diễn như sau:

$$W = \{\ldots w_{-3}, w_{-2}, w_{-1}, w_0, w_1, w_2, w_3 \ldots\}$$

Theo cách biểu diễn như trên, W là ngữ cảnh của từ đa nghĩa w bên trong một cửa sổ (-3, +3), trong đó w_0 chính là từ đa nghĩa. Ứng với mỗi giá trị của i trong đoạn [-3, +3], w_i chính là từ xuất hiện ở vị trí thứ i có liên quan đến từ w_0.

Theo nghiên cứu của Ide [5], một số loại đặc trưng có hiệu quả đối với bài toán WSD có thể kể đến như: Túi từ, cụm từ, thứ tự từ....Dựa trên các nghiên cứu [6][7] về việc lựa chọn đăc trưng hiệu quả cho bài toán WSD chúng tôi đề xuất sử dụng loại đặc trưng Túi từ. Đây là loại đặc trưng có dạng sau:

$F_1 (l, r) = \{w_l, \ldots, w_{+r}\}$. Nó là tập tất cả các từ trong một cửa sổ quanh từ nhập nhằng w_0, từ vị trí $-l$ đến vị trí $+r$. Dựa vào đặc điểm là các từ càng nằm xa từ nhập nhằng cả về 2 phía thì càng ít ý nghĩa và ít có ảnh hưởng đến từ nhập nhằng. Do đó, chúng tôi đề xuất chọn $l = r-3$. Khi đó, chúng tôi thu được loại đặc trưng $F_1 (-3, +3)$ bao gồm 7 phần tử (đặc trưng) như sau:

$$F_1 (-3, +3) = \{w_{-3}, w_{-2}, w_{-1}, w_0, w_1, w_2, w_3\}$$

Như vậy, đặc trưng này sẽ sử dụng được chúng tôi sử dụng trong các giải thuật phân lớp dựa trên luật như: Decision list [8], Decision tree [9] [10], CMAR.

5. THỰC NGHIỆM

5.1. Chuẩn bị dữ liệu

Đối với ngôn ngữ tiếng Anh, các nghiên cứu đã thực nghiệm trên một số kho ngữ liệu chuẩn như Senseval-1, Senseval-2, Senseval-3, DOS, ...Tuy nhiên, theo hiểu biết của chúng tôi, kho ngữ liệu cho bài toán WSD đối với tiếng Việt chưa có. Vì vậy cần thiết phải xây dựng kho ngữ liệu này cho phần thực nghiệm.

Trước hết, chúng tôi lựa chọn ra 10 từ nhập nhằng có nhiều nghĩa trong tiếng Việt. Tiếp theo, chúng tôi thu thập dữ liệu từ hơn 50 các trang báo tin tức tiếng Việt như: www.vnexpress.net, www.dantri.com.vn,... Sau đó, trích rút ngữ cảnh cho 10 từ nhập nhằng đã chọn. Từ những ngữ cảnh này, chúng tôi tiến hành gán nhãn bằng tay và thu được kho ngữ liệu D được mô tả như bảng 1.

Bảng 1. Thống kê dữ liệu được gán nhãn

Stt	Từ	Từ loại	Số nghĩa	Số mẫu
1	Cầu	Danh từ	2	1174
2	Cầu	Động từ	2	1085
3	Cắt	Động từ	8	1203
4	Khai	Động từ	4	3459
5	Ngã	Động từ	3	1426
6	Pha	Động từ	2	592
7	Phát	Động từ	8	1941
8	Quay	Động từ	6	3038
9	Sắc	Danh từ	4	2000
10	Tán	Động từ	3	575

Từ kho ngữ liệu được gán nhãn bằng tay *D*, chúng tôi chia kho ngữ liệu này thành hai phần theo tỉ lệ 3:1, thu được hai kho ngữ liệu D_1 và D_2 tương ứng. Kho ngữ liệu D_1 được sử dụng cho giai đoạn huấn luyện còn kho ngữ liệu D_2 được sử dụng cho giai đoạn kiểm tra trong các mô hình.

5.2. Công cụ

Chúng tôi sử dụng phần mềm WEKA 3.7[1] cho phần thực nghiệm. Nó cung cấp hầu hết các giải thuật học máy hiện nay.

5.3. Kết quả thực nghiệm

Trong phần này, chúng tôi trình bày kết quả thực nghiệm trên các mô hình sau: PART, J.48, và CMAR. Sử dụng tập dữ liệu D_1 để huấn luyện lên các mô hình PART, J.48, CMAR. Sau đó chúng tôi

[1] Xem chi tiết về phần mềm WEKA tại http://www.cs.waikato.ac.nz/ml/weka/downloading.html

sử dụng tập dữ liệu D_2 để kiểm tra cho các mô hình này. Từ bảng 2 cho thấy kết quả độ chính xác trung bình cho 10 từ nhập nhằng đối với các mô hình PART, J.48, CMAR lần lượt là 83.207%, 85.424%, 87.064%. Hơn nữa, kết quả của mô hình CMAR tốt hơn khi so sánh với các mô hình PART, J.48 lần lượt là 3.857 %, 1.640 % tương ứng.

Bảng 2. Bảng kết quả thực nghiệm

STT	Từ	Từ loại	D_1	D_2	PART	J.48	CMAR
1	Cần	Danh từ	880	294	90.476	90.476	94.217
2	Cần	Động từ	814	271	88.561	89.668	91.775
3	Cắt	Động từ	902	301	80.731	82.392	84.053
4	Khai	Động từ	2594	865	86.012	88.093	89.133
5	Ngã	Động từ	1070	356	78.370	81.161	82.865
6	Pha	Động từ	111	118	73.649	75.676	77.027
7	Phát	Động từ	1456	485	81.856	83.711	84.536
8	Quay	Động từ	2279	759	71.937	81.686	82.082
9	Sắc	Danh từ	1500	500	90.200	90.800	91.200
10	Tán	Động từ	431	144	90.278	90.278	93.750
Trung Bình					83.207	85.424	87.064

6. KẾT LUẬN

Trong bài báo này, chúng tôi đã đề xuất áp dụng phương pháp khai phá luật kết hợp cho việc giải quyết bài toán WSD. Kết quả thực nghiệm trên 10 từ nhập nhằng trong tiếng Việt mà chúng tôi lựa chọn ngẫu nhiên cho thấy độ chính xác tăng lần lượt là 3.857%, 1.640% khi so sánh với kết quả các phương pháp PART, J.48 tương ứng. Kết quả này cho thấy rằng các luật kết hợp khai phá từ những ngữ cảnh của từ nhập nhằng có hiệu quả trong việc giải quyết bài toán này. Hơn nữa, kết quả phương pháp CMAR cho thấy nó hiệu quả cao hơn so với các phương pháp dựa trên luật thông thường như PART và J.48. Trong tương lai, chúng tôi sẽ ứng dụng các luật kết hợp để cải thiện hiệu năng của mô hình học máy thống kê.

TÀI LIỆU THAM KHẢO

[1]. Agrawal, R., Imieli´nski, T., Swami, A., "Mining association rules between sets of items in large databases," in Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, New York, USA, 1993.

[2]. Liu, B., Hsu, W., Ma, Y., "Integrating classification and association rule mining," in Knowledge Discovery and Data Mining, New York, USA, 1998.

[3]. Li, W., Han, J., Pei, J., "Cmar: Accurate and efficient classification based on multiple class-association rules," in Proceedings of the 2001 IEEE International, Washington, DC, USA, 2001.

[4]. Yin, X., Han, J., "Classification based on Predictive Association Rules," in SDM, San Francisco,CA,USA, 2003.

[5]. Ide, N., Véronis J., "Introduction to the special issue on word sense disambiguation: the state of the art," Computational Linguistics, vol. 24, no. 1, pp. 2-40, 1998.

[6]. Le, A.C., Shimazu, A., "High word sense disambiguation using naive bayesian classifier with rich features," in The 18th Pacific Asia Conference on Language, Information and Computation, Waseda University, Tokyo, Japan, 2004.

[7]. Dinh, P.H., Nguyen, N.K., Le, A.C., "Combining statistical machine learning with transformation rule learning for vietnamese word sense disambiguation," in IEEE - RIVF, Ho Chi Minh, Vietnam, 2012.

[8]. R. L. R, Learning Decision Lists, Boston, Netherlands: Kluwer Academic Publishers, 1987.

[9]. Pedersen, T., "A decision tree of bigrams is an accurate predictor of word sense," in Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies, Stroudsburg, PA, USA, 2001.

[10]. J. Quinlan, C4.5: Programs for Machine Learning, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.