Chương 3 HỌC MÁY VÀ SO KHỚP ONTOLOGY

3.1 Các phương pháp học máy

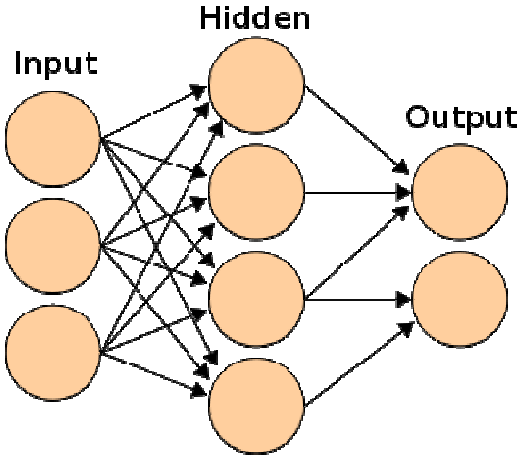
Học máy là một lĩnh vực thuộc ngành trí tuệ nhân tạo liên quan đến việc thiết kế và phát triển các thuật toán cho phép máy tính cải thiện hiệu quả qua thời gian dựa trên dữ liệu. Tuỳ thuộc vào tính chất của dữ liệu huấn luyện, các thuật toán máy học được chia thành ba nhóm. Nhóm thứ nhất là nhóm các thuật toán học có giám sát (supervised learning), huấn luyện trên tập mẫu được gán nhãn, thường được sử dụng trong các bài toán phân lớp hoặc nội suy. Nhóm thứ hai là các thuật toán học không giám sát (unsupervised learning), sử dụng các thuật toán gom cụm để khai thác các cấu trúc vốn có trong dữ liệu chưa gán nhãn. Nhóm các phương pháp học bán giám sát (semi-supervised learning), sử dụng cả các mẫu gán nhãn và chưa gán nhãn trong quá trình gán nhãn. Các thuật toán này quan tâm đến các tập dữ liệu mà tập mẫu gán nhãn chỉ chiếm một phần nhỏ (từ một đến vài mẫu trong mỗi lớp), trong đó a) không đạt được đủ số mẫu cần thiết để đạt được độ tin cậy cao và b) không cho phép tích hợp các thông tin *biết trước* vào trong quá trình học. Những tiểu mục dưới đây sẽ tóm lược một số kiến thức cơ bản về hai loại học có giám sát và bán giám sát.

3.1.1 Học có giám sát

Trong các thuật toán học có giám sát, dữ liệu huấn luyện bao gồm các cặp đối tượng đầu vào (thường là các vector) và kết xuất mong muốn tương ứng. Các kết xuất có thể là một giá trị liên tục hoặc có thể dự đoán nhãn lớp của đối tượng đầu vào. Nhiệm vụ của chương trình học có giám sát là dự đoán giá trị của hàm cho bất kỳ đối tượng đầu vào sau khi nhìn qua một số mẫu huấn luyện (các cặp đầu vào và kết xuất mục tiêu). Để đạt điều này, chương trình học phải tổng quát hoá từ những dữ liệu cho trước và đưa ta đến những tình huống chưa thấy theo một cách thức "hợp lý". Các chương trình phân loại được dùng rộng rãi là Mạng Nơ-

ron Nhân tạo, Support Vector Machine, k-láng giềng gần nhất, Naïve Bayes, Mô hình Hỗn hợp Gauss.

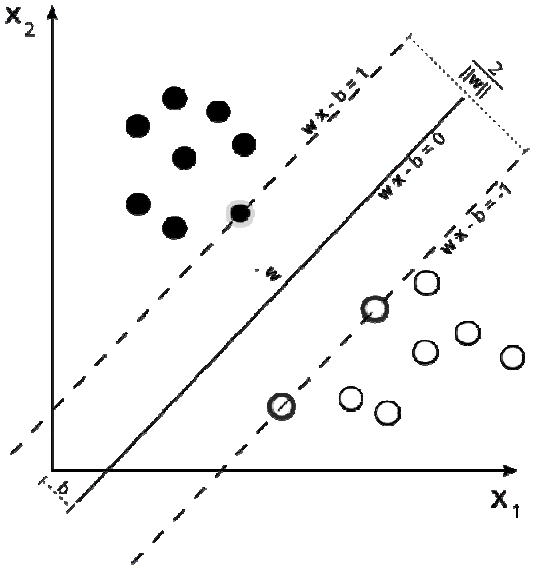
Mạng Nơ-ron Nhân tạo là một mô hình toán học hay mô hình tính toán dựa trên mạng nơ-ron sinh học. Nó bao gồm một nhóm các nơ-ron nhân tạo nối với nhau và xử lý thông tin dùng một cách tiếp cận kết nối để tính toán. Hình 3.1 mô tả một mạng nơ-ron lan truyền tiến nhiều lớp, loại mạng nơ-ron phổ biến nhất được sử dụng trong các bài toán phân lớp. Mạng nơ-ron này nhận tín hiệu đầu vào, các giá trị của vector đầu vào qua các nút ở lớp nhập, lan truyền các tín hiệu qua nút ẩn và cho kết xuất của mạng qua các nút xuất. Quá trình trên được gọi là lan truyền tiến và dùng trong pha phân lớp đối tượng. Quá trình học được thực hiện bằng sự lan truyền ngược sai số để điều chỉnh trọng số kết nối giữa node thuộc các lớp. Mạng nơ-ron nhân tạo có ưu điểm có thể giúp xác định các hàm số và các siêu phẳng phân biệt phức tạp nhưng mô hình của nó là một hộp đen đối với người sử dụng, ý nghĩa của các tham số trong mô hình lại không thể dễ dàng hiểu thấu đáo được.



Hình 3.1. Mạng lan truyền tiến nhiều lớp

Các *Support Vector Machine (SVM)* là một họ các thuật toán học có giám sát liên hệ với nhau. Xem dữ liệu đầu vào như hai tập vector trong không gian n-chiều, một *Support Vector Machine* sẽ xây dựng một siêu phẳng phân biệt trong không gian đó sao cho nó tối đa hoá biên lề giữa hai tập dữ liệu. Để tính lề, hai

siêu phẳng song song được xây dựng, mỗi cái nằm ở một phía của siêu phẳng phân biệt và chúng được đẩy về phía hai tập dữ liệu (xem ví dụ minh hoạ trong Hình 3.2). Một cách trực quan, một phân biệt tốt thu được bởi siêu phẳng có khoảng cách lớn nhất đến các điểm lân cận của cả hai lớp, vì nói chung lề càng lớn thì sai số tổng quát hoá của bộ phân lớp càng tốt hơn. SVM ban đầu là một thuật toán phân lớp tuyến tính, nhưng nhờ việc áp dụng của các hàm *kernel*, thuật toán có thể giúp tìm ra các siêu phẳng phân biệt phi tuyến trong không gian đặc trưng biến đổi. Và đây chính là điểm nổi bật của các SVM.



Hình 3.2. Siêu phẳng lề tối đại và các biên cho một SVM được huấn luyện với các mẫu từ hai lớp

Thuật toán phân lớp Naïve Bayes là một thuật toán phân lớp xác suất đơn giản dựa trên việc áp dụng định lý Bayes với giả định độc lập mạnh. Dựa vào bản chất rõ ràng của mô hình xác suất, các bộ phân lớp Naïve Bayes có thể được huấn luyện rất hiệu quả trong một môi trường học có giám sát. Trong nhiều ứng dụng thực tế, việc ước lượng tham số cho các mô hình Naïve Bayes sử dụng phương pháp khả suất tối đại; nói cách khác, người ta có thể dùng mô hình Naïve Bayes mà không cần tin vào xác suất Bayes hay dùng bất kỳ phương pháp Bayes nào. Dù thiết kế chất phác và các giả định quá đơn giản, các bộ phân lớp Naïve Bayes thường hoạt động tốt hơn mong đợi trong nhiều tình huống thế giới thực

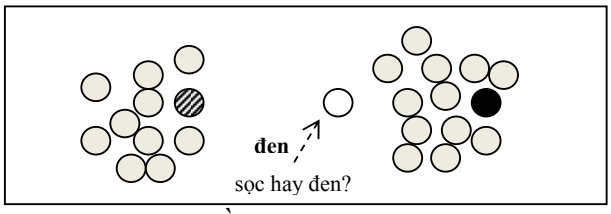
phức tạp. Một ưu điểm của thuật toán phân lớp Naïve Bayes là nó đòi hỏi ít dữ liệu huấn luyện để ước lượng các tham số cần thiết để phân lớp.

Thuật toán k-láng giềng gần nhất là thuật toán đơn giản nhất trong số các thuật toán học máy. Một đối tượng được phân loại bằng một cuộc bỏ phiếu đa số giữa các láng giềng của nó và đối tượng sẽ được gán lớp có nhiều có nhiều đối tượng chung nhất trong số k láng giềng gần nhất. k là một số nguyên dương, thường là số nhỏ. Nếu $k = 1$, đối tượng sẽ đơn giản được gán cho lớp của láng giềng gần nó nhất. Trong bài toán phân loại hai lớp, việc chọn k là số lẻ sẽ hữu ích giúp tránh được trường hợp số phiếu bầu bằng nhau.

3.1.2 Học bán giám sát

Về cơ bản, các thuật toán học bán giám sát sử dụng các mẫu dữ liệu chưa gán nhãn để làm giàu cho tập huấn luyện bằng cách từ từ gán nhãn cho chúng dựa vào ước lượng từ tập mẫu gán nhãn ban đầu. Hình 3.3 minh hoạ một ví dụ trực quan cho phương pháp học bán giám sát. Bởi vì chúng ta chỉ có một mẫu đen và một mẫu sọc biểu diễn cho hai lớp, khó mà quyết định mẫu trắng chưa gán nhãn sẽ thuộc lớp nào. Nhưng với sự hiện diện của các mẫu xám chưa biết, mẫu trắng có thể được phân vào lớp đen với độ chính xác cao hơn. Một số phương pháp học bán giám sát tiêu biểu là: EM (Expectation Maximization) với mô hình sinh hỗn hợp, tự huấn luyện, huấn luyện cộng tác, *transductive support vector machine* và các phương pháp đồ thị. Các nghiên cứu tổng quan về các phương pháp này được giới thiệu trong [23], [16]. Các thuật toán học bán giám sát dựa trên giả định phân phối của dữ liệu chưa biết và phân phối của dữ liệu đã biết là như nhau hoặc giả định nhất quán (consistent assumption), các điểm dữ liệu ở gần nhau trong không gian metric hoặc có cấu trúc gần giống nhau sẽ có cùng nhãn.

Mô hình sinh có lẽ là phương pháp học bán giám sát sớm nhất. Nó giả định phân phối của các điểm dữ liệu thuộc các phân lớp là phân phối hỗn hợp đồng nhất, ví dụ tuân theo mô hình hỗn hợp Gauss. Với số lượng lớn dữ liệu chưa gán nhãn, các thành phần hỗn hợp có thể được xác định; sau đó một cách lý tưởng



Hình 3.3. Ví dụ về trường hợp học bán giám sát.

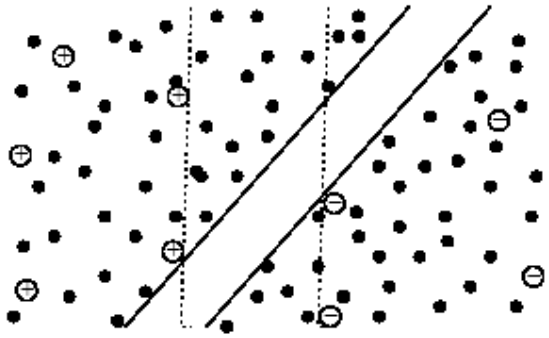
chúng ta chỉ cần một mẫu gán nhãn trên mỗi thành phần cũng đủ xác định phân phối hỗn hợp.

Tự huấn luyện là kỹ thuật được sử dụng được dùng nhiều nhất cho học bán giám sát. Trong tự huấn luyện một bộ phân lớp đầu tiên được huấn luyện với một lượng nhỏ dữ liệu gán nhãn. Bộ phân lớp sau đó được dùng để phân loại các dữ liệu chưa gán nhãn. Thông thường những điểm được gán nhãn với độ tin cậy cao nhất cùng với những nhãn dự đoán của nó sẽ được thêm vào tập huấn luyện. Bộ phân lớp được huấn luyện lại và thủ tục trên lặp lại. Lưu ý rằng bộ phân lớp dùng dự đoán của chính nó để dạy lại nó. Mô hình sinh và thuật toán EM có thể xem là một trường hợp đặc biệt của tự huấn luyện mềm. Người ta có thể nghĩ rằng một lỗi phân lớp có thể tăng cường thêm chính nó. Một số thuật toán cố gắng loại lỗi này bằng cách "không học" những điểm chưa gán nhãn nếu độ tin cậy dự đoán xuống dưới một ngưỡng nào đó.

Huấn luyện cộng tác giả định rằng (i) các đặc trưng có thể được chia thành hai tập; (ii) mỗi tập đặc trưng phụ là đủ để huấn luyện một bộ phân lớp tốt; (iii) hai tập là độc lập có điều kiện cho trước phân lớp. Đầu tiên hai bộ phân lớp độc lập được huấn luyện với dữ liệu gán nhãn, trên hai tập đặc trưng phụ tương ứng. Mỗi bộ phân lớp sau đó sẽ phân lớp dữ liệu chưa gán nhãn và "dạy" bộ phân lớp kia với một vài mẫu chưa gán nhãn (vùng với nhãn dự đoán của nó) mà chúng cảm thấy tin cậy nhất. Mỗi bộ phân lớp được huấn luyện với mẫu huấn luyện bổ sung cho bởi bộ phân lớp kia và quá trình lặp lại.

Transductive support vector machine (TSVM) là một mở rộng của support vector machine chuẩn với dữ liệu chưa gán nhãn. Trong một SVM chuẩn chỉ có dữ liệu gán nhãn được dùng và mục tiêu là tìm một biên tuyến tính có lề tối đại

trong không gian. Trong TSVM dữ liệu chưa gán nhãn cũng được dùng. Mục tiêu là tìm một gán nhãn của các dữ liệu chưa gán nhãn, sao cho tồn tại một biên tuyến tính có lề tối đại trên cả dữ liệu gán nhãn ban đầu và dữ liệu chưa gán nhãn. Biên quyết định có sai số tổng quát hoá nhỏ nhất giới hạn trên dữ liệu chưa gán nhãn. Hình 3.4 minh hoạ trực quan cho trường hợp TSVM, dữ liệu chưa gán nhãn hướng dẫn biên tuyến tính ra xa khỏi vùng có mật độ dữ liệu dày. Chỉ với dữ liệu gán nhãn, biên lề tối đại là đường chấm chấm. Với thêm các dữ liệu chưa gán nhãn (các điểm đen), biên lề tối đại là đường thẳng màu đen.



Hình 3.4. Một ví dụ về Tranductive SVM

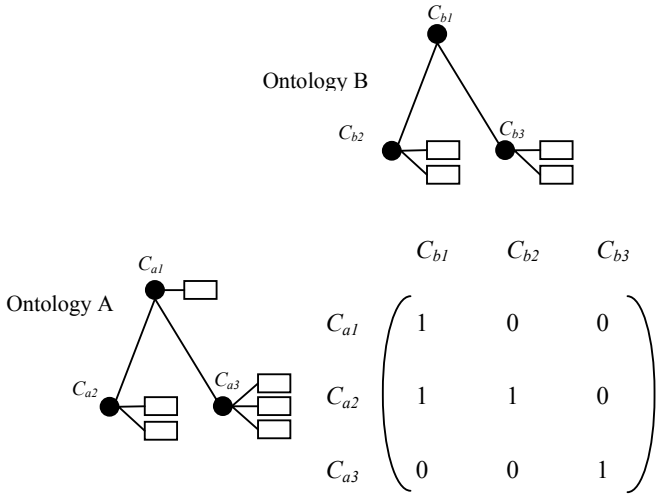
3.2 Học máy trong So khớp Ontology

Những kỹ thuật học máy rút trích được tri thức tự động từ dữ liệu. Do đó, những kỹ thuật này có ý nghĩa khi chúng ta cần giải quyết các bài toán mà lượng dữ liệu nhiều đến mức tràn ngập, không cho phép xử lý bằng tay và các hệ thống tự động cũng chưa đưa ra được kết quả cao, chẳng hạn như trong bài toán so khớp ontology [17]. Tiểu mục đầu tiên trong phần này trình bày biểu diễn bài toán so khớp ontology như một bài toán học máy có thể được giải quyết trong một mô hình học tổng quát Cách biểu diễn và mô hình học này được giới thiệu trong [11]. Tiểu mục tiếp theo giới thiệu các công trình liên quan đến việc nghiên cứu học máy trong bài toán so khớp ontology cùng với vấn đề được giải quyết trong luận văn này.

3.2.1 Bài toán So khớp Ontology như là một Bài toán học máy

Trong nghiên cứu này, luận văn quan tâm đến bài toán so khớp ontology với khái niệm tương ứng đơn giản, nghĩa là quan hệ giữa hai khái niệm được định nghĩa là quan hệ tương đương với độ tin cậy nhận giá trị 0 hoặc 1. Để giải quyết bài toán so khớp ontology, hệ thống tổ hợp các khái niệm giữa những ontology khác nhau. Trong trường hợp này, vấn đề là xác định giá trị của những cặp tổ hợp này. Nói cách khác, bài toán so khớp ontology bao gồm việc định nghĩa giá trị của các cặp khái niệm trong một ma trận cặp khái niệm, như trình bày trong Hình 3.5. Các dòng của ma trận biểu diễn các khái niệm của Ontology A, đó là C_{a1}, C_{a2} và C_{a3} và các cột của ma trận biểu diễn các khái niệm của Ontology B: C_{b1}, C_{b2} và C_{b3}. Giá trị của ma trận biểu diễn giá trị của ánh xạ. Giá trị 1 khi hai khái niệm có thể được ánh xạ và giá trị 0 khi hai khái niệm không thể được ánh xạ. Ví dụ, giá trị ở dòng thứ hai và cột thứ ba của ma trận biểu diễn giá trị của ánh xạ đối cho C_{a2} của Ontology A và C_{b3} của Ontology B. Ánh xạ cụ thể này là không hợp lệ bởi vì giá trị trong ma trận là 0.

Câu hỏi tiếp theo là cần thông tin gì để suy ra được ma trận. Như đã trình bày trong Chương 2, kỹ thuật cơ bản để xác định được ánh xạ giữa hai cặp khái niệm của hai ontology là sử dụng các độ đo tương tự. Chúng ta có thể sử dụng một độ đo khái niệm, ví dụ độ tương tự dựa trên tên, sử dụng so sánh chuỗi, hoặc các độ đo khác. Tuy nhiên, một độ đo tương tự duy nhất là không đủ để xây dựng được ma trận bởi tính đa dạng của các ontology. Ví dụ, xét trường hợp khái niệm "bank" giữa hai ontology. Các khái niệm trên dường như là một cặp tương ứng nếu dùng độ đo tương tự dựa trên chuỗi. Tuy nhiên, khi một khái niệm trong một ontology có khái niệm cha là "finance" và một khái niệm trong ontology kia có khái niệm cha là "construction", hai khái niệm này không phải là một tương ứng đúng vì chúng diễn tả những khái niệm khác nhau. Trong trường hợp như thế, một độ đo tương tự khác của các khái niệm. Do đó, hệ thống cần dùng nhiều độ đo tương tự để xác định các ánh xạ đúng.



Hình 3.5. Biểu diễn ma trận của bài toán so khớp ontology [11]

Như vậy để xác định giá trị cho ma trận so khớp, đầu tiên cần định nghĩa một vector tương tự sử dụng nhiều độ đo tương tự. Kết quả là ta có thể xây dựng được một bảng biểu diễn cho bài toán này như trình bày trong Bảng 3.1. Cột *ID* trong bảng đại diện cho một cặp khái niệm: *Class* biểu diễn giá trị của tương ứng và các cột ở giữa biểu diễn độ tương tự giữa các khái niệm. Ví dụ, dòng đầu tiên của bảng biểu diễn tương ứng cho C_{a1} và C_{b1} có giá trị tương tự 0.75 cho độ đo tương tự 1. Khi biết một số ánh xạ, ví dụ $C_{a1} \Leftrightarrow C_{b1}$ và $C_{a1} \Leftrightarrow C_{b2}$, hệ thống có thể dùng những ánh xạ này để xác định độ quan trọng của các độ đo tương tự. Sau đó, hệ thống có thể quyết định giá trị ánh xạ cho những cặp chưa biết ví dụ $C_{a5} \Leftrightarrow C_{b7}$ bằng cách dùng độ quan trọng của các độ đo tương tự. Bảng ví dụ 3.1 này tương tự như bài toán trong một hệ thống học máy có giám sát. Do đó, bài toán so khớp ontology có thể được chuyển thành một bài toán học máy.

Bảng 3.1. Biểu diễn dạng bảng của bài toán so khớp ontology

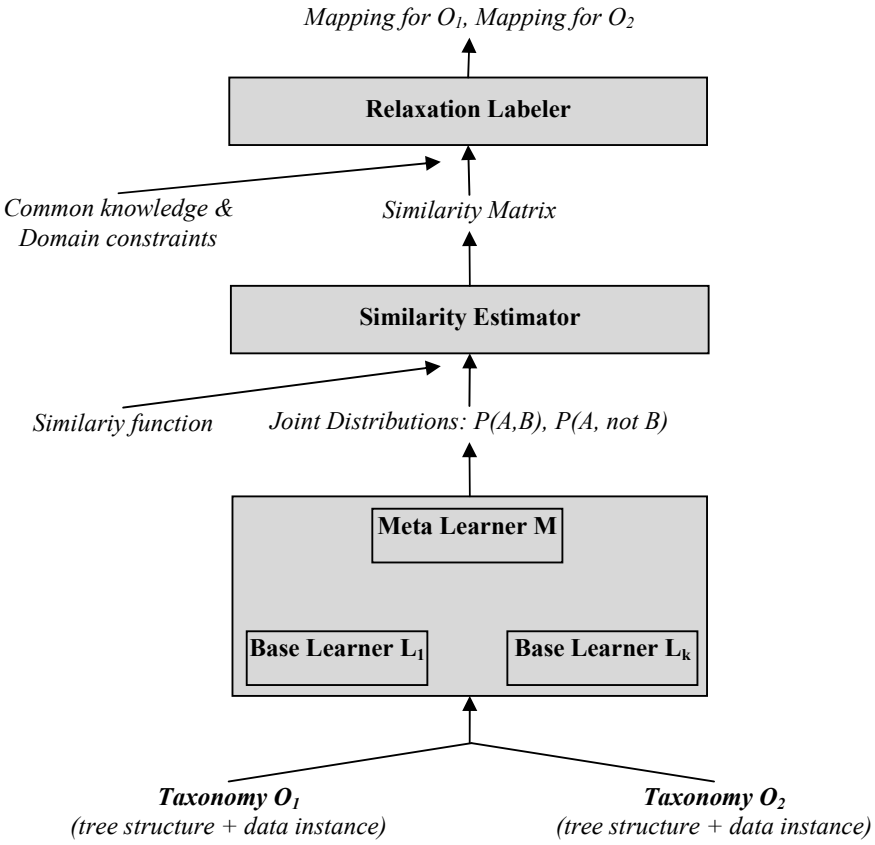
ID	Độ tương tự 1	Độ tương tự 2	…	Độ tương tự n	Lớp
$C_{a1} \Leftrightarrow C_{b1}$	0.75	0.4	…	0.38	1 (Dương)
$C_{a1} \Leftrightarrow C_{b2}$	0.52	0.7	…	0.42	0 (Âm)
…	…	…	…	…	…
$C_{a5} \Leftrightarrow C_{b7}$	0.38	0.6	…	0.25	?
…	…	…	…	…	…

3.2.2 Các nghiên cứu có liên quan

Ngoài mô hình học tổng quát từ [11] như được trình bày ở trên, cách tiếp cận học máy cũng được giới thiệu trong một vài công trình có liên quan đến bài toán so khớp ontology. Agrawal và Srikant [1] giới thiệu mô hình ENB (Enhanced Naïve Bayes) cho bài toán tích hợp các catalog hàng hoá. ENB là thuật toán cải tiến của thuật toán học cơ sở Naïve Bayes, trong đó các tác giả sử dụng các thông tin bổ sung về quan hệ giữa lớp đề hỗ trợ cho việc phân lớp các thể hiện vào các lớp của catalog. Kết quả phân tích và thử nghiệm cho thấy mô hình học cải tiến giúp cải thiện đáng kể độ chính xác của việc tích hợp dữ liệu.

Wang và cộng sự [19] giới thiệu hệ thống cũng giới thiệu một hệ thống so khớp ontology trong đó sử dụng nội dung của các thể hiện để xây dựng độ đo tương tự giữa các khái niệm. Tiếp đó, sử dụng nhân lực để gán nhãn bằng tay cho các cặp khái niệm chọn lọc, họ xây dựng một tập dữ liệu huấn luyện mẫu và sử dụng phương pháp Markov Random Field để làm bộ học phân lớp cho bài toán so khớp các bộ chỉ mục thư viện tại Thư viện Quốc gia Hà Lan. Trong hệ thống này, các tác giả sử dụng thông tin là các trường siêu dữ liệu mô tả cho các đối tượng sách và đa phương tiện làm cơ sở để tính độ đo tương tự. Thông tin này được dùng riêng trong trường hợp của tác giả nhưng có thể dễ dàng tích hợp vào các hệ thống học máy tổng quát như [11], các thông tin này có sẵn trong một số bài toán so khớp khác.

Doan và cộng sự [7] giới thiệu GLUE là hệ thống so khớp ontology trong đó sử dụng kỹ thuật học trong một số bước để xây dựng độ tương tự giữa các khái niệm. GLUE cũng sử dụng nhiều bộ học bao gồm các bộ học trên các loại dữ liệu khác nhau và một bộ siêu học để lựa chọn đặc trưng tương tự cho các bước so khớp tiếp theo. Hình 3.6 mô tả kiến trúc tổng quát của GLUE.



Hình 3.6. Kiến trúc của GLUE [7]

Jeong và cộng sự [14] giới thiệu một mô hình học cho bài toán tổng quát cho bài toán so khớp các lược đồ XML. Mô hình này cũng tương tự như mô hình được giới thiệu trong [11] bao gồm việc xây dựng vector tương tự nhiều đặc trưng và áp dụng các chiến lược học khác nhau. Các tác giả cũng thử nghiệm các phương pháp học khác nhau trên hệ thống bao gồm học cả học có giám sát và bán giám sát.

Các thuật toán học máy có giám sát cần sử dụng một tập dữ liệu đã được gán nhãn để huấn luyện mô hình, việc này thường gây tốn kém vì chi phí nhân công cho việc gán nhãn cao. Hơn nữa, do đặc thù đa dạng của các môi trường ứng dụng so khớp ontology thực tế, hệ thống học cần sử dụng một tập dữ liệu huấn luyện riêng nhận từ người dùng cuối cho từng bài toán. Do đó, việc giới hạn kích thước tập huấn luyện là cần thiết để bảo đảm sự hài lòng của người dùng. Những

người dùng cuối thường không sẵn lòng để gán nhãn hàng ngàn mẫu dữ liệu khác nhau như yêu cầu của các hệ thống học máy. Trong trường hợp số mẫu huấn luyện được giới hạn đến mức ít nhất, hệ thống sử dụng phương pháp học bán giám sát kết hợp với học chủ động để giải quyết vấn đề số mẫu huấn luyện ít hơn nhiều so với số mẫu cần dự đoán.

APPEL [8] cũng là một hệ thống học máy tương tự như [11], nhưng hệ thống này đòi hỏi việc sử dụng các ontology khác cũng như yêu cầu người dùng thẩm định là một số cặp so khớp hạt giống được phát sinh tự động trước sử dụng chúng làm tập huấn luyện cho mô hình. Hệ thống này có thể đáp ứng về mặt hiệu quả đối với chương trình nhưng gây khó khăn đối với những người dùng không chuyên do phải cung cấp một số tham số chuyên môn như độ tin cậy của tương ứng.

Có một điểm lưu ý khi sử dụng phương pháp học bán giám sát là cần thiết lập một môi trường thích hợp để sử dụng. Qua thử nghiệm, Jeong và cộng sự [14] nhận thấy các thuật toán học bán giám sát không thực sự cho kết quả cải thiện đáng kể so với các thuật toán học có giám sát. Điều này có thể lý giải do môi trường thử nghiệm không thật sự thích hợp với các thuật toán học bán giám sát, cụ thể số mẫu gán nhãn không thực sự vượt trội so với số mẫu gán nhãn (190 mẫu chưa gán nhãn trên 60 mẫu gán nhãn).

Ngoài ra, việc mẫu chưa gán nhãn có thể là giảm hiệu quả học trong các thuật toán học bán giám sát cũng được ghi nhận trong [6]. Tian và cộng sự [18] xem xét hiện tượng này qua việc khảo sát hiệu quả của các thuật toán học trong các điều kiện phân phối xác suất của các tập dữ liệu có gán nhãn (L) và tập dữ liệu chưa gán nhãn (U). Với tình huống giả định về phân phối dữ liệu thoả, tức là $P_L = P_U$, dữ liệu chưa gán nhãn giúp nâng cao hiệu quả học của các các học bán giám sát. Trong trường hợp $P_L \neq P_U$, việc thay đổi của hiệu quả là không đoán trước. Tuy nhiên, ngược với những ghi nhận trên, Zhou và cộng sự [22] đề xuất một mô hình học cộng tác trong bài toán truy vấn ảnh với phản hồi người dùng.

Thử nghiệm cho thấy mô hình được đề xuất cho hiệu quả cao hơn các mô hình học có giám sát do ảnh hưởng của kích thước tập huấn luyện nhỏ.

Với những thông tin trên, luận văn đề xuất mở rộng mô hình học tổng quát trong [11] thành một hệ thống học linh hoạt trong đó bổ sung phương pháp học bán giám sát kết hợp học chủ động vào mô hình để xử lý cho trường hợp phản hồi người dùng.