Hệ thống chuyên gia có ứng dụng 41 (2014) 2372–2378

Danh sách nội dung có sẵn tại ScienceDirect

Hệ thống chuyên gia với các ứng dụng

trang chủ tạp chí: www.elsevier.com/locate/eswa

Học tập bán giám sát kết hợp đồng đào tạo với học tập tích cực

Yihao Zhang [một], Junhao Wen b, ,Xibin Wang [một],Zhuo Jiang [một]

[một] Cao đẳng Khoa học Máy tính, Đại học Trùng Khánh, Trùng Khánh 400030, Trung Quốc
[b] Cao đẳng kỹ thuật phần mềm, Đại học Trùng Khánh, Trùng Khánh 400030, Trung Quốc

thông tin bài viết

Từ khóa:
Học tập bán giám sát
Đồng đào tạo
Ước tính độ tin cậy
Học tập tích cực
Phiên bản thông tin

trừu tượng

Đồng đào tạo là một mô hình bán giám sát tốt, mô hình này yêu cầu tập dữ liệu được mô tả bởi hai quan điểm của các đối tượng địa lý. Có một đặc điểm đáng chú ý được chia sẻ bởi nhiều thuật toán đồng đào tạo: Các trường hợp không được gắn nhãn nên được dự đoán với độ tin cậy cao, vì điểm số tin cậy cao thường ngụ ý rằng dự đoán tương ứng là đúng. Thật không may, nó không phải lúc nào cũng có thể cải thiện hiệu suất phân loại với các cá thể không được gắn nhãn có độ tin cậy cao này. Trong bài báo này, một thuật toán học tập bán giám sát được đề xuất kết hợp các lợi ích của cả đồng đào tạo và tích cực học hỏi. Thuật toán áp dụng đồng đào tạo để chọn các trường hợp đáng tin cậy nhất theo hai dấu hiệu của độ tin cậy cao và hàng xóm gần nhất để thúc đẩy bộ phân loại, cũng khai thác các trường hợp nhiều thông tin nhất với chú thích của con người để cải thiện hiệu suất phân loại. Thử nghiệm trên một số Bộ dữ liệu UCI và tác vụ xử lý ngôn ngữ tự nhiên, chứng tỏ phương pháp của chúng tôi đạt được cải tiến đáng kể hơn nhờ hy sinh cùng một lượng nỗ lực của con người.

1. Giới thiệu

Học bán giám sát rất hữu ích trong nhiều ứng dụng thực tế, học từ cả dữ liệu được gắn nhãn và dữ liệu không được gắn nhãn và tự động khai thác dữ liệu không được gắn nhãn để cải thiện việc học hiệu suất mà không có sự can thiệp của con người (Chapelle, Scholkopf, & Zien, 2006; Zhu, 2008). Đồng đào tạo là một chương trình Bán giám sát nổi tiếng mô hình học tập bắt đầu từ công việc cơ bản của Blum và Mitchell (Blum và Mitchell, 1998), được đề xuất cho các bài toán phân loại nhị phân trong đó có hai quan điểm khác nhau. Thuật toán đồng đào tạo stan dard yêu cầu hai quan điểm (Blum và Mitchell, 1998), nghĩa là, các thuộc tính có thể được phân chia tự nhiên rally thành hai bộ, mỗi bộ đều đủ để học và độc lập có điều kiện đối với cái khác được gán nhãn lớp. Đào tạo đồng hoạt động theo cách lặp đi lặp lại mà hai bộ phân loại được đào tạo riêng biệt trên các chế độ xem khác nhau và các dự đoán của một trong hai bộ lọc clas trên các cá thể không được gắn nhãn được sử dụng để tăng cường tập huấn luyện của người khác (Zhang & Zhou, 2011; Zhu, 2008).

Có một đặc điểm đáng chú ý được chia sẻ bởi nhiều đồng đào tạo thuật toán: các trường hợp không được gắn nhãn đã chọn sẽ được dự đoán với độ tin cậy cao, vì điểm số tin cậy cao thường ngụ ý rằng dự đoán tương ứng là đúng (Blum và Mitchell, Năm 1998; Mihalcea, 2004). Thật không may, không phải lúc nào tôi cũng có thể chứng minh hiệu suất phân loại với độ tin cậy cao này các trường hợp không được gắn nhãn. Tang và cộng sự. (2007) đề xuất một chiến lược mới rằng việc cập nhật các bộ phân loại thông qua đồng đào tạo, bổ sung các phiên bản nega tive gần với siêu mặt phẳng của bộ phân loại sao cho trình phân loại sẽ học cách phân biệt tốt hơn các trường hợp này. Giảm bớt công việc trên cơ sở đào tạo đồng tiêu chuẩn, một số nội dung có liên quan các phương pháp đã được phát triển. Wang và Zhou (2010) đã phân tích quá trình đồng đào tạo và được xem nó như là một tổ hợp nhãn kết hợp trên hai chế độ xem. Yu, Krishnapuram, Rosales và Rao (2011) đã đề xuất một mô hình đồ họa vô hướng của Bayes để đồng đào tạo, mô hình này có thể xử lý một cách thanh lịch các mẫu dữ liệu bị thiếu lượt xem. Sun và cộng sự. (2011) đã đề xuất một quy tắc thuật toán đồng đào tạo dựa trên thực thể, không yêu cầu kiến thức trước về phân phối giai cấp rất quan trọng trong nhịp điệu thuật toán đồng huấn luyện tiêu chuẩn. Thật không may, đồng đào tạo được kêu gọi chính xác khi tập huấn luyện được gắn nhãn là nhỏ và không chắc chắn liệu tập huấn luyện stan dard có hoạt động hay không trên các tập huấn luyện có nhãn nhỏ (Du và cộng sự, 2011).

Trong bài báo này, một thuật toán phân loại bán giám sát mới đã được đề xuất kết hợp các lợi ích của cả hai chương trình đồng đào tạo và học tập tích cực, và những đóng góp lớn gấp đôi:

(1) Đầu tiên, trong mỗi vòng đồng đào tạo, một vài trong số các các trường hợp được chọn ra từ dữ liệu không được gắn nhãn cho vòng học tiếp theo và các trường hợp đáng tin cậy nhất là được lựa chọn theo hai tiêu chí về độ tin cậy cao và hàng xóm gần nhất. Cụ thể, mức độ đóng góp được định nghĩa là tiêu chí của một số trường hợp thông tin được chọn, mà không chỉ xem xét các trường hợp không chắc chắn nhất nhưng cũng xem xét sự khác biệt không chắc chắn giữa cá thể và hàng xóm gần nhất của nó.

Tác giả tương ứng. Điện thoại: +86 13983146919.
Địa chỉ email: yihaozhang@cqu.edu.cn (Y. Zhang), jhwen@cqu.edu.cn (J. Wen), binxiwang@cqu.edu.cn (X. Wang), jiangzhuo1986@gmail.com (Z. Jiang).

0957-4174 / $ - xem vấn đề phía trước
http://dx.doi.org/10.1016/j.eswa.2013.09.035

(2) Thứ hai, học tập tích cực sử dụng khung truy vấn mà một người học tích cực truy vấn các trường hợp mà nó ít cách gắn nhãn nhất định, đóng góp do thuật toán xác định của chúng tôi mức độ là tiêu chí lựa chọn của các cá thể thông tin, đã đạt được cải tiến đáng kể hơn để hy sinh cùng một lượng nỗ lực của con người và làm việc tốt trên các tập huấn luyện có nhãn.

Phần còn lại của bài báo này được tổ chức như sau. Phần 2 đánh giá một số vấn đề trong đồng đào tạo và học tập tích cực. Sau đó Sec tion 3 giới thiệu bản phác thảo của thuật toán và trình bày chi tiết về nhịp điệu thuật toán. Phần 4 báo cáo kết quả thử nghiệm trên một số của bộ dữ liệu trong thế giới thực và phân tích thêm các lý do cơ bản cho thuật toán. Cuối cùng, Phần 5 kết luận và chỉ ra một số các vấn đề cho công việc sau này.

2. Các vấn đề về đồng đào tạo và học tập tích cực

Đồng đào tạo là một thuật toán bán giám sát, đa chế độ xem sử dụng tập dữ liệu được gắn nhãn ban đầu để tìm hiểu bộ phân loại yếu trong mỗi quan điểm (Blum và Mitchell, 1998). Sau đó, mỗi bộ phân loại được áp dụng cho phần còn lại của các trường hợp không được gắn nhãn và đào tạo đồng phát hiện các trạng thái mà mỗi bộ phân loại đưa ra các điểm dự đoán tự tin nhất. Những trường hợp có độ tin cậy cao này được gắn nhãn nhãn ước tính và được thêm vào tập dữ liệu được gắn nhãn. Dựa trên tập huấn luyện mới, một bộ phân loại mới được lặp lại trong vài lần lặp lại. Cuối cùng, một giả thuyết cuối cùng được tạo ra bởi một sơ đồ bỏ phiếu kết hợp các dự đoán của các bộ phân loại đã học trong mỗi lượt xem.

Đồng đào tạo, một mô hình học tập bán giám sát tốt, đã gần đây đã thu hút sự chú ý và quan tâm đáng kể (Zhou & Li, 2010). Chương trình đào tạo đồng tiêu chuẩn giả định rằng dữ liệu có thể được ghi chép bởi hai bộ tính năng hoặc chế độ xem riêng biệt và nó hoạt động tốt khi hai quan điểm thỏa mãn tính đầy đủ và độc lập giả định (Blum và Mitchell, 1998). Tuy nhiên, hai các giả định thường không được biết đến hoặc không được đảm bảo trong thực tế, và xem việc chia nhỏ là không đáng tin cậy trong các tập huấn luyện nhỏ được dán nhãn nhất định. Thông thường hơn, hầu hết các tập dữ liệu được giám sát được mô tả bởi một tập hợp các thuộc tính (một chế độ xem). Để khai thác lợi thế của việc đồng đào tạo, Goldman và Zhou (Goldman và Zhou, 2000) đề xuất một thuật toán thuật toán không khai thác phân vùng đặc trưng; thuật toán sử dụng hai thuật toán học tập có giám sát khác nhau để đào tạo hai bộ định âm clas. Zhou và Li (2005) đề xuất phương pháp huấn luyện ba người, sử dụng ba bộ phân loại được tạo từ các mẫu bootstrap của tập huấn luyện ban đầu. Du và Ling (2011) có kết luận rằng hiệu quả của đồng đào tạo là hỗn hợp. Đó là, nếu hai quan điểm được đưa ra và được biết là thỏa mãn hai giả định, đồng đào tạo hoạt động tốt; Nếu không, dựa trên các tập hợp đào tạo được gắn nhãn nhỏ, hãy xác minh việc nhập các giả định hoặc tách một chế độ xem thành hai chế độ xem là không đáng tin cậy; do đó, không chắc chắn liệu chương trình đồng đào tạo tiêu chuẩn sẽ hoạt động hay không.

Học tích cực, một trường con của học máy, có thể thực hiện đặt cược với ít đào tạo hơn bằng cách chọn biểu mẫu dữ liệu mà nó học được. Nó cố gắng vượt qua nút cổ chai về nhãn bằng cách đặt các truy vấn trong hình thức của các cá thể không được gắn nhãn sẽ được gắn nhãn bởi một tiên tri (Giải quyết, 2010) (ví dụ: một trình chú thích của con người). Bằng cách này, người học tích cực nhằm mục đích đạt được độ chính xác cao bằng cách sử dụng ít trường hợp được gắn nhãn như có thể, do đó giảm thiểu chi phí lấy dữ liệu được dán nhãn. Ý tưởng chính đằng sau hầu hết các thuật toán học tập tích cực là chọn các trường hợp không chắc chắn nhất để phân loại. Do đó, một chìa khóa khía cạnh của học tập tích cực là đo lường sự không chắc chắn về phân loại của các cá thể không được gắn nhãn. Zhu (2003) đã đề xuất một chiến lược học tập bán giám sát mới, kết hợp học tập tích cực và học bán giám sát theo mô hình trường ngẫu nhiên Gaussian. Yang và cộng sự. (2009) đề xuất khung Bayes về khoảng cách hoạt động học chỉ số bằng cách chọn các cặp ví dụ không được gắn nhãn đó với độ không đảm bảo đo lớn nhất trong khoảng cách tương đối. Lughofer (2012) pro đã đưa ra một chiến lược học tập tích cực mới cho các bộ phân loại theo hướng dữ liệu, điều này cần thiết để giảm chú thích và giám sát ef đồn của người vận hành trong các hệ thống phân loại ngoại tuyến và trực tuyến, như toán tử chỉ phải gắn nhãn một tập hợp con tinh tế của dòng ngoại tuyến dữ liệu đào tạo. Li, Shi và Liu (2012) đã đề xuất một phương pháp tiếp cận tìm hiểu tích cực chung kết hợp một chiến lược truy vấn tổng hợp mới và phân biệt đối xử hiện có, phù hợp một cách thích nghi với sự khác biệt về phân biệt và thể hiện tính mạnh mẽ hơn những sử dụng chiến lược đơn lẻ.

Từ những phân tích trên, đồng đào tạo là một kỹ thuật quan trọng để cải thiện độ chính xác của dự đoán khi dữ liệu được gắn nhãn khan hiếm. Tuy nhiên, thuật toán này thường không được đảm bảo hoạt động tốt trong ứng dụng thế giới thực. Thứ nhất, đồng đào tạo yêu cầu tập dữ liệu có thể được tách hai quan điểm, và thỏa mãn các giả định đầy đủ và không sâu sắc. Trong thực tế, những điều kiện đó không dễ để Hoàn thành. Thứ hai, mặc dù đồng đào tạo thường chọn các trường hợp có mức độ liên quan cao được gắn nhãn với nhãn lớp ước tính và thêm chúng vào các tập huấn luyện, điều này không đảm bảo những các trường hợp có độ tin cậy cao được chọn có giá trị hơn để ứng biến độ chính xác của dự đoán. Trong bài báo, một thuật toán bán giám sát kết hợp đồng đào tạo với học tập tích cực đã được đề xuất, có thể tận dụng lợi ích của hai thuật toán và giảm nỗ lực chú thích của các nhà khai thác.

3. Kết hợp đồng đào tạo với học tập tích cực

Trong phần này, chúng tôi cung cấp mô tả cấp cao về thuật toán học bán giám sát và có thể mô tả khung của nó như Hình 1. Thuật toán (SSLCA) kết hợp đồng đào tạo với học tập tích cực có thể được chia thành ba bước: thứ nhất, dữ liệu được chia thành hai chế độ xem để áp dụng đào tạo đồng đào tạo hai chế độ xem tiêu chuẩn, học trình phân loại h1 và trình phân loại h2 chỉ dựa trên hai chế độ xem của dữ liệu được gắn nhãn; thứ hai, dữ liệu unla beled cũng được chia thành hai chế độ xem để ước tính kết quả của chúng bằng cách sử dụng bộ phân loại riêng biệt; thứ ba, các trường hợp đáng tin cậy nhất hoặc các trường hợp thông tin được chọn dựa trên một số chiến lược. Các hầu hết các trường hợp cung cấp thông tin được chọn theo hai tiêu chí dựa trên độ tin cậy cao và người hàng xóm gần nhất, sau đó được đưa vào một nhóm khác để có thêm chú thích.

3.1. Phương pháp ước tính độ tin cậy

3.1.1. Phương pháp Naive Bayes

Naïve Bayes hình thành ước tính hậu kỳ tối đa cho lớp xác suất có điều kiện cho từng tính năng từ khóa đào tạo được gắn nhãn dữ liệu D. Các xác suất trước của mỗi lớp được tính theo kiểu mô phỏng bằng cách đếm qua các trường hợp. Xác định P (cj) biểu thị xác suất của lớp cj, và | D | biểu thị số lượng phiên bản trong dữ liệu đào tạo:

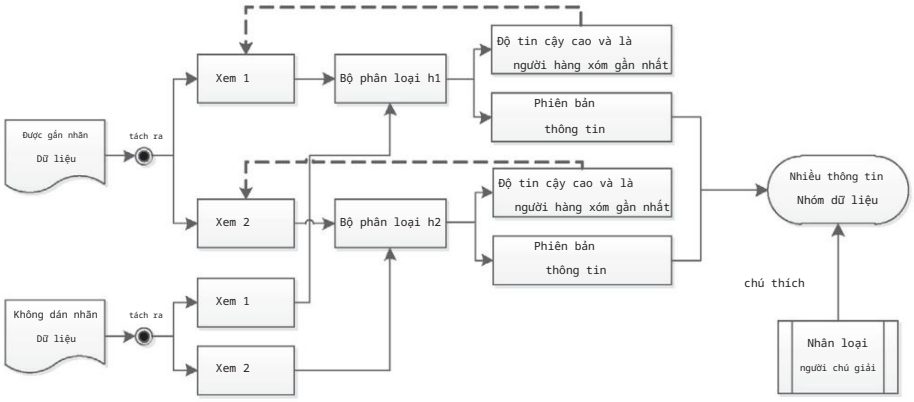
$$Pðcjþ ¼ \frac{1 þ PjDj i¼1Pðcjþ}{jDj}$$

Sau đó, ước tính xác suất posteriori cho mỗi trường hợp là tính theo giả thiết độc lập, xác định ai biểu thị từng tính năng trong mỗi trường hợp, n biểu thị số trong số các tính năng:

$$Pðcjjaiþ / Pðcjþ Pðaijcjþ ¼ Pðcjþ Yn \quad Pðaijcjþ$$
$$i¼1$$

3.1.2. Phương pháp tối đa hóa kỳ vọng

Tối đa hóa kỳ vọng (EM) là một thống kê lặp lại kỹ thuật để ước tính khả năng xảy ra tối đa trong các vấn đề với



Hình 1. Khung của thuật toán SSLCA.

dữ liệu không đầy đủ. Với một mô hình tạo dữ liệu và dữ liệu có một số giá trị bị thiếu, EM sẽ tối đa hóa cục bộ khả năng của các tham số và đưa ra ước tính cho các giá trị bị thiếu. Mô hình tổng hợp Bayes ngây thơ cho phép áp dụng EM để ước lượng tham số. Trong kịch bản của chúng tôi, các nhãn lớp của dữ liệu không được gắn nhãn được coi là các giá trị bị thiếu. Trong triển khai, EM là một quá trình hai bước lặp đi lặp lại. Các giao diện ước tính tham số ban đầu được đặt bằng cách sử dụng các byte ngây thơ tiêu chuẩn từ các cá thể được gắn nhãn. Bước E tính toán xác suất sau của các tham số của chúng ta là ci , với xi và sử dụng cài đặt hiện tại của các tham số của chúng ta. Sử dụng quy tắc Bayes, chúng tôi thu được:

$$pðci ¼ jjxi; U; l; RÞ ¼ \frac{pðxijci ¼ j; l; RÞpðci ¼ j; UÞ}{Pk\ 1¼1pðxijci ¼ 1; l; RÞpðci ¼ 1; UÞ}$$

Ở đây, p (xi | ci = j; l, R) được đưa ra bằng cách đánh giá mật độ của một Gauss ian với trung bình lj và hiệp phương sai Pj tại ví dụ xi, p (ci = j; U) là gi ven bởi Uj, khi đó Các cá thể có độ tin cậy cao đã được thêm vào nhóm dữ liệu được gắn nhãn riêng biệt. M-step ước tính các tham số bộ phân loại mới bằng cách sử dụng tất cả dữ liệu được gắn nhãn trong nhóm. Cho wðjÞ : ¼ pðci ¼ jjxi; U; l; RÞ trong mỗi lần lặp, tỷ lệ của mỗi lớp Uj, lj trung bình và hiệp phương sai Pj được tính như sau:

$$Uj ¼ \frac{1}{m} \sum_{i¼1}^{m} wðjÞ$$

$$lj ¼ \frac{\sum_{i¼1}^{m} wðjÞ\ xi}{\sum_{i¼1}^{m} wðjÞ}$$

$$Rj ¼ \frac{\sum_{i¼1}^{m} wðjÞ ðxi - lj Þðxi - lj Þ^{T}}{\sum_{i¼1}^{m} wðjÞ}$$

3.2. Phương pháp tách Entropy

Entropy split là một phương pháp heuristic đơn giản để chia các khung nhìn đơn lẻ thành hai khung nhìn (Du et al., 2011); đầu tiên, là tính toán entropy của từng đối tượng trong một chế độ xem dựa trên toàn bộ tập dữ liệu, tương tự như tính toán entropy khi quyết định đối tượng nào nên được chọn làm gốc của cây quyết định . thuộc lớp ci trong tập dữ liệu D, khi đó pi có thể được định nghĩa là pi = | ci, D | / | D |, do đó, thông tin kỳ vọng về phân loại trong tập dữ liệu D có thể được tính như sau:

$$InfoðDÞ¼ \sum_{i¼1}^{m} pilog2ðpiÞ$$

Hãy để | Dj | / | D | biểu thị trọng số của phép chia có giá trị rời rạc j, thuộc tính A có v giá trị khác nhau {a1, a2 , av}, do đó entropy của thuộc tính A có thể được tính như sau:

$$InfoAðDÞ ¼ \sum_{j¼1}^{v} \frac{jDjj}{jDj} InfoðDjÞ$$

Theo trực giác, entropy càng lớn thì khả năng tiên đoán càng cao. Chúng tôi chỉ cần gán các đối tượng có entropy cao nhất, thứ nhất, v.v. cho chế độ xem đầu tiên, sau đó gán các đối tượng có entropy cao nhất cho chế độ xem thứ hai, thứ tư, v.v. Cơ sở lý luận là phân phối đồng đều các đặc trưng entropy cao trong hai chế độ xem và do đó, cả hai chế độ xem có nhiều khả năng là đủ.

3.3. Chọn các phiên bản thông tin

Học tích cực thường chọn mẫu không chắc chắn để nhập truy vấn, người học tích cực truy vấn các trường hợp mà nó ít chắc chắn nhất về cách gắn nhãn. Cách tiếp cận này thường đơn giản đối với các mô hình học tập theo xác suất. Ví dụ, khi sử dụng một mô hình xác suất xác suất để phân loại nhị phân, việc lấy mẫu độ không đảm bảo chỉ đơn giản là truy vấn cá thể có xác suất dương tính sau gần nhất là 0,5.

Trong thuật toán của chúng tôi, chúng tôi xác định mức độ đóng góp là tiêu chí lựa chọn của các cá thể thông tin, không chỉ xem xét mức độ không chắc chắn nhất của các trường hợp mà còn xem xét sự khác biệt về độ không chắc chắn giữa cá thể đó và láng giềng gần nhất của nó, vì vậy mức độ đóng góp có thể được xác định là:

$$ContributionðConf; xiÞ ¼ \frac{1 þ a X}{N\ Confðxi; cÞ} \qquad ½Confðxi; cÞ Confðx; cÞ$$

$$x2Nðxi Þ$$

Trong đó, N biểu thị số lượng phân loại, Conf (xi, c) biểu thị độ tin cậy của đối tượng xi thuộc loại c, a biểu thị trọng số của mức độ đóng góp, N (xi) biểu thị hàng xóm gần nhất của đối tượng xi. Nó dựa trên ý tưởng, nếu độ tin cậy của cá thể cao, nhưng độ tin cậy của người hàng xóm của nó thấp, điều này không đáp ứng giả thuyết phân cụm; nếu không, những người đó có khả năng cao hơn là các cá thể cung cấp thông tin vì mức độ đóng góp cao.

3.4. Thuật toán SSLCA

Thuật toán SSLCA là một thuật toán bán giám sát, đa chế độ xem; khung được mô tả như Hình 1. Trong phần này, chúng tôi cung cấp một mô tả chi tiết, được thể hiện trong Hình 2.

4. Thử nghiệm

4.1. Bộ dữ liệu thử nghiệm

Để đánh giá hiệu suất của thuật toán SSLCA, chúng tôi sử dụng sáu bộ dữ liệu UCI (Witten và Frank, 2005) và dữ liệu Coreference của Trung Quốc về tác vụ xử lý ngôn ngữ tự nhiên trong thế giới thực.

Given:

-a learning algorithm ℓ

-the labeled data L and unlabeled data U

-the number k of iterations to be performed

SSLCA algorithm:

-split labeled data L and unlabeled data U respectively into view V_1 and view V_2

Loop for k iterations

-use ℓ, $V_1(L)$ and $V_2(L)$ to create classifiers h_1 and h_2

For each class c_i, do

-let E_1 and E_2 be the e unlabeled instances on each classifiers h_1 and h_2

-selected the high confidence and nearest neighbor instances E_{hn} for c_i, label them according to h_1 and h_2, respectively, and add them to L

-selected the informative instances E_{in}, and add them to data pool P

-remove E_{hn} and E_{in} form U

End For each class c_i

End Loop for k iterations

-label the instances of data pool P, and add them to L

- create the classifier based on labeled data L

Hình 2. Thuật toán SSLCA.

Bộ dữ liệu UCI: Kho lưu trữ UCI hiện đang duy trì 239 bộ dữ liệu như một dịch vụ cho cộng đồng học máy. Chúng tôi chọn sáu các tập dữ liệu đều có điểm chung với hai lớp: vú-w, credit-a, bệnh tiểu đường, tầng điện ly, kr-vs-kp, sonar. Đối với mỗi tập dữ liệu, chúng tôi đã giữ nhãn và đưa chúng vào nhóm dữ liệu được gắn nhãn L, xóa nhãn và đưa chúng vào nhóm dữ liệu không được gắn nhãn U, tỷ lệ giữa dữ liệu được gắn nhãn và dữ liệu không được gắn nhãn có thể được đặt trong thử nghiệm tiếp theo.

Bộ dữ liệu Coreference của Trung Quốc: chúng tôi sử dụng Lancaster Corpus của Tiếng Quan Thoại làm dữ liệu Coreference, sau đó xây dựng làm nổi bật và gắn nhãn mối quan hệ hội nghị bằng opera thủ công. Trong trích xuất tính năng, chúng tôi trích xuất 14 tính năng theo ngữ nghĩa, chứa đặc điểm đơn hoặc số nhiều, giới tính tính năng, tính năng cấu trúc câu, et al. Khác, chúng tôi trích xuất từ và tính năng Part-of-Speech trong vòng năm chiến thắng của antecedent và anaphor (Zhang, Guo, Yu, Zhang, & Yao, 2009). Trong thử nghiệm tiếp theo, chúng tôi cũng tách hai các loại đối tượng địa lý thành hai chế độ xem và kết hợp hai chế độ xem này cùng nhau để tạo tập dữ liệu một chế độ xem.

4.2. Thiết lập thí nghiệm

Hiệu suất của thuật toán SSLCA được so sánh với hai thuật toán học bán giám sát. Thuật toán so sánh đầu tiên là tiêu chuẩn đồng đào tạo (Blum và Mitchell, 1998). Hơn nữa, SSLCA được so sánh với một thuật toán học bán siêu hiển thị nổi tiếng khác TSVM (Joachims, 1999). Không giống như tiêu chuẩn đồng đào tạo ban đầu đào tạo một bộ phân loại trên dữ liệu được gắn nhãn và sau đó tăng cường thêm bộ đào tạo được gắn nhãn của nó bằng cách thêm một số phiên bản không được gắn nhãn mới được gắn nhãn với hầu hết các dự đoán đáng tin cậy về sở hữu. Thuật toán bán giám sát TSVM ban đầu đào tạo một bộ phân loại trên dữ liệu được gắn nhãn và dữ liệu không được gắn nhãn, khai thác cụm cấu trúc của dữ liệu và coi nó như kiến thức trước đây về việc học nhiệm vụ. Thuật toán SSLCA sử dụng đồng đào tạo ban đầu đào tạo một phân loại tốt hơn trên dữ liệu được gắn nhãn sốt, không chỉ sử dụng đồng đào tạo để đưa ra các phiên bản có độ tin cậy cao nhằm thúc đẩy bộ phân loại, mà cũng khai thác các phiên bản thông tin cho trình chú thích của con người.

Đối với bất kỳ thuật toán so sánh nào, một số loại thuật toán học được sử dụng để thực hiện quy nạp bộ phân loại. Đặc biệt, thuật toán Naïve Bayes, thuật toán SMO và thuật toán TSVM được sử dụng, mà hai thuật toán trước đây đến từ nền tảng weka và svmlight được sử dụng làm thuật toán TSVM, mà đến từ tác phẩm của Martin Theobald (Theobald, 2013). Fur thermore, ba thuật toán được sử dụng làm thuật toán cơ sở cho mục đích tham khảo. Các thuật toán Naïve Bayes và SMO chỉ đào tạo các phân loại trên các phiên bản đào tạo được gắn nhãn ban đầu trong khi TSVM đào tạo bộ phân loại trên các cá thể được gắn nhãn cùng với không gắn nhãn cho giả định nhập cụm. Đối với đồng đào tạo, học tập tích cực và nhịp điệu thuật toán SSLCA được trang bị bất kỳ bộ cảm ứng phân loại nào, 100 độc lập các lần chạy được thực hiện dưới mọi cấu hình của '. Đồng đào tạo và các thuật toán học tập tích cực được trình bày chi tiết trong Phần 2; Thuật toán SSLCA được trình bày chi tiết trong Phần 3.

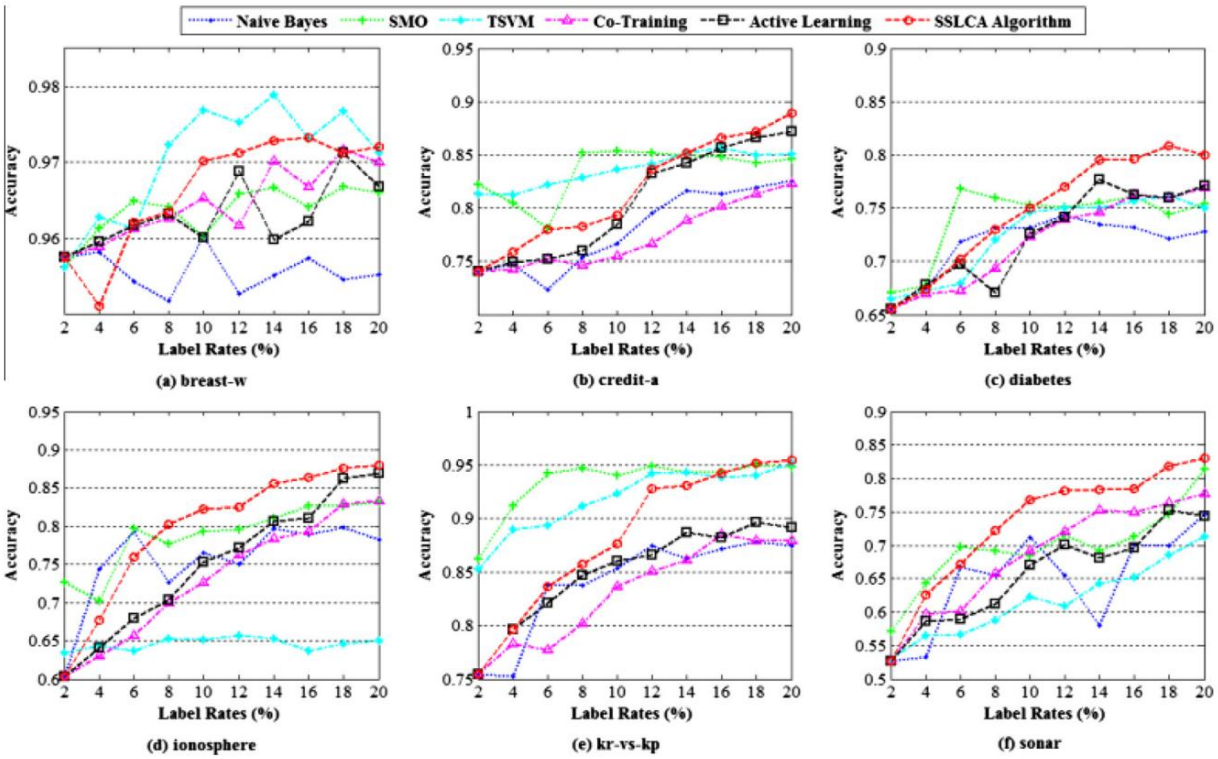
4.3. Kết quả thử nghiệm trên bộ dữ liệu UCI

Bây giờ chúng tôi trình bày kết quả thử nghiệm của chúng tôi. Sáu bộ dữ liệu chúng tôi được sử dụng từ Kho lưu trữ Học máy UCI. Đối với tất cả sáu dữ liệu bộ chúng tôi chọn để tăng số lượng dữ liệu được gắn nhãn được cung cấp, sử dụng phần còn lại làm dữ liệu chưa được gắn nhãn. Hình 3 và 4 minh họa cách mỗi so sánh thuật toán thực hiện với các phương pháp ước lượng độ tin cậy khác nhau, với số lượng các trường hợp huấn luyện được gắn nhãn trong các nếp gấp và các bộ phân loại cơ sở khác nhau.

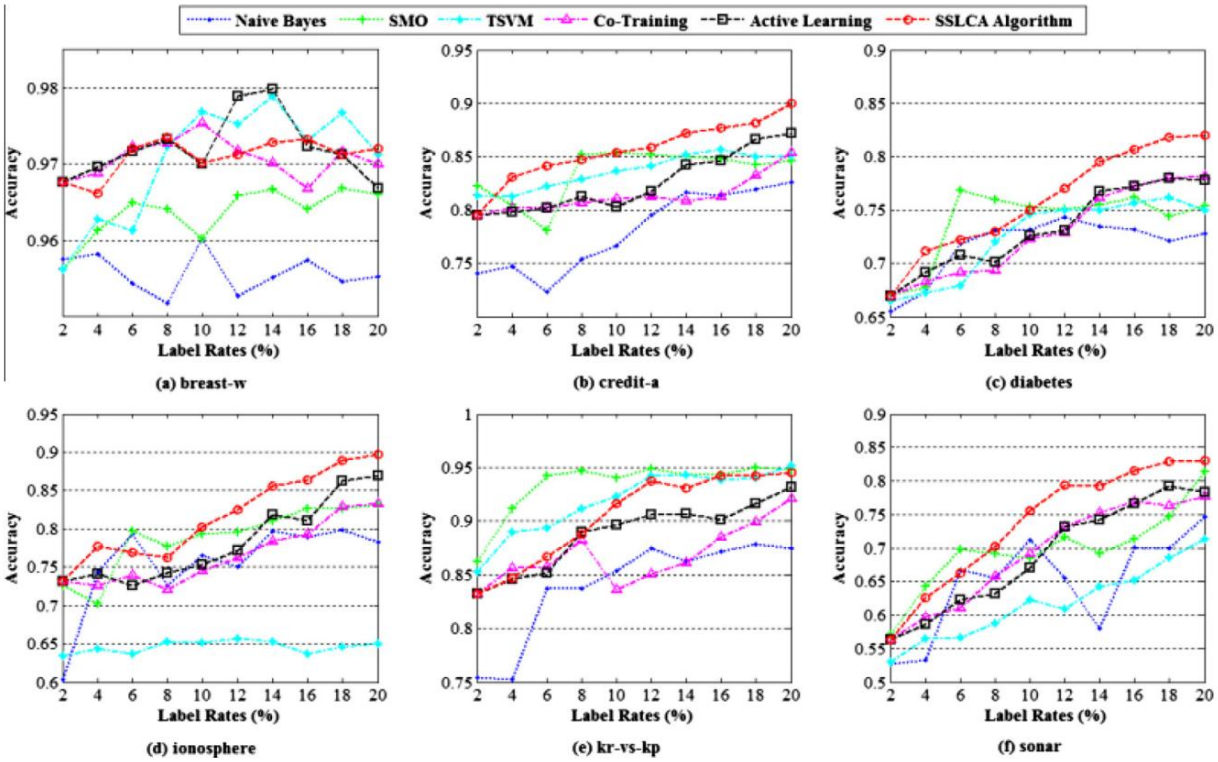
Trong Hình 3, Naïve Bayes được sử dụng làm ước lượng độ tin cậy phương pháp; Ngoài ra, nó còn được sử dụng làm bộ phân loại cơ sở của đồng đào tạo, hoạt động học và thuật toán SSLCA. Trên tập dữ liệu vú, TSVM nhận được hiệu suất tốt nhất với sự gia tăng của các phiên bản được gắn nhãn. Chi tiết, trên bộ dữ liệu tín dụng, thuật toán SSLCA có kết quả tốt hơn khi tỷ lệ dữ liệu được gắn nhãn là hơn 14%. Về bệnh tiểu đường, tầng điện ly và tập dữ liệu sonar, thuật toán SSLCA vượt trội hơn so với các thuật toán với tỷ lệ dữ liệu được gắn nhãn tăng đến mức nhất định. Trên tập dữ liệu kr-vs-kp, thuật toán SMO và thuật toán TSVM nhận được hoạt động tốt nhất khi dữ liệu được gắn nhãn nhỏ và thuật ngữ SSLCA có được hiệu suất tốt nhất khi sự gia tăng của tập dữ liệu được gắn nhãn.

Trong Hình 4, Tối đa hóa kỳ vọng được sử dụng làm độ tin cậy phương pháp ước lượng và thuật toán SMO được sử dụng làm cơ sở bộ phân loại của đồng đào tạo, học tập tích cực và thuật toán SSLCA. Trên tập dữ liệu vú-w, TSVM có được hiệu suất tốt nhất với sự gia tăng trong số các phiên bản được gắn nhãn. Về bệnh tiểu đường, tập dữ liệu tầng điện ly và sóng siêu âm, thuật toán SSLCA có được hiệu suất tốt hơn đáng kể so với các thuật toán khác có tỷ lệ dữ liệu được gắn nhãn tăng lên mức độ nhất định. Trên bộ dữ liệu credit-a và kr-vs-kp, thuật toán SMO và thuật toán TSVM có được hiệu suất tốt nhất khi dữ liệu được gắn nhãn là nhỏ và thuật toán SSLCA có được hiệu suất tốt nhất với sự gia tăng của tập dữ liệu được gắn nhãn.

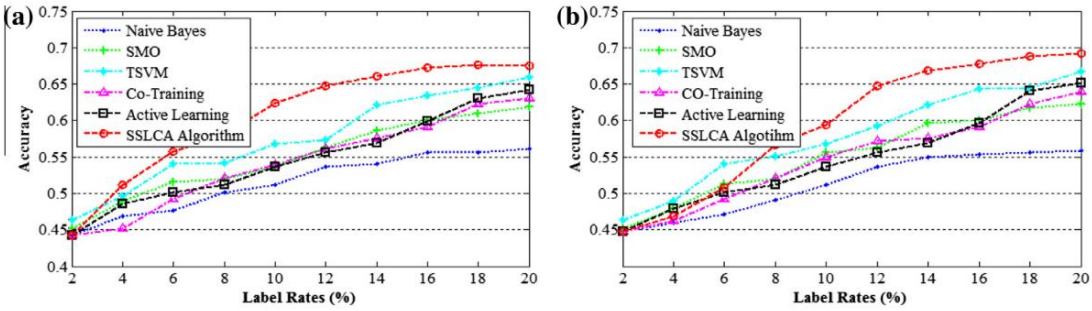
Như thể hiện trong Hình. 3 và 4, rõ ràng là thuật toán SSLAC là ưu việt hơn thuật toán được giám sát truyền thống trên được chọn ngẫu nhiên bộ dữ liệu, nó cũng vượt trội so với thuật toán đồng đào tạo truyền thống và thuật toán học tập tích cực trên bốn tập dữ liệu, trên hai tập dữ liệu còn lại, ít nhất nó có hiệu suất gần như tương tự. Tóm lại, Thuật toán SSLCA có hiệu quả về mặt thống kê theo tỷ lệ trên sáu tập dữ liệu được chọn ngẫu nhiên. Thuật toán áp dụng đồng đào tạo để chọn ra các trường hợp đáng tin cậy nhất để thúc đẩy trình phân loại, nhưng cũng không loại bỏ các trường hợp không chắc chắn, thường là các cá thể cung cấp thông tin cho bộ phân loại. Chúng tôi dán nhãn những phiên bản thông tin đó và thêm chúng vào nhóm dữ liệu được gắn nhãn L để thúc đẩy trình phân loại, giúp đạt được sự cải thiện đáng kể so với các thuật toán so sánh khác để hy sinh cùng một lượng nỗ lực của con người.



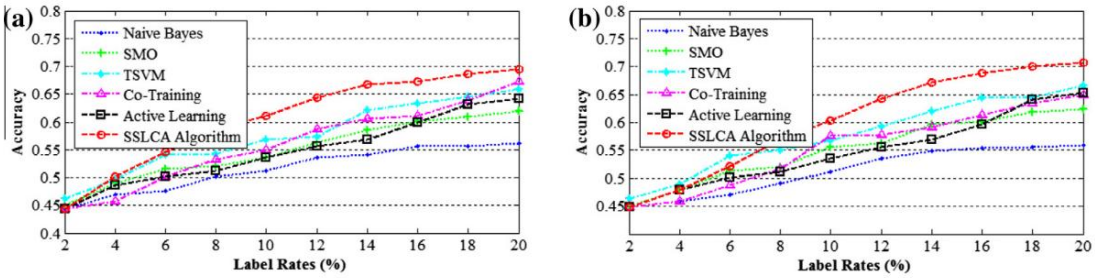
Hình 3. Độ chính xác phân loại của mỗi thuật toán so sánh thay đổi khi số lượng các phiên bản huấn luyện được gắn nhãn tăng lên, trong đó Naïve Bayes được sử dụng làm phương pháp ước tính độ tin cậy.



Hình 4. Độ chính xác phân loại của mỗi thuật toán so sánh thay đổi khi số lượng các phiên bản huấn luyện được gắn nhãn tăng lên, trong đó Tối đa hóa kỳ vọng được sử dụng làm phương pháp ước tính độ tin cậy.



Hình 5. Độ chính xác phân loại của mỗi thuật toán so sánh thay đổi khi số lượng phiên bản huấn luyện được gắn nhãn tăng lên, trong đó phương pháp tách entropy được sử dụng làm khung nhìn tách ra.



Hình 6. Độ chính xác phân loại của mỗi thuật toán so sánh thay đổi khi số lượng các phiên bản huấn luyện được gắn nhãn tăng lên, trong đó các chế độ xem phân tách tự nhiên được sử dụng làm hai quan điểm.

4.4. Ứng dụng vào Coreference của Trung Quốc

Tập dữ liệu Coreference của Trung Quốc được lấy từ Lancaster Corpus of Mandarin Chinese. Chúng tôi trích xuất 14 đặc điểm ngữ nghĩa và tính năng Word và Part-of-Speech trong năm cửa sổ, tạo thành hai chế độ xem dữ liệu thử nghiệm tương ứng trong đồng đào tạo và thuật toán SSLCA, chúng tôi cũng kết hợp hai quan điểm này cùng nhau để xây dựng tập dữ liệu một lần xem cho các thuật toán nhập so sánh khác.

Hình 5 và 6 minh họa cách mỗi thuật toán so sánh trên mỗi biểu mẫu với các phương pháp ước tính độ tin cậy khác nhau, khi số lượng các trường hợp huấn luyện được gắn nhãn tăng lên và cơ sở khác nhau bộ phân loại. Trong Hình 5, phương pháp tách entropy được sử dụng làm khung nhìn phương pháp phân tách, (a) Naïve Bayes được sử dụng làm phương pháp ước tính độ tin cậy và bộ phân loại cơ sở, (b) Tối đa hóa kỳ vọng là được sử dụng làm phương pháp ước tính độ tin cậy và SMO được sử dụng làm phân loại cơ sở của đồng đào tạo, học tập tích cực và SSLCA thuật toán. Trong hình . _ _

Như thể hiện trong Hình 5, thuật toán SSLAC không khác với so sánh các thuật toán khi dữ liệu được gắn nhãn là nhỏ, đặc biệt, Thuật toán TSVM vượt trội hơn tất cả các thuật toán so sánh khi tỷ lệ được gắn nhãn nhỏ hơn 6% trong biểu đồ (b), nhưng thuật toán SSLAC là luôn vượt trội hơn tất cả các thuật toán so sánh khi tỷ lệ của anh ta là hơn 8%. Trên thực tế, thông qua việc quan sát Hình 6 , chúng ta có thể nhận được cùng một kết luận rằng thuật toán SSLAC luôn vượt trội so với tất cả các thuật toán so sánh với sự gia tăng của dữ liệu được gắn nhãn khi Naïve Bayes hoặc SMO được kết hợp làm cơ sở bộ phân loại.

Ngoài cuộc thảo luận ở trên, chúng tôi còn so sánh thêm về sự tái tạo của Figs. 5 và 6, có thể thấy rằng thuật toán SSLCA thậm chí còn có kết quả tốt hơn trong Hình 6. Điều này cho thấy rằng chế độ xem phân tách meth od được sử dụng trong thuật toán đồng huấn luyện có thể đóng một vai trò quan trọng. Hơn nữa, các chế độ xem được phân chia tự nhiên theo các đặc điểm ngữ nghĩa và tính năng Word và Part-of-Speech đáng tin cậy hơn so với entropy phương pháp phân tách, nghĩa là, khi tồn tại đủ và dư thừa , việc sử dụng chúng một cách thích hợp sẽ mang lại lợi ích cho việc học theo hình thức trong đồng đào tạo và thuật toán SSLCA.

5. Kết luận

So với đào tạo đồng truyền thống yêu cầu tập dữ liệu có thể được tách hai quan điểm, và thỏa mãn các giả định đầy đủ và không sâu sắc. Trong thực tế, những điều kiện đó không dễ để Hoàn thành. Ngoài ra, đồng đào tạo thường chọn các trường hợp có độ tin cậy cao được gắn nhãn với nhãn lớp ước tính và thêm chúng vào các tập huấn luyện, không đảm bảo các trường hợp có độ tin cậy cao đã chọn này có giá trị hơn để cải thiện dự đoán sự chính xác. Trong bài báo này, chúng tôi đã trình bày một chiến lược học hỏi bán giám sát mới để khai thác dữ liệu không được gắn nhãn, kết hợp đào tạo đồng thời và học tập tích cực. Kết quả thí nghiệm chứng minh rằng thuật toán của chúng tôi đạt được sự cải thiện đáng kể khi hy sinh cùng một lượng nỗ lực của con người.

So với thuật toán đồng đào tạo truyền thống, phương pháp của chúng tôi không chỉ áp dụng đồng đào tạo để chọn ra sự tự tin đáng tin cậy nhất các trường hợp để tăng cường trình phân loại, nhưng cũng không loại bỏ các trường hợp không chắc chắn, thường là các trường hợp cung cấp thông tin vào bộ phân loại. So với nhịp điệu học tập tích cực truyền thống, chúng tôi xác định mức độ đóng góp là tiêu chí lựa chọn của các trường hợp mang tính thông tin, không chỉ xem xét phần yếu nhất của các trường hợp, mà còn xem xét sự khác biệt về độ không đảm bảo giữa cá thể và hàng xóm gần nhất của nó. Hơn nữa, nhịp điệu thuật toán của chúng tôi rất hữu ích để khai thác các trường hợp thông tin, điều này có thể được xác minh bằng thực tế rằng thuật toán SSLCA tỏ ra vượt trội hơn hẳn so với việc học tập tích cực với bộ phân loại cơ sở khác nhau. So sánh với đồng đào tạo được gọi chính xác khi tập hợp đào tạo nhỏ, thuật toán của chúng tôi hoạt động tốt trên các tập huấn luyện.

Trong tương lai, điều rất quan trọng là phải tiến hành các phân tích lý thuyết sâu sắc hơn về hiệu quả của phương pháp tiếp cận của chúng tôi. Hơn nữa, việc thiết kế cách tiếp cận phù hợp hơn với tư cách là tiêu chí lựa chọn của các trường hợp thông tin đáng được nghiên cứu. Cách tiếp cận tách đặc trưng hiệu quả khác cũng cần nỗ lực nghiên cứu và thảo luận.

Sự nhìn nhận

Các tác giả xin chân thành cảm ơn các phản biện ẩn danh đã có những ý kiến đóng góp quý báu để cải thiện đáng kể chất lượng của bài báo này. Nghiên cứu được báo cáo trong bài báo này đã được hỗ trợ một phần bởi Tổ chức Khoa học Quốc gia Trung Quốc theo Grant số 61379158, Ph.D. Chương trình Foundation của Bộ Giáo dục Trung Quốc số 20120191110028, Chương trình Nghiên cứu Cơ bản Trọng điểm Quốc gia Trung Quốc (973) No. 2013CB328903.

Người giới thiệu

Blum, A., & Mitchell, T. (1998). Kết hợp dữ liệu được gắn nhãn và không được gắn nhãn với đào tạo đồng [C]. Trong Kỷ yếu hội nghị thường niên lần thứ 11 về lý thuyết học tập tính toán (trang 92–100).

Chapelle, O., Scholkopf, B., & Zien, A. (2006). Học tập bán giám sát. Cambridge, MA: MIT Press.

Du, J., Ling, X., & Zhou, ZH (2011). Khi nào đồng đào tạo hoạt động trong dữ liệu thực [J]. Giao dịch IEEE về Kiến thức và Kỹ thuật Dữ liệu, 23 (5), 788–799.

Goldman, S & Zhou, Y. (2000). Tăng cường học tập có giám sát với dữ liệu không được gắn nhãn [C]. Trong Kỷ yếu hội nghị quốc tế lần thứ 17 về máy học (trang 327–334).

Joachims, T. (1999). Suy luận chuyển đổi cho máy vectơ hỗ trợ phân loại văn bản [C]. Trong Kỷ yếu hội nghị quốc tế lần thứ 6 về học máy (trang 200–209).

Li, H., Shi, Y., & Liu, Y. (2012). Phát hiện khái niệm video giữa nhiều miền: Một phương pháp học tập tích cực mang tính phân biệt và chung chung [J]. Hệ thống Chuyên gia với Ứng dụng, 39, 12220–12228.

Lughofer, E. (2012). Học tích cực kết hợp để giảm nỗ lực chú thích của các toán tử trong hệ thống phân loại [J]. Nhận dạng mẫu, 45 (2), 884–896.

Mihalcea, R. (2004). Đồng đào tạo và tự đào tạo để Phân biệt Nhận thức Từ [C] Đang tiếp tục CONLL-04.

Giải quyết, B. (2010). Khảo sát văn học tích cực học tập. Báo cáo Kỹ thuật Khoa học Máy tính 1648, Đại học Wisconsin tại Madison, ngày 26 tháng 1 năm 2010.

Sun, A., Liu, Y., & Lim, EP (2011). Phân loại web của các thực thể khái niệm bằng cách sử dụng co training [J]. Hệ thống Chuyên gia với Ứng dụng, 38 (12), 14367–14375.

Tang, F., Brennan, S., Zhao, Q., Tao, H., Cruz U S., & Cruz, S. (2007). Đồng theo dõi bằng cách sử dụng máy vectơ hỗ trợ bán giám sát [C]. Sắp tới hội nghị quốc tế lần thứ 11 về thị giác máy tính (trang 1–8).

Theobald, M. (2013). Chương trình của thuật toán svmlight. http: //www.mpi inf.mpg.de/~mtb/svmlight/JNI_SVM-light-6.01.zip.

Wang, W., & Zhou, ZH (2010). Một phân tích mới về đồng đào tạo [C]. Kỷ yếu hội nghị quốc tế lần thứ 27 về học máy, Haifa, Israel.

Witten, IH, & Frank, E. (2005). Khai phá dữ liệu: Các công cụ và kỹ thuật học máy thực tế. San Francisco: Morgan Kaufmann. http: //prdownloads.sourceforge. net / weka / datasets-UCI.jar ..

Yang, L., Jin, R., & Sukthankar, R. (2009). Học theo chỉ số khoảng cách chủ động của Bayes [C]. Trong Kỷ yếu hội nghị lần thứ 23 về sự không chắc chắn trong, trí tuệ nhân tạo (trang 442–449).

Yu, S., Krishnapuram, B., Rosales, R., & Rao, R. (2011). Đồng đào tạo Bayes [J]. Tạp chí Nghiên cứu Máy học, 12, 2649–2680.

Zhang, YH, Guo, JY, Yu, ZT, Zhang, ZK, & Yao, XM (2009). Nghiên cứu về độ phân giải lỗi tham chiếu của Trung Quốc dựa trên mô hình và quy tắc entropy cực đại [J], Trong bài giảng về khoa học máy tính. Tập 5854 (trang 1–8).

Zhang, ML & Zhou, ZH (2011). COTRADE: Tự tin đồng đào tạo về chỉnh sửa dữ liệu [J]. Giao dịch IEEE trên Hệ thống, Con người và Điều khiển học-Phần B: Điều khiển học, 41 (6), 1612–1626.

Zhou, ZH, & Li, M. (2005). Tri-training: Khai thác dữ liệu không được gắn nhãn bằng ba bộ phân loại [J]. Giao dịch IEEE về Kiến thức và Kỹ thuật Dữ liệu, 17 (11), 1529–1541.

Zhou, ZH, & Li, M. (2010). Học tập bán giám sát bởi sự bất đồng [J]. Hệ thống thông tin và tri thức, 24 (3), 415–439.

Zhu, XJ (2008). Khảo sát tài liệu học tập bán giám sát, Khoa học Máy tính TR 1530, Đại học Wisconsin tại Madison, ngày 19 tháng 7.

Zhu, XJ, Lafferty, J., & Ghahramani, Z. (2003). Kết hợp giữa học chủ động và học bán giám sát bằng cách sử dụng trường gaussian và hàm điều hòa [C] Trong Kỷ yếu hội thảo ICML-2003 về sự liên tục từ dữ liệu được gắn nhãn đến không được gắn nhãn, Washington DC.