Sử dụng Đồng đào tạo để nâng cao năng lực học tập chủ động

Aktif O¨ grenmeyi G uc¸lendirmek ic¸in Es¸-o¨grenme Kullanılması

Payam V. AZAD
Kỹ thuật máy tính Departmenet
Istanbul Technichal Univ ersity
Email: vakil@itu.edu.tr

Yusuf Yaslan
Kỹ thuật máy tính Departmenet
Đại học Công nghệ Istanbul
Email: yyaslan@itu.edu.tr

Tóm tắt — Học chủ động và đồng đào tạo là các trường hợp học bán giám sát, cả hai đều được sử dụng khi dữ liệu được gắn nhãn khan hiếm. Học tập tích cực cố gắng cải thiện mô hình học tập bằng cách truy vấn trên dữ liệu không được gắn nhãn và thách thức chính ở đó là tìm ra truy vấn phiên bản tối ưu. Và đồng đào tạo cố gắng khai thác hai các bộ tính năng khác nhau để phóng to số lượng dữ liệu được gắn nhãn mà không bất kỳ nhu cầu để có được thông tin bên ngoài. Một số nghiên cứu đã thử kết hợp hai phương pháp này và tận dụng tốt nhất chúng và chúng đạt được những kết quả đáng ghi nhận. Nhưng chúng tôi đã chứng kiến rằng việc sử dụng đồng đào tạo và học tập tích cực trong kiến trúc trình tự mang lại hiệu quả tốt hơn khi chúng đang làm việc song song. Sử dụng chúng theo trình tự có nghĩa là chúng tôi đã sử dụng các kỹ thuật đồng đào tạo để chỉ tìm ra các truy vấn tốt nhất để học tập tích cực chứ không phải trong chính quá trình học tập. Chúng tôi sẽ chứng minh rằng nó có kết quả tốt hơn so với học tập tích cực đơn thuần và đồng đào tạo và thậm chí cả các kiến trúc song song hiện tại. Đối với điều này chúng tôi đã sử dụng các kỹ thuật khác nhau để chia dữ liệu thành hai bộ dữ liệu riêng biệt; chúng tôi cũng sẽ thảo luận về nó cùng với truy vấn của chúng tôi phương pháp lựa chọn.

Ozetc¸e —Aktif o¨grenme (học tập tích cực) ve es¸-o¨grenme (đồng đào tạo), az sayıda etiketli Veriye sahip oldugumuzda etiketsiz olan veriden en iyi sonucu ortaya c¸ıkaran yontemlerdir. Aktif o¨grenme etiketsiz olan Veriyi tahmin etmeye c¸alısarak o¨grenme modelini iyiles¸tirmeye c¸alıs¸ır. Buradaki en onemli zorluk en etkili sorgulamanın bulunmasıdır. Es¸-o¨grenme ise herhangi bir ek kaynaga ihtiyac¸ duymadan etiketli Veriyi c¸ogaltmak ic¸in birbirinden bagımsız ozellik k umelerinden ve kendi bas¸ına yeterli faydalanır. Bu iki yontemi aynı anda kullanan pek c¸ok bas¸arılı aras¸tırma yapılmıs¸tır. Fakat biz c¸alıs¸mamızda gord uk ki; es¸-o¨grenme ve aktif o¨grenme birbiri ardına kullanıldıgında aynı anda kullanılmalarından daha bas¸arılı sonuc¸lar vermektedir. C¸alıs¸mamızda es¸-o¨grenme yontem ve yapısını, do grudan o¨grenme as¸amasında kullanmak yerine, aktif o¨grenme ic¸in en iyi sorgulamanın bulunmasında kullanıldık ve bu onerdi sadece aktif o¨grenme ve es¸-o¨grenme kullanıldıgında ve hatta paralel mimarilerde kullanıldıgında domainse edilen bas¸arıdan daha iyi sonuc¸lar verdigini gosterdik. Buna ek olarak, veri kumesini birbirinden farklı iki kumeye b olmek ic¸in farklı olc¸utler kullandık ve uyguladıgımız her sorgu sec¸me yontemi ic¸in bu olc¸utleri degerlendirdik.

Từ khóa — Học tập tích cực, đồng đào tạo, học máy, học bán giám sát.

Anahtar Kelimeler — aktif o¨grenme, es¸-o¨grenme, makine o¨grenmesi, yarı-denetimli o¨grenme.

978-1-5090-6494-6 / 17 / $ 31,00 c 2017 IEEE

I. GIỚI THIỆU

Các thuật toán học bán giám sát là tập hợp các thuật toán nằm giữa các thuật toán học tập có giám sát sử dụng chỉ dữ liệu được gắn nhãn và các thuật toán học tập không được giám sát chỉ sử dụng dữ liệu không được gắn nhãn. Các thuật toán bán giám sát đang sử dụng cả dữ liệu được gắn nhãn và không được gắn nhãn trong quy trình học tập. Tích cực học như một thành viên của gia đình này, cố gắng làm giàu được gắn nhãn tập dữ liệu bằng cách truy vấn các trường hợp không được gắn nhãn từ người ngoài nguồn. Trong học tập tích cực, thách thức lớn nhất là tìm ra truy vấn tốt nhất. Nó có nghĩa là chọn các phiên bản không được gắn nhãn tốt nhất để truy vấn nhãn của nó và thêm chúng vào tập hợp được gắn nhãn, theo một cách rằng với các truy vấn tối thiểu sẽ có được sự cải thiện lớn nhất [1].

Trong tài liệu đồng đào tạo gốc [2] Blum et al. Đã đề xuất một phương pháp để phóng to tập dữ liệu được gắn nhãn và do đó cải thiện thuật toán học, khai thác hai tập dữ liệu riêng biệt. Trong này công việc mà họ đã phân loại các trang web internet, mà họ đã một số trang được gắn nhãn và một số lượng lớn các trang không được gắn nhãn trong tay. Họ đã cố gắng sử dụng hai loại khác nhau và gần như tập hợp các tính năng độc lập cho mỗi trang; bộ tính năng đầu tiên được trích xuất từ nội dung của trang và bộ tính năng thứ hai thu thập từ các trang đã liên kết đến trang này.

Một số công trình đã thực hiện khai thác các thuật toán này như [3], [4] nơi cả hai đều đã làm việc trên các tập dữ liệu không liên quan được chia nhỏ; nhưng Zhang và cộng sự. [5] đã giới thiệu một thuật toán mới mà họ có được gọi là SSCLA mà họ đã cố gắng thực hiện cả đồng đào tạo và học tích cực song song và bằng cách chia nhỏ một tập dữ liệu thành hai cái riêng biệt và cũng tính toán một số đo để tìm những trường hợp nhiều thông tin nhất, cải thiện cả hai cùng nhau.

Trong bài báo này, chúng tôi đã cố gắng sử dụng ý tưởng này; nhưng thay vì lắp ráp các thuật toán này song song, chúng tôi đã sử dụng chúng trong sự phối hợp. Ý tưởng là đầu tiên phân vùng tập dữ liệu đơn lẻ thành hai bộ tính năng sử dụng một số chỉ số, sau đó bằng cách đồng đào tạo, hãy thử để tìm độ chắc chắn và mức độ đóng góp của từng dự đoán ví dụ, sau đó sử dụng các giá trị này, chúng tôi chọn phiên bản tốt nhất để truy vấn. Trong các thí nghiệm, chúng tôi đã sử dụng các phương pháp khác nhau để phân chia tập dữ liệu của chúng tôi, bao gồm cả mức tăng thông tin, chi-square và ANOVA.

TABLO I: DataSets

	số phiên bản tính năng	số lớp số 698 2 3195 2	207 2
cung Cự Giải		10	
cờ vua		37	
sonar		61	
tầng	389	16	2
điện ly tín dụng	350	35	2

II. PHƯƠNG PHÁ P

Thuật toán của chúng tôi là một thuật toán lặp được đề xuất gồm bốn bước, 1) chia dữ liệu được gắn nhãn và không được gắn nhãn thành hai tập hợp khác nhau L1, L2 và U1, U2 bằng cách sử dụng phương pháp tách, ví dụ: độ lợi thông tin hoặc chi-square. 2) huấn luyện hai tập dữ liệu có nhãn này bằng một thuật toán l và tạo hai bộ phân loại h1 và h2. 3) Sử dụng các bộ phân loại này để dự đoán nhãn của dữ liệu không được gắn nhãn. 4) Dựa trên một số tiêu chí, chọn trường hợp không được gắn nhãn nhiều thông tin nhất, và truy vấn họ buộc tội nhãn của trường hợp này và thêm nó vào tập hợp được gắn nhãn. Trong lần lặp tiếp theo của thuật toán, chúng tôi sẽ có dữ liệu mới được gắn nhãn trong tập hợp được gắn nhãn của chúng tôi và nó sẽ dẫn đến mô hình mạnh mẽ hơn và dự đoán tốt hơn.

A. Phân vùng tính năng

Thuật toán đồng đào tạo đang hoạt động dựa trên một logic rằng chúng ta có hai tập hợp dữ liệu được gắn nhãn và chưa được gắn nhãn độc lập và tự túc khác nhau. Tại nghiên cứu đồng đào tạo ban đầu [2] dữ liệu vốn có từ hai nguồn độc lập khác nhau, các liên kết đến một trang web và nội dung của trang web. Nhưng trong khi trong bộ dữ liệu của chúng tôi, Zhang et al. đề xuất [5] chúng tôi đang cố gắng phân vùng dữ liệu của mình thành hai tập hợp khác nhau và giả định một cách ngây thơ rằng chúng độc lập với nhau. Chúng tôi đã sử dụng ba phương pháp khác nhau: Tăng thông tin, thống kê Chi-Square và ANOVA (Phân tích phương sai) để chia tập dữ liệu của chúng tôi thành hai tập con.

1) Tăng thông tin: Bằng cách sử dụng thu được thông tin, một tập dữ liệu đã được chia thành hai tập con có lượng thông tin khá giống nhau. Trước tiên, chúng tôi đã tính toán mức tăng thông tin của từng tính năng và sắp xếp chúng theo mức tăng thông tin của chúng, sau đó chỉ định tính năng đầu tiên, thứ ba, thứ năm, v.v. cho tập hợp đầu tiên và gán tính năng thứ hai, thứ tư, thứ sáu, v.v. cho tập dữ liệu thứ hai. Bằng cách này, chúng ta sẽ có 2 bộ tính năng khác nhau mang lượng thông tin gần như giống nhau.
Nhược điểm lớn nhất của phương pháp này (hoặc các phương pháp phân vùng khác) là khả năng mất thông tin tồn tại trong các đối tượng liên kết, tức là đối tượng có dữ liệu có giá trị khi chúng được liên kết với nhau [6].

2) Chi-Square: Thống kê Chi-Square (X2) tính toán tính độc lập của hai biến ngẫu nhiên [7]. Biện pháp này đã được sử dụng rộng rãi để lựa chọn đối tượng theo cách mà đối tượng địa lý độc lập nhất với nhãn lớp là đối tượng địa lý kém giá trị nhất. Chúng tôi đã sử dụng ý tưởng này để tách dữ liệu của mình và tìm ra các tính năng có giá trị nhất.

3) ANOVA: ANOVA (Phân tích phương sai) [8] là tính toán sự thay đổi của tập dữ liệu và phân tích mức độ tương tự của các biến thể này trong hai tập dữ liệu. Nó dẫn đến việc tìm bao nhiêu hai tập dữ liệu mang cùng một ý nghĩa. Trong khi sử dụng giá trị này, chúng tôi giữa mỗi đối tượng địa lý và nhãn lớp, các đối tượng địa lý có đóng góp nhiều nhất cho nhãn lớp có thể được suy ra.

B. Đào tạo người phân loại

Ở bước này, chúng tôi đã sử dụng Naive Bias làm bộ phân loại cơ sở. Là một phương pháp xác suất là tài sản cụ thể nhất của Naive Bayes đối với chúng tôi. Nó cung cấp độ tin cậy thực nghiệm về các kết quả giúp chúng tôi tính toán thẳng thắn về độ chắc chắn của các dự đoán. Sử dụng thuật toán này riêng biệt trên hai tập hợp tính năng, chúng tôi tạo bộ phân loại để dự đoán dữ liệu không được gắn nhãn với sự chắc chắn.

Có hai bộ dữ liệu khác nhau, đưa ra hai dự đoán khác nhau (p1, p2) cho mỗi trường hợp và tâm sự của họ (xác suất trong trường hợp của chúng tôi: prob1, prob2). Chúng tôi đã sử dụng một số loại biểu quyết để có được dự đoán xác định cho từng trường hợp theo cách mà xác suất hoặc độ chắc chắn của mỗi dự đoán cũng được tính đến. Nếu cả hai bộ phân loại đều dự đoán cùng một lớp, lựa chọn cuối cùng là hiển nhiên nhưng trong trường hợp chúng có các dự đoán khác nhau, chúng tôi chấp nhận quyết định của một bộ phân loại dự đoán xác suất cao hơn cho mỗi trường hợp.

Đối với đào tạo phương pháp học tập lặp đi lặp lại đã được sử dụng; nó có nghĩa là chúng ta đã phù hợp một phần mô hình với tập huấn luyện vì nó sẽ cung cấp cho chúng ta sức mạnh để thêm lặp đi lặp lại các trường hợp mới mà không cần phải huấn luyện lại toàn bộ mô hình từ đầu. Nó đào tạo mô hình theo lô, lô đầu tiên trong trường hợp của chúng tôi là toàn bộ tập dữ liệu được gắn nhãn và các lô tiếp theo là các phiên bản được chọn từ tập hợp không được gắn nhãn bằng thuật toán học tập tích cực.

C. Lựa chọn truy vấn

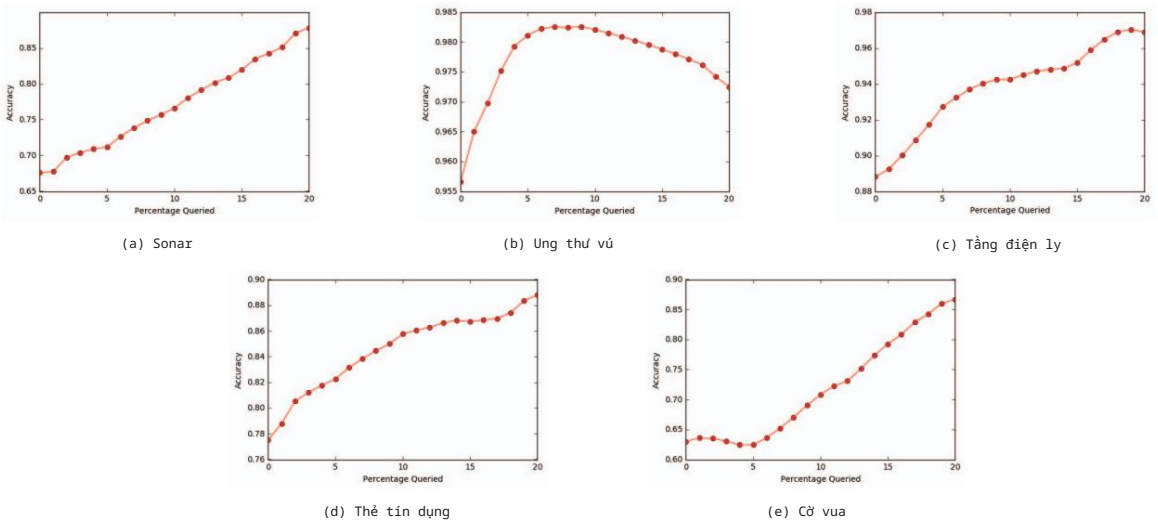
Người học tích cực nói chung sử dụng trường hợp không chắc chắn nhất để truy vấn nhằm tìm kiếm thông tin nhiều nhất. Sự không chắc chắn này là điều dễ hiểu đối với các mô hình học tập theo xác suất. Trong các mô hình này ở phân loại nhị phân, những trường hợp mà xác suất chúng thuộc cả hai lớp là gần 0,5 là những trường hợp không chắc chắn nhất.

Nhưng chúng tôi đã sử dụng cách tiếp cận đã được đề xuất bởi SSLCA. Chúng tôi đang cố gắng tìm ra phiên bản nhiều thông tin nhất bằng cách sử dụng K-Nearest Neighbors. Trong cách tiếp cận này, những trường hợp đó mang nhiều thông tin nhất mà bản thân chúng là những thông tin nhất định nhưng những người hàng xóm của chúng có độ chắc chắn rất thấp; điều đó chống lại giả thuyết phân cụm và vùng lân cận. Nó có nghĩa là nó có thể đóng góp thông tin rất lớn vào tập dữ liệu nếu nhãn đã được đoán. Tại thời điểm này, chúng tôi đang sử dụng phương pháp đồng đào tạo để tính toán giá trị đóng góp như sau:

$$\text{Đóng góp (Conf, xi)} = \frac{1}{M \quad \text{Conf (xi, c))} +} \quad \alpha \quad [\text{Conf (xi, c)} - \text{Conf (x, c)}] \quad (1)$$

$$x \ N \ (xi)$$

Trong đó, M biểu thị số lượng bộ phân loại, trong trường hợp của chúng ta là hai, Conf (xi, c) biểu thị độ tin cậy của ví dụ xi thuộc loại c, a biểu thị trọng số của mức độ đóng góp, N (xi) biểu thị mức độ láng giềng gần nhất của ví dụ xi.
Nó dựa trên ý tưởng rằng nếu độ tin cậy của đối tượng cao, nhưng độ tin cậy của người hàng xóm của nó thấp, thì



(a) Sonar (b) Ung thư vú (c) Tầng điện ly

(d) Thẻ tín dụng (e) Cờ vua

S, ekil 1: Kết quả thuật toán được đề xuất sử dụng dữ liệu không được gắn nhãn

không đáp ứng giả thuyết phân cụm; nếu không, những người có khả năng cao hơn là các trường hợp cung cấp thông tin vì mức độ đóng góp. những trường hợp này là giá trị nhất thể hiện để truy vấn từ toán tử [5].

Sau đó, trong mỗi lần lặp lại, một phần của các dự đoán kém tin cậy nhất hoặc các trường hợp có tỷ lệ đóng góp cao nhất sẽ được truy vấn từ tiên tri. Hiện tại, chúng tôi đã xem xét phần này để có kích thước chỉ bằng 1, điều đó có nghĩa là chúng tôi đã truy vấn chỉ một và thêm nó để huấn luyện trong mỗi lần lặp lại. Và sau đó cập nhật mô hình của chúng tôi bằng cách gọi phù hợp một phần (học lặp đi lặp lại) chức năng thêm các phiên bản mới vào mô hình của chúng tôi mà không cần để tìm hiểu các trường hợp đã học trước đó.

TABLO II: Kết quả của các thuật toán khác nhau với 70% được đào tạo dữ liệu so với thuật toán của chúng tôi với 50% + 20% dữ liệu đào tạo

	cung cự giải	cờ	tầng điện ly	tín dụng	sonar
AdaBoost	0,957	vua	0,738	0,778	0,869
Cây quyết định	0,926	0,964	0,710	0,832	0,861
Quy trình Gaussian	0,955	**0,942**	0,752	0,534	0,857
SVM tuyến tính	0,967	0,984	0,676	0,828	0,849
Naive Bayes	0,959	0,916	0,6090	0,621	0,810	0,865
Những người hàng xóm gần nhất	0,959	0,894	0,676	0,677	0,800
Mạng thần kinh	0,941	0,966	0,752	0,685	0,894
QDA	0,955	0,667	0,593	0,429	0,849
RBF SVM	0,896	0,515	0,759	0,567	0,739
Rừng ngẫu nhiên	0,951	0,842	0,669	0,770	0,902
AL + NB	0,972	0,867	0,878	0,888	0,969

III. THỰC HIỆN

Các thử nghiệm được thực hiện trên 5 tập dữ liệu khác nhau từ UCI. UCI kho lưu trữ hiện đang duy trì các tập dữ liệu khác nhau như một dịch vụ cho cộng đồng học máy. Chúng tôi đã chọn năm tập dữ liệu mà tất cả đều là vấn đề phân loại nhị phân. Các bộ dữ liệu này là: 1) Wisconsin Ung thư vú [9], 2) Phê duyệt Tín dụng [10], 3) Tầng điện ly [11], 4) Cờ vua (King-Rook vs. King-Pawn) [12] và 5) Ghế dài kết nối (sonar) [13]. Đối với mỗi tập dữ liệu, chúng tôi 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 các thí nghiệm tiếp theo (Bảng. ??).

Đối với tất cả các tập dữ liệu, chúng tôi đã triển khai lớp gấp 10 lần xác nhận và chia 10% cho thử nghiệm ở mỗi lần. Chúng ta có chia 90% đoạn tàu còn lại thành hai nhóm 50% đoàn tàu và xác nhận 40%. Sau đó, bắt đầu học tập tích cực lặp đi lặp lại cho 20% dữ liệu. Điều đó dẫn đến giảm 20% so với xác thực và sau khi được truy vấn, hãy thêm nó vào tập huấn luyện do đó ở cuối trong quá trình lặp lại, chúng tôi sẽ có 70% đào tạo, 20% xác thực và 10% test trong tay.

IV. CÁ C KẾT QUẢ

A. Kết quả tham khảo

Như đã thảo luận, tập dữ liệu được đào tạo với 50% dữ liệu và sau đó tăng nó thành 70% dữ liệu. Chúng tôi đã làm tất cả xử lý bằng trình phân loại Naive Bayes rất đơn giản (sau này chúng tôi sẽ cải thiện chính người huấn luyện) vì đơn giản và có kết quả đồng nhất độc lập với quyền lực của bộ phân loại.

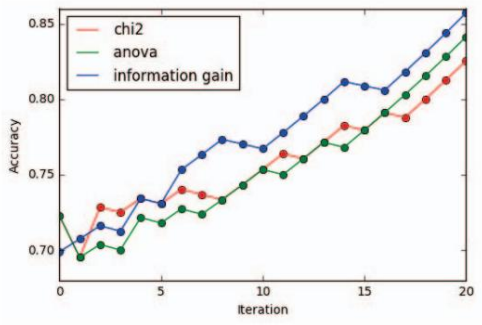
Như một tài liệu tham khảo, chúng tôi đang giữ phần này và cố gắng kiểm tra thuật toán của chúng tôi so với các thuật toán đã được thiết lập tốt khác. Vì vậy, chúng tôi đã sử dụng 70% bộ dữ liệu của mình để đào tạo bằng cách sử dụng một số các thuật toán mạnh mẽ như Random Forrest, SVN, Gaussian Quy trình và Mạng thần kinh cho mục đích này (Bảng ??). Chắc chắn rằng chúng tôi không nhằm mục đích so sánh thuật toán của chúng tôi với các thuật toán này bởi vì chúng có đặc điểm khác nhau và chúng tôi có các truy vấn tùy chỉnh không có sẵn trong các thuật toán. Nhưng nó chỉ cho chúng ta một cái nhìn sơ lược về liệu 70% dữ liệu được chọn làm đào tạo một cách ngẫu nhiên, có thể so sánh với chọn 50% dữ liệu, đào tạo nó bằng thuật toán đơn giản và sau đó tùy chỉnh chọn thêm 20% phần trăm dữ liệu và đạt 70%.

Như bạn có thể thấy, chúng tôi cũng bao gồm kết quả của Naive Bayes mà không có Học tập tích cực, trong hầu hết các trường hợp, rất yếu

hơn các thuật toán khác nhưng khi nó được kết hợp với phiên bản của chúng tôi của học tập tích cực, nó vượt qua các thuật toán mạnh mẽ khác một cách dễ dàng trong tất cả các bộ dữ liệu ngoại trừ cờ vua. Lý do cho việc không được mong đợi kết quả trong tập dữ liệu cờ vua là các tính năng của tập dữ liệu này là tương quan cao và cố gắng tách đặc điểm của nó thành hai tập hợp con, chúng tôi đang mất thông tin này đặt trong các mối quan hệ. Chúng ta có thể thấy rằng trong tập dữ liệu này Quy trình Gaussian [14] hoạt động rất tốt vì nó cố gắng xem xét tất cả các mối quan hệ giữa các tính năng trong một ma trận lớn hoặc có mạng nơ-ron với cấu trúc phức tạp của nó cố gắng trích xuất mối quan hệ giữa các bên trong Tính năng, đặc điểm.

B. Kết quả phân vùng tính năng

Như đã đề cập, Mức tăng thông tin, Thống kê Chi-Square và ANOVA được sử dụng để phân vùng tập dữ liệu thành hai tập hợp con. Trong hầu hết mọi trường hợp, Mức tăng thông tin đạt được tốt hơn các kết quả. Vì vậy, trong phần còn lại của báo cáo, chúng tôi sẽ chỉ sử dụng Thông tin Thu được. Để tham khảo, chúng tôi chỉ cho thấy sự khác biệt ở Hình 2 đối với tập dữ liệu sonar thu được với ba phương pháp phân vùng khác nhau.



S¸ ekil 2: So sánh các phương pháp phân vùng

C. Kết quả thuật toán được đề xuất

Kết quả trung bình cho xác nhận chéo 10 lần cho mỗi tập dữ liệu có sẵn trong hình 1. Kết quả cho biết cải thiện đạt được khi thử nghiệm trong khi từng phần trăm dữ liệu được thêm vào từ tập xác nhận thành tập huấn luyện.

Ngoài ra, chúng tôi đã cố gắng so sánh kết quả của mình với kết quả SSLCA trong Bảng III. Mặc dù họ không báo cáo kết quả của họ trong giá trị định lượng; Chúng tôi chỉ có thể trích xuất đại khái kết quả thành số từ các số liệu của họ. Từ những con số nó có thể giả định rằng ngoại trừ tập dữ liệu cờ vua, chúng tôi luôn có cải thiện kết quả so với SSLCA. Lưu ý rằng dữ liệu không được gắn nhãn kích thước phân vùng và loại xác thực chéo không có sẵn cho SSLCA dẫn đến [5].

V. THẢO LUẬN

Từ những thí nghiệm ban đầu cho thấy rằng đơn giản Naive Bayes chỉ sau 20% truy vấn từ oracle hoạt động tốt hơn nhiều hơn hầu hết các thuật toán phức tạp khác có cùng một số phiên bản để đào tạo. Điều này cho thấy hiệu quả và tầm quan trọng của các chiến lược truy vấn. Trong tương lai chúng ta có thể làm việc về các thuật toán truy vấn và tính toán bí mật tốt hơn và Ngoài ra, chúng tôi sẽ thử điều này với các thuật toán học tập khác. chúng tôi đang có kế hoạch mở rộng nghiên cứu này cho phức tạp hơn bộ dữ liệu không chỉ có các lớp nhị phân mà còn cho nhiều nhãn bộ dữ liệu.

TABLO III: So sánh kết quả của chúng tôi với SSLCA.

	SSLCA		Thuật toán của chúng tôi
	khởi đầu	chấm dứt	bắt đầu kết thúc
cung cự giải	0,958	0,97	0,957 0,972
cờ vua	0,75	0,95	0,62 0,87
tầng	0,53	0,83	0,68 0,88
điện	0,745	0,88	0,80 0,89
ly tín dụng sonar	0,6	0,875	0,88 0,97

Ngoài ra còn có khả năng sử dụng các thuật toán hồi quy và sau đó sử dụng kết quả hồi quy để rút ra cả hai phân loại và sự tự tin; chỉ cần có một số hiệu chuẩn (chuẩn hóa đầu ra thành phạm vi 0-1) để làm cho nó hoạt động hợp lý và đúng cách [15]. Trong các tác phẩm trong tương lai, chúng tôi cũng đang có kế hoạch cũng được sử dụng phương pháp này.

KAYNAKC¸ A

[1] B. Dàn xếp, “Khảo sát văn học học tập tích cực,” Đại học Wisconsin, Madison, tập. 52, không. 55-66, tr. 11 năm 2010.

[2] A. Blum và T. Mitchell, “Kết hợp dữ liệu được gắn nhãn và không được gắn nhãn với sự đồng đào tạo, ”trong Kỷ yếu của hội nghị thường niên lần thứ 11 về Lý thuyết học tập tính toán. ACM, 1998, trang 92–100.

[3] I. Muslea, S. Minton và CA Knoblock, “Hoạt động + bán giám sát learning = học tập đa góc nhìn mạnh mẽ, ”trong ICML, vol. 2, 2002, trang 435–442.

[4] C.-H. Mao, H.-M. Lee, D. Parikh, T. Chen và S.-Y. Huang, “Phương pháp tiếp cận dựa trên học tập tích cực và đồng đào tạo có giám sát để phát hiện xâm nhập từ nhiều chế độ xem,” trong Kỷ yếu của hội nghị chuyên đề ACM 2009 trên Máy tính Ứng dụng. ACM, 2009, trang 2042–2048.

[5] Y. Zhang, J. Wen, X. Wang, và Z. Jiang, “Học tập bán giám sát kết hợp đồng đào tạo với học tập tích cực, “Expert Systems with Ap plications, vol. 41, không. 5, trang 2372–2378, 2014.

[6] CM Bishop, “Nhận dạng mẫu”, Học máy, tập. 128, tr. 1–58, 2006.

[7] Z. Zheng, X. Wu và R. Srihari, “Lựa chọn tính năng cho phân loại văn bản trên dữ liệu không cân bằng,” Bản tin Khám phá ACM Sigkdd, tập. 6, không. 1, trang 80–89, 2004.

[8] I. Guyon và A. Elisseeff, “Giới thiệu về biến và tính năng lựa chọn, ”Tạp chí nghiên cứu máy học, tập. 3, không. Tháng 3, pp. 1157–1182, 2003.

[9] WH Wolberg và OL Mangasarian, “Phương pháp đa bề mặt của tách mẫu để chẩn đoán y tế áp dụng cho tế bào học vú. ” Kỷ yếu Viện Hàn lâm Khoa học Quốc gia, tập. 87, không. 23, tr. 9193–9196, 1990.

[10] JR Quinlan, “Đơn giản hóa cây quyết định,” tạp chí quốc tế về nghiên cứu con người-máy, tập. 27, không. 3, trang 221–234, 1987.

[11] VG Sigillito, SP Wing, LV Hutton và KB Baker, “Phân loại của radar trả về từ tầng điện ly bằng cách sử dụng mạng nơ-ron, ”Johns Thông báo kỹ thuật APL của Hopkins, tập. 10, không. 3, trang 262–266, 1989.

[12] AD Shapiro, Cảm ứng có cấu trúc trong hệ thống chuyên gia. Addison-Wesley Longman Publishing Co., Inc., 1987.

[13] RP Gorman và TJ Sejnowski, “Phân tích các đơn vị ẩn trong một lớp mạng được huấn luyện để phân loại các mục tiêu sonar, ”Mạng lưới thần kinh, tập. 1, không. 1, trang 75–89, 1988.

[14] CE Rasmussen, “Các quy trình Gaussian cho học máy,” 2006.

[15] R. Caruana và A. Niculescu-Mizil, "Một so sánh thực nghiệm của các thuật toán học tập theo phương pháp su pervised", trong Kỷ yếu của quốc tế lần thứ 23 hội thảo về Máy học. ACM, 2006, trang 161–168.