TRƯỜNG ĐẠI HỌC KỸ THUẬT ISTANBUL F TỐT NGHIỆP KHOA HỌC

KỸ THUẬT VÀ CÔNG NGHỆ

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

M.Sc. ĐỀ TÀI

Payam VAKIL ZADEH AZAD

Khoa Kỹ thuật Máy tính

Chương trình Kỹ thuật Máy tính

Tháng 10 năm 2017

TRƯỜNG ĐẠI HỌC KỸ THUẬT ISTANBUL F TỐT NGHIỆP KHOA HỌC

KỸ THUẬT VÀ CÔNG NGHỆ

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

M.Sc. ĐỀ TÀI

Payam VAKIL ZADEH AZAD (504111553)

Khoa Kỹ thuật Máy tính

Chương trình Kỹ thuật Máy tính

Cố vấn luận văn: Yar. Doç. Tiến sĩ Yusuf YASLAN

Tháng 10 năm 2017

˙ISTANBUL TEKN˙IK ÜN˙IVERS˙ITES˙I F FEN B˙IL˙IMLER˙I ENST˙ITÜSÜ

Aktif Ögrenmeyi Güçlendırmek için E¸s-ö grenme Kullanılması

YÜKSEK L˙ISANS TEZ˙I

Payam VAKIL ZADEH AZAD
(504111553)

Bilgisayar Mühendisligi Anabilim Dalı

Chương trình Bilgisayar Mühendisligiı

Tez Danı¸smanı: Vâng. Doç. Tiến sĩ Yusuf YASLAN

Ekim 2017

Payam VAKIL ZADEH AZAD, một M.Sc. sinh viên Khoa Khoa học Cao học ITU
Kỹ thuật và Công nghệ 504111553 đã bảo vệ thành công luận án mang tên “Chúng tôi đồng đào tạo để trao quyền cho việc học chủ động”, mà anh / cô ấy đã chuẩn bị sau khi hoàn thành các yêu cầu được quy định trong các luật liên quan, trước bồi thẩm đoàn có nhiệm vụ ký tên dưới đây.

Cố vấn luận án :	Vâng. Doç. Tiến sĩ Yusuf YASLAN Đại học Kỹ thuật Istanbul	.....
Thành viên Ban giám khảo:	PGS.TS Hatice Köse Đại học Kỹ thuật Istanbul	.....
	Trợ lý GS.TS Serap Kırbız ¸Sim¸sek ..... Đại học MEF	
		.....

Ngày gửi: 6 tháng 9 năm 2017
Ngày Quốc phòng: 6 tháng 10, 2017

LỜI TỰA

Tôi muốn cảm ơn giáo sư của tôi, người đã dẫn dắt tôi trong tất cả các bước của nghiên cứu này và luôn là một người bạn tốt đối với tôi.

Tháng Mười

Payam VAKIL ZADEH AZAD

MỤC LỤC

Trang

CÁC TỪ VIẾT TẮT

ANOVA: Phân tích phương sai

CEAL	: Đồng đào tạo Học tập tích cực nâng cao
DT	: Cây quyết định
EDF	: Định dạng dữ liệu Châu Âu
EM	: Tối đa hóa kỳ vọng
EMD	: Phân rã chế độ thực nghiệm
FFT	: Biến đổi Fourier nhanh
GNB	: Gaussian Naive Bayes
IG	: Tăng thông tin
KDE	: Ước tính mật độ hạt nhân
KNN	: K-Nearest Neighbor
MLP	: Perceptron nhiều lớp
NB	: Naive Bayes
NN	: Mạng thần kinh
RF	: Rừng ngẫu nhiên
SVM	: Máy hỗ trợ Vector
SSLCA	: Học tập bán giám sát kết hợp Đồng đào tạo với Học tập tích cực
UCI	: Đại học California, Irvine

DANH MỤC BẢNG BIỂU

DANH MỤC CÁC HÌ NH

Trang

TÓM LƯỢC

Một lượng lớn dữ liệu hiện có sẵn và điều quan trọng là phải trích xuất kiến thức bằng cách sử dụng lượng dữ liệu khổng lồ này. Tuy nhiên, chỉ một phần nhỏ của dữ liệu này được phân loại và gắn nhãn là kiến thức. Các thuật toán học máy thường được sử dụng để trích xuất kiến thức từ các bộ dữ liệu này; một nhóm các thuật toán học máy khai thác cả dữ liệu được gắn nhãn và không được gắn nhãn để nâng cao quy trình gắn nhãn được gọi là Học bán giám sát. Trong phương pháp bán giám sát, chúng tôi phát triển một mô hình, sau đó dựa trên các phiên bản có nhãn khan hiếm, cố gắng mở rộng và cải thiện nó bằng cách sử dụng nhiều phiên bản không được gắn nhãn. Học tập tích cực và Đồng đào tạo là hai thành viên nổi bật của thuật toán Học tập bán giám sát mà trên đó đã có rất nhiều nghiên cứu.

Đồng đào tạo là quá trình học nhãn cho các phiên bản chưa được gắn nhãn từ tập dữ liệu nhiều chế độ xem. Đồng đào tạo học hai bộ phân loại khác nhau cho hai chế độ xem đối tượng địa lý khác nhau, sau đó, các cá thể không được gắn nhãn sẽ được gắn nhãn khi hai bộ phân loại được xác nhận trên nhãn lớp của một cá thể. Một số trường hợp được gắn nhãn này được chọn và thêm vào tập huấn luyện để nâng cao mô hình học tập. Quá trình này được lặp lại cho đến khi đạt được tiêu chí chấm dứt hoặc đạt được độ chính xác về phân loại.

Trong khi đó, Active Learning là một quy trình sử dụng kiến thức giới hạn của con người về chú giải (tiên tri) để cải thiện các mô hình. Trong các thuật toán này, một mô hình đơn giản được huấn luyện đầu tiên bằng cách sử dụng một tập dữ liệu nhỏ được gắn nhãn, sau đó một số cá thể không được gắn nhãn thông tin được chọn lặp đi lặp lại và được gắn nhãn bởi một nhà tiên tri để cải thiện mô hình. Thách thức quan trọng nhất ở giai đoạn này là tìm ra các trường hợp sẽ thực hiện cải tiến tối ưu trên toàn bộ mô hình bằng cách biết nhãn của chúng.

Các thuật toán đề xuất trong luận văn này sử dụng các kỹ thuật Co-training để phát hiện các truy vấn tối ưu cho Active Learning. Tuy nhiên, hai thách thức nảy sinh ở đây. Vì Co-training sử dụng hai chế độ xem tính năng độc lập, nên trước tiên cần phải chia các tập hợp tính năng thành hai chế độ xem đối tượng địa lý khác nhau. Thách thức thứ hai là chọn các phiên bản để truy vấn từ oracle bằng cách sử dụng kết quả Đồng đào tạo.

Đồng đào tạo hoạt động dựa trên giả định có hai hoặc nhiều quan điểm độc lập và đủ. Điều này có nghĩa là mỗi khung nhìn của tập dữ liệu phải đủ để huấn luyện một mô hình và chúng phải độc lập với nhau. Các giả định này thường là không thể tưởng tượng được đối với các tập dữ liệu chế độ xem một lần điển hình. Do đó, phương pháp tốt nhất để đạt được những giả định này là chia chế độ xem đối tượng địa lý đơn lẻ thành hai nhóm chế độ xem đối tượng địa lý có lượng thông tin gần nhau nhất có thể. Lượng thông tin được tính bằng bốn phương pháp chủ yếu được sử dụng để lựa chọn đối tượng địa lý. Trước đây, thuật toán Học tập bán giám sát kết hợp Đồng đào tạo với Học tập tích cực (SSLCA) chỉ được áp dụng Mức tăng thông tin để đạt được

các chế độ xem tính năng khác nhau. Trong luận án này, các thuật toán Chi-Square, Analysis of Variance (ANOVA) và ReliefF được áp dụng cho mục đích này bên cạnh Độ lợi thông tin.

Để chọn các phiên bản từ tập dữ liệu không được gắn nhãn đến nhãn truy vấn, hai cách tiếp cận được sử dụng. Đầu tiên là tính toán sự đóng góp của các trường hợp dựa trên mỗi trường hợp và sự không chắc chắn của các trường hợp lân cận của nó. Thứ hai là sử dụng phương pháp học tập không có giám sát và phân cụm các trường hợp không chắc chắn để tìm các trường hợp ứng viên tốt nhất cho truy vấn. Hai cách tiếp cận này hình thành nên các phương pháp được đề xuất của chúng tôi, đó là Học tập tích cực nâng cao đồng đào tạo (CEAL) và Học tập cùng tích cực. Gần đây, Co-Active Learning cũng đã được đề xuất và áp dụng cho hai chế độ xem tính năng. Chúng tôi đã mở rộng thuật toán đó bằng cách áp dụng tính năng tách đối tượng để nó có thể được sử dụng trong các tập dữ liệu chế độ xem đối tượng địa lý.

Kết quả thử nghiệm được thực hiện trên tám bộ dữ liệu điểm chuẩn do Kho lưu trữ Máy học UCI (Đại học California, Irvine) cung cấp. Các bộ dữ liệu này là bộ dữ liệu rất phổ biến trong nghiên cứu học máy đã được hàng trăm nghiên cứu tham khảo và sử dụng. Đối với các mô hình đào tạo, năm thuật toán học máy từ các loại thuật toán khác nhau đã được sử dụng. Từ phương pháp thống kê Gaussian Naive Bayes được sử dụng, đó là một thuật toán tổng hợp. Từ các thuật toán dựa trên cây Cây quyết định và Rừng ngẫu nhiên đã được sử dụng. Từ Mạng thần kinh, Perceptron nhiều lớp và từ các phương pháp dựa trên vùng lân cận K-Nearest Neighbor đã được sử dụng. Gaussian Naive Bayes đã cung cấp cho chúng tôi những chắc chắn thống kê trực tiếp và là một thuật toán nhanh lặp đi lặp lại. Multilayer Perceptron cũng cung cấp sự chắc chắn ở một lớp trước lớp cuối cùng (softmax). Tuy nhiên, các thuật toán khác đã được thực hiện ở dạng hồi quy và bằng cách hiệu chỉnh kết quả hồi quy để nằm trong khoảng 0-1 chắc chắn đã được suy ra. Một số thử nghiệm được thực hiện để so sánh các phương pháp phân vùng, chiến lược lựa chọn truy vấn và phương pháp học máy. Độ chính xác của phân loại được so sánh với việc chúng tôi triển khai SSLCA, Học tập tích cực và Đồng đào tạo như các thuật toán cơ bản. Kết quả thử nghiệm cho thấy rằng trong hầu hết các trường hợp, CEAL vượt qua các phương pháp khác và phương pháp phân vùng tốt nhất là Mức tăng thông tin với một biên độ rất nhỏ.

Aktif Ögrenmeyi Güçlendirmek için E¸s-ö grenme Kullanılması

ÖZET

Son günlerde, yüksek miktarlarda verify eri¸silebilir hale gelmi¸stir, fakat bunların çok küçük bir kısmı sınıflandırılmı¸s ve etiketlenmi¸stir. Ayrıca, bu yüksek miktardaki veriden bilginin çıkarılması ise oldukça önemlidir. Etiketli ve etiketsiz Verinin her ikisinden de faydalanarak etiketleme proşedürünü geli¸stiren bir grup makine ögrenmesi algoritması Yarı-Denetimli Ö grenme olarak adlandırılmaktadır. Yarı-denetimli ögrenme yöntemlerinde, nadir bulunabilen etiketli örnekler üzerinden bir model egitilir ve daha sonra bu model çokça bulunabilen etiketsiz Verilerden faydalanarak geni¸sletilir ve iyile¸stirilir. Aktif Ögrenme ve E¸s-ö grenme yöntemleri üzerine ara¸stırmalar yapılan Denetimli Ö grenme algoritmalarından önde gelen iki yöntemdir.

Aktif Ögrenme ve E¸s-ö grenme, bilinen yarı denetimli ö grenme yöntemlerindendir ve literatürde bu konular üzerinde çok sayıda çalı¸sma yapılmı¸stır. Ayrıca cô ấy iki yöntemi birlikte kullanan farklı ara¸stırmalar da literatürde yer almaktadır. Bu çalı¸smada, bu iki yöntem, iki farklı mimaride birle¸stirelek, Aktif Ögrenme ve E¸s-ö grenmein en ba¸sarılı uygulamalarından biri olan SSLCA yöntemi ile kar¸sıla¸stırılmı¸stır. Bu çalı¸sma için yeniden geli¸stirilen SSLCA yöntemiyle pek çok farklı senaryoda kiểm tra edilmi¸s, ba¸sarıları degerlendirilmi¸stir. Aktif Ö grenmeyi gerçekle¸stirirken E¸s-ö grenmenin de dahil edilmesi ile, üç farklı ¸sekilde yöntemler kar¸sıla¸stırılmı¸stır: 1) mimariye dayalı 2) bölümleme yöntemine dayalı 3) algoritmaya dayalı.

Bu çalı¸smadaki temel fikir az sayıda etiketli verify ve çok sayıda etiketsiz Verinin bulundugu problemlerde, etiketsiz verify içerisinden ba¸sarımı arttırabilecek örneklerin seçimidir. Yarı-denetime tabi olan bu tür etiketleme sorununa E¸s-ögrenme ve aktif ögrenmenin birle¸stirilmesiyle çözüm üretilmeye çalı¸sılmı¸stır. Klasik E¸s-ö grenme yakla¸sımlarının aksine, tekli görünümlü verify kümeleri üzerinde çalı¸sılmı¸stır. Önerilen yakla¸sım Beklenen Hata Azaltma veya Varyans Azaltma gibi birçok aktif ögrenme yakla¸sımlarının aksine, yogun hesaplama ve i¸slem gücü gerektirmemektedir.

Belirtilen problemlerdeki ba¸samın incelenmesi için, bu çalı¸smada önerilen E¸s-ögrenmeyle Zenginle¸stirilmi¸s Aktif Ö grenme (Đồng đào tạo Học tập Tích cực Nâng cao - CEAL), daha önceki çalı¸smalarda ba¸sarılı oldugu gösterilen CoActive ve SSLCA olmrak üzere üç farklullan yöntmı kullanıölm. SSLCA yakın zamanda önerilmi¸s yarı denetimli ögrenme problemleri için ba¸sarılı bir algoritma oldu gu için bu çalı¸sma referans algoritma olarak kullanılmı¸stır. CoActive yöntemi çalı¸smada hedeflenen tekli görünümlü verify kümeleri üzerinde uygulanamadıgı için, bu yöntem de gi¸stirilerek vấn đề uygulanmı¸stır.

E¸s-ögrenme do gi¸cal ¸sekilde kullanılmaktadır. Çalı¸smada kullanılan verify kümesi çalı¸sma yöntemine göre dolayı bunların iki görünüme ayırılması gerekmektedir. Bu nedenle öncelikle, tekli görünümlü verify kümesi (etiketsiz ve etiketli kümeler), ögrenmeye katkılarına dayanarak iki ba gımsız öznitelik kümesine ayrılmı¸stır. Ögrenme kümesindeki her bir özelli gin sahip oldu gu bilgi, Bilgi Kazancı (Tăng thông tin), Chi-square, ANOVA ve ReliefF olmak üzere dört farklı metrik ile ölçülmü¸stür. Bilgi kazançları bilgi teorisine dayalı bir yöntemdir. Esas olarak, karar agaçlarının (cây quyết đ ịnh) özelliklerini inceleyerek agaç olu¸sumu sırasında dallarda bulunacak öznitelik seçimi için kullanılır. Bir özelligin ta¸sıdı gı bilgi miktarı, Sistemin entropisinde neden olaca gı de gi¸sim ile ilgilidir. Diger yöntemler, Chi-Square (χ) ve ANOVA, öznitelik seçimi için kullanılan yöntemlerdir ve her bir öznitelik ile etiketin arasındaki ili¸skiye dayanarak çalı¸smaktadır. Chi-Square, öznitelik ve etiketler arasındaki bagımlılı gı hesaplar, ANOVA ise öznitelikler ve etiketler arasındaki kovaryansı hesaplar. Veri kümelerini bölmek için kullandıgımız son yöntem ise di ger bir öznitelik seçimi yöntemi olan Cứu trợ algoritmasının güncellenmi¸s bir versiyonu olan Cứu trợF'tir. Bu algoritma yinelemeli bir algoritma olup, özniteliklere agırlık vermekte ve cô ấy adımda rastgele seçilen bir örnege en yakın aynı etiketli ve en yak ven farklı etiketli örnekleri kullanaküted gırlıklekleri kullanaküted gırlıkleküted gırlıklek. Bu metriklerle yapılan ölçümler sonucunda, ögrenmeye etkilerine göre öznitelikler her iki kümeye de e¸sdo gru ögrenmeye çalı¸sılmı¸stır. Bir diger deyi¸sle, cô ấy iki kümenin de e¸sit bir ¸sekilde bilgiye ve güce sahip olması amaçlanmı¸stır. Daha sonra, Gaussian Naive Bayes, K-Nearest Neighbor, Karar Agacı, Rừng Ngẫu nhiên ve Çok Katmanlı Algılayıcı (Sinir A gı) olmak üzere, bu etiketli görünümler üzerinde çe¸sitli ögrenme algoritmaları e gitimi ve tahmini için kullanılarak, etiketlenmemi¸s örneklerin her bir sınıf için o sınıfına ait olma ihtimali bulunmu¸stur.

Bu noktadan itibaren üç yöntem birbirlerinden ayrılmaktadır. SSLCA tekrarlayan bir algoritma olup; öncelikli olarak etiketli verify kümesi üzerinde sınıflandırıcı egiterek etiketsiz Veri üzerindeki en belirsiz ve en çok emin olunan örnekleri bulmaya çalı¸sır. En belirsiz örnegi aktif ö grenme prosedürüne gönderir ve en emin olunan örne gi ise birlikte egitim prosedürüne gönderir. Bu iki algoritmadan gelen yeni Verileri etiketli Veri kümesine katarak sınıflandırıcıyı günceller. Diger bir yöntem ise CoActive olup; etiketsiz verify kümesi içerisinden örnekler sınıf sayısı kadar demet olu¸sturularak bulunur. Demetler iki farklı sınıflandırıcının kararlarının uyu¸smadıgı örnekler üzerinde olu¸sturulur. Her bir demetin medyanına en yakın üye seçilerek etiketlenir ve egitim kümesine katılır. Bu çalı¸smada önerilen CEAL algoritması ise, etiketsiz Veri kümesi içerisinde en çok bilgi içeren örnekleri bulmak için bir katkı degeri kullanır ve en yüksek katkı degerine sahip etiketsiz Veriyi aktif ö grenme için etiketleyiciye gönderir.

Kiểm tra sonuçları farklı makine ögrenme algoritmaları kullanılarak domainse edilmi¸stir. Çalı¸smaya referans olan algoritma, olasılıkları dogrudan veren, üretken istatistiksel bir mô hình olan Gaussian Naive Bayes'tir. Gaussian Naive Bayes, makine ögrenmesi algoritmaları arasında güçlü bir algoritma olarak anılmamaktadır. Bu yüzden Karar Agacı, Rừng Ngẫu nhiên ve Çok Katmanlı Perceptron gibi daha güçlü algoritmalar da test edilmi¸stir. Karar Agacı ve Random Forest algoritmalarının çıktırlarında bir sınıfa dahil olma olasılık degerlerini domainse edebilmek için bu algoritmalar regresyon yapılarak sonuçlar olasılık degerine çrrilmi. Sinir aglarında ise, her

olma olasılıklarını domainse edebilmek için benzer bir kalibrasyon kullanılmı¸stır.

Yöntemler üzerindeki ilk kar¸sıla¸stırma, mimari farkı üzerinde yapılmı¸stır. Çalı¸smada referans alınan mimari, Aktif Ögrenme ve E¸s-ö grenme i¸slevlerini birbirine paralel olarak uygulayan SSLCA'dır ve her yinelemede her ikisini birden kullanmaktadır. Önerilen mimariler ise CEAL ve modifiye edilen CoActive adlı yönteme ait mimarilerdir. CEAL'de sıralı Aktif Ögrenme ve E¸s-ö grenme kullanılmı¸stır. Fakat SSLCA'da oldugu gibi e¸s-ö grenme kümeleri ayrı ayrı bulunmaktadır. CoActive yöntemi ise

vi iyi sorguyu bulmak için benzer yöntemler kullanılmı¸s, ancak kümeleme yapılmı¸stır.

Diger kar¸sıla¸stırma, bölümleme algoritmalarına dayanarak yapılmı¸stır. E¸s-ö grenme, verify kümelerinin özünde iki bagımsız ve kendine yeterli görünüme ayrıldı gı varsayımına dayanılarak yürütülmektedir. Dolayısıyla özellikleri iki alt kümeye bölmek amacıyla, en bilgilendirici örnekleri domainse etmek ve özniteliklerin alt bölümlere oldukça adil dagılmasını sa glamak için Bilgi Kazancı, Chi-Square, ANOVA ve ReliefF özellik seçme yöntemleri kullanılmı¸stır ve kar¸sıla¸stırılmı¸stır.

Deney sonuçları göstermektedir ki, çogu test ko¸sulunda CEAL di ger yöntemlerden üstün gelmektedir ve en iyi bölümleme yöntemi ise az bir fark ile Bilgi Kazancı yöntemidir.

1. GIỚI THIỆU

Trong nhiều vấn đề về học máy, tồn tại một lượng lớn dữ liệu không được gắn nhãn để khai thác ngoài dữ liệu có nhãn sẵn có. Trong hầu hết các trường hợp trong các vấn đề thế giới thực, thu thập dữ liệu thô rẻ hơn và dễ dàng hơn nhiều so với việc lấy nhãn của chúng. Nhãn là thường được chỉ định thông qua các nỗ lực từ một nhân viên con người và nhân viên này thường phải là một chuyên gia. Vấn đề này có thể được nhìn thấy trong nhiều lĩnh vực như xử lý hình ảnh, nhận dạng giọng nói và tin sinh học. Ví dụ, [1] tuyên bố rằng việc xác định các vùng bị methyl hóa trong DNA cần được điều tra lâu bởi một chuyên gia được đào tạo chuyên sâu. Việc tìm kiếm một chuyên gia về con người là rất khó khăn và tốn kém.

Tuy nhiên, việc phát triển các mô hình học máy thích hợp đòi hỏi một lượng lớn dữ liệu được gắn nhãn. Do đó, ý tưởng đào tạo một mô hình phức tạp với ít nhất các trường hợp được gắn nhãn là hấp dẫn. Nhiều thuật toán học tập bán giám sát đã được đề xuất có thể sử dụng dữ liệu không được gắn nhãn; hai trong số những ý tưởng này là Học tập tích cực [2] và Đồng đào tạo [3]. Trong Học tập tích cực, chúng tôi cố gắng tìm số lượng tối thiểu các trường hợp đại diện cho toàn bộ dữ liệu theo cách tốt nhất và loại bỏ việc dán nhãn các trường hợp dư thừa; sau đó, chỉ cần yêu cầu nhà tiên tri (tức là một chuyên gia về tác nhân con người) chỉ gắn nhãn các trường hợp này.

Cách tiếp cận khác là Đồng đào tạo; trong đó chúng tôi khai thác một số lượng lớn các các trường hợp. Đồng đào tạo là một thuật toán lặp đi lặp lại và học hai bộ phân loại khác nhau trên hai chế độ xem đối tượng địa lý khác nhau. Về cơ bản, nó phân loại các mẫu dữ liệu trên một tập hợp không được gắn nhãn và thêm các ví dụ chắc chắn nhất vào tập huấn luyện. Có rất nhiều phần mở rộng của hai thuật toán này trong tài liệu.

Luận án này đề xuất một thuật toán Active Learning sử dụng Co-training để chọn phiên bản không gắn nhãn hứa hẹn nhất để lấy nhãn của nó từ oracle. Đề xuất thuật toán đã được thử nghiệm trên tám tập dữ liệu sẽ được điều tra kỹ lưỡng trong Phần Kết quả Thực nghiệm. Nó cũng được so sánh với các đối tác của nó, Đồng đào tạo,

Học tập tích cực, học tập cùng hoạt động và bán giám sát kết hợp Đồng đào tạo
và các thuật toán Active Learning (SSLCA). Phần tiếp theo thảo luận về các chi tiết kỹ thuật
về thuật toán Co-training và Active Learning sẽ được thảo luận.

1.1 Tổng quan tài liệu về Đồng đào tạo và Học tập tích cực

Phần này tóm tắt các công trình trước đây về Đồng đào tạo và Học tập tích cực
các thuật toán và sự kết hợp trước đây của các phương pháp này đã được thảo luận trong tài liệu.

1.1.1 Thuật toán đồng đào tạo

Đồng đào tạo là một cách tiếp cận được Blum và Mitchell đề xuất vào năm 1998 [3]. Blum et
al. đã thảo luận về cách thu thập dữ liệu có thể được thực hiện tự động, nhưng việc gắn nhãn nó yêu cầu
nỗ lực tốn kém của con người và thu thập một lượng lớn dữ liệu không được gắn nhãn ít hơn nhiều
đắt hơn việc dán nhãn cho chúng. Trong hầu hết các trường hợp, một lượng đáng kể nỗ lực của chuyên gia
là bắt buộc, dẫn đến việc chúng ta có một tập hợp lớn dữ liệu không được gắn nhãn và một tập hợp nhỏ
dữ liệu được gắn nhãn. Ý tưởng cơ bản của Đồng đào tạo là thúc đẩy một cách để khai thác
dữ liệu để nâng cao các mô hình được đào tạo với dữ liệu được gắn nhãn.

Việc triển khai thuật toán Đồng đào tạo ban đầu được thực hiện trên trang web
vấn đề phân loại. Trong khi khám phá các trang web được nắm bắt, hai tập hợp thông tin
đã có sẵn; tập hợp đầu tiên là các tính năng được trích xuất từ mỗi từ của trang web và
nhóm thứ hai là các tính năng được trích xuất từ các từ trong các trang web có siêu liên kết đến
trang mạng. Blum và cộng sự. có một số lượng nhỏ các trang web được gắn nhãn L ngược lại với
một số lượng lớn các trang web không được gắn nhãn, U, đã được lấy tự động. Họ
các tính năng đã sử dụng (Bag of Words) thu được từ hai tập hợp các trường hợp được gắn nhãn L1 và
L2 để huấn luyện hai mô hình Naive Bayes khác biệt h1 và h2. Sau đó, họ sử dụng các mô hình này
để dự đoán các trường hợp không được gắn nhãn U1 và U2. Do đó, hai dự đoán t1 và t2 và
xác suất của những dự đoán này được giữ lại cho mỗi trường hợp không được gắn nhãn.

Các trường hợp có xung đột trong dự đoán của hai mô hình đã được chọn làm
các trường hợp xung đột. Các trường hợp xung đột là những trường hợp có sự khác biệt cao nhất
giữa các xác suất của các dự đoán; các mô hình có thể được khởi động bằng cách lấy

dự đoán từ trình phân loại đáng tin cậy hơn, đặt nhãn của đối tượng địa lý khác
tập hợp cá thể đó và thêm cá thể này vào các cá thể được gắn nhãn. Hai bộ phân loại
được giữ lại trong lần lặp tiếp theo bằng cách sử dụng các phiên bản mới được thêm vào và Đồng đào tạo có
cùng ý tưởng về Tối đa hóa kỳ vọng (EM), ngoại trừ hai bộ dữ liệu khác nhau [4].

Naive Bayes được sử dụng làm bộ phân loại trong thuật toán Đồng đào tạo ban đầu; nó là một
mô hình thống kê vốn đã trích xuất xác suất của mỗi trường hợp thuộc về
bất kỳ lớp nào và dự đoán cuối cùng thu được bằng cách sử dụng xác suất này [5]. Naive Bayes là một
thuật toán gia tăng giúp Đồng đào tạo chạy nhanh hơn và các thuật toán gia tăng là
các thuật toán trong đó việc thêm một phiên bản không có nghĩa là chúng ta phải đào tạo lại toàn bộ
mô hình, chúng ta chỉ cần xem xét hiệu ứng của các trường hợp mới [6]. Blum và cộng sự. minh họa
rằng có hai bộ tính năng là cần thiết để chạy Đồng đào tạo và chúng phải
thỏa mãn hai điều kiện sau:

1) Mỗi bộ tính năng cần phải đủ, có nghĩa là mỗi bộ tính năng là đủ
để đào tạo một mô hình đại diện phù hợp một mình và

2) Mỗi tập dữ liệu phải độc lập với tập dữ liệu khác.

Lưu ý rằng các điều kiện tiên quyết này được thỏa mãn trong bài toán phân loại trang web. Ở trong
vấn đề này, có hai bộ dữ liệu trong đó mỗi bộ đào tạo đủ một mô hình và chúng
khá độc lập với nhau. Có các công trình khác áp dụng Co-training
sử dụng các thuật toán khác nhau tương tự như Naive Bayes, chẳng hạn như Máy vectơ hỗ trợ
(SVM) và cây quyết định [7]. Họ đã chỉ ra rằng thay vì các mô hình chung chung, một
cũng có thể sử dụng các mô hình phân biệt mạnh mẽ hơn và có thể thay thế
các mô hình trong nghiên cứu học máy gần đây [7].

Kể từ đó, đã có nhiều công việc trong lĩnh vực này. Pierce và cộng sự. [8] đã thảo luận
Đồng đào tạo sử dụng trong xử lý ngôn ngữ tự nhiên và cho thấy rằng Đồng đào tạo là một
mô hình phân biệt đối xử nhiều hơn so với mô hình tổng hợp. Mô hình sinh là mô hình
có thể mô tả mô hình tạo ra tập dữ liệu vấn đề. Ví dụ
các mô hình sinh sản là Naive Bayes và EM [9]. Trong các mô hình này, một mô hình thống kê
(kết hợp các bản phân phối) được tạo ra có thể mô tả dữ liệu ban đầu

bằng cách tốt nhất. Ngược lại, các mô hình phân biệt đối xử không liên quan gì đến việc
dữ liệu đến từ đâu. Họ cố gắng tìm các nhãn mà không có bất kỳ ý tưởng nào về cách dữ liệu có
đã được tạo. Các mô hình phân biệt ví dụ là SVM, hồi quy tuyến tính và Neural
Các mạng [9]. Mặc dù các thuật toán Đồng đào tạo và EM khai thác dữ liệu không được gắn nhãn,
Đồng đào tạo tốt hơn EM trong hầu hết các trường hợp vì EM cố gắng phù hợp với mô hình
đối với giả định của nó về dữ liệu. Trong trường hợp những giả định này hơi sai,
nó cố gắng học các mô hình sai và trong mỗi lần lặp lại sai lệch so với thực tế. Mặc dù
vấn đề này tồn tại trong Đồng đào tạo, nó ít nghiêm trọng hơn vì thay vì sử dụng tập hợp này
của các giả định như EM, Co-training sử dụng đầu ra của bộ phân loại để đào tạo lại mô hình [4].

Một vấn đề khác trong việc sử dụng thuật toán Đồng đào tạo là điều kiện có hai
các bộ tính năng độc lập cho từng vấn đề. Đây là một điều kiện tiên quyết khá khó khăn.
Trước đây, để giải quyết vấn đề này, hai bộ phân loại khác nhau đã được sử dụng trên một
tập dữ liệu [10]. Do đó, các giảng viên yếu kém khác nhau được sử dụng để đào tạo và dự đoán các trường hợp,
và trong trường hợp này, chúng tôi không giới hạn số lượng tập dữ liệu phân biệt độc lập.
Do đó, bằng cách hoạt động tương tự như AdaBoost hoặc bất kỳ thuật toán khởi động nào khác,
số lượng dự đoán có thể được tăng lên và có thể đưa ra quyết định cuối cùng bằng cách
quần thể của họ.

1.1.2 Học tập tích cực

Học chủ động hoặc học truy vấn là một thuật toán học máy bán giám sát
đã được sử dụng trong nhiều ứng dụng khác nhau. Trong Học tập tích cực, ngoài
một tập hợp các bản sao được gắn nhãn và không được gắn nhãn, chúng tôi có quyền truy cập vào một số loại chú thích
có thể cung cấp cho chúng tôi nhãn của từng trường hợp. Các nhãn này có thể được sử dụng để mở rộng tập hợp
các trường hợp được gắn nhãn và do đó, trao quyền cho mô hình được đào tạo từ tập hợp này. Điều này có nghĩa là
mà chúng ta có thể truy vấn các nhãn của các cá thể không được gắn nhãn từ một tiên tri. Bên thứ ba này
nguồn có thể là người vận hành hoặc bất kỳ hình thức nào khác của hệ thống ghi nhãn. Các thách thức
là việc lấy nhãn từ một nhà tiên tri rất tốn kém ở nhiều định dạng (ví dụ: nó có thể
tốn rất nhiều thời gian, hoặc chúng tôi có thể chỉ có quyền truy cập vào một số truy vấn hạn chế).

Thách thức chính của Học tập chủ động là tìm ra truy vấn tối ưu. Tối ưu
truy vấn là quá trình tìm kiếm một cá thể trong đó việc có nhãn của nó sẽ làm giảm
Lỗi khái quát hóa nhiều nhất. Nói cách khác, chúng tôi cố gắng tìm kiếm thông tin
các trường hợp để truy vấn có lợi cho việc đạt được nhiều cải tiến nhất trong mô hình của chúng tôi với
số lượng truy vấn ít nhất. Cải tiến này có thể giảm thiểu lỗi chung,
tối đa hóa điểm F chung hoặc bất kỳ chỉ số chất lượng nào khác của mô hình
quan trọng với chúng tôi.

Bằng cách sử dụng khái niệm thứ nguyên Vapnik-Chervonenkis (thứ nguyên VC), Cohn et
al. [11] cho thấy rằng để đạt được độ chính xác của ε trong Học tập tích cực, chúng ta chỉ cần
O (log (N)) cá thể để có được nhãn trong đó N là số cá thể. Các
trong trường hợp xấu nhất, Học chủ động sẽ tốt như học thụ động (a
phương pháp học máy thông thường), nhưng mục tiêu của Active Learning là tìm
các phương pháp tiếp cận theo kinh nghiệm tốt hơn nhiều so với trường hợp xấu nhất.

Từ góc độ lý thuyết, Active Learning có thể được chia thành hai trường hợp. Hoặc
chúng tôi giả định rằng chúng tôi có một hàm đích đại diện cho dữ liệu được gắn nhãn và không được gắn nhãn
hoàn hảo (trường hợp đã nhận ra), hoặc chúng tôi không có giả định như vậy và chúng tôi cố gắng cải thiện
bộ phân loại của chúng tôi trong mỗi bước để có được hiệu suất tốt hơn (trường hợp bất khả tri) [12]. Từ
về góc độ thực tế, chúng tôi có các loại kịch bản Học tập tích cực khác nhau. Các
hai trường hợp phổ biến nhất là 1) dựa trên luồng, trong đó các phiên bản không được gắn nhãn đến
trong một luồng và cần phải quyết định trực tuyến xem có nên truy vấn hay không và 2) dựa trên nhóm,
nơi chúng tôi có quyền truy cập vào tất cả các phiên bản chưa được gắn nhãn bất kỳ lúc nào; do đó, chúng tôi có thể quyết định
các trường hợp tốt nhất để truy vấn, có thể kiểm tra tất cả các trường hợp. Trong nghiên cứu này, dựa trên nhóm
các tình huống được điều tra vì chúng là dạng vấn đề phổ biến nhất
xảy ra trong thế giới thực [2].

1.1.2.1 Các chiến lược lựa chọn truy vấn

Các cách tiếp cận khác nhau để chọn phiên bản không được gắn nhãn tốt nhất để truy vấn được thảo luận
trong văn học. Bảy cách tiếp cận quan trọng nhất được xem xét dưới đây [2, 12, 13].

Trong các cách tiếp cận sau, trường hợp đã chọn sẽ được hiển thị dưới dạng x ·

Dựa trên sự không chắc chắn

Các phương pháp truy vấn cơ bản và phổ biến nhất trong Active Learning là
các phương pháp dựa trên độ không đảm bảo. Cách triển khai dễ nhất của một phương pháp như vậy là
thực hiện nó bằng cách sử dụng các mô hình di truyền thống kê, bởi vì chúng mang lại khả năng
của mỗi lớp vốn có cho mọi trường hợp dưới dạng xác suất. Khác
các thuật toán cũng được sử dụng rộng rãi cho mục đích này bằng cách thực hiện chuẩn hóa và
hiệu chuẩn kết quả hồi quy của thuật toán [2].

Trong các phương pháp độ không đảm bảo, các trường hợp không được gắn nhãn không chắc chắn nhất được phát hiện và
chất vấn. Trong các bài toán phân loại nhị phân, độ không chắc chắn của một mô hình cao khi
xác suất sau của hai lớp quá gần nhau. Điều này có thể được đại diện
sử dụng công thức sau:

$$x = \text{argmin}x \mid P(y = + \mid x) \quad P(y = - \mid x) \mid \tag{1.1}$$

Trong phương trình này, P (y = + | x) là xác suất của trường hợp x thuộc lớp +
và P (y = - | x) là xác suất để nó thuộc lớp -.

Nhưng trong trường hợp chúng ta có nhiều hơn hai lớp thì công thức trước đó phải là
đã cập nhật. Trong trường hợp này, lựa chọn sẽ là trường hợp có tối thiểu là tối đa
xác suất sau. Nó có nghĩa là trường hợp này có xác suất thấp nhất cho dự đoán
nhãn mác. Phiên bản thông tin này được chọn bởi các phương trình sau.

$$x = \text{argmin}xP(y\hat{} \mid x) \tag{1.2}$$

$$y\hat{} = \text{argmax}yP(y \mid x) \tag{1.3}$$

trong đó y là lớp có xác suất lớn nhất. Do đó, x sẽ là
thành viên có xác suất hậu kỳ tối thiểu tối đa. Trong trường hợp này, chúng tôi thua
thông tin xác suất của các lớp không phải là lớp sau cực đại. Để vượt qua
vấn đề này lấy mẫu lề đã được đề xuất [2] với công thức được đưa ra trong

Phương trình 1.4.

$$x = \operatorname{argmin}_x P(\hat{y}1 \mid x) \quad P(\hat{y}2 \mid x) \tag{1.4}$$

trong đó ˆy1 và ˆy2 là lớp có khả năng xảy ra cao nhất và lớp có khả năng xảy ra cao thứ hai tương ứng. Trong cách tiếp cận này, chúng tôi mong muốn tìm ra trường hợp có sự khác biệt lớn nhất giữa lớp sau tối đa có thể xảy ra với lớp sau tối đa thứ hai lớp có thể xảy ra. Nói cách khác, đây là trường hợp mà mô hình của chúng tôi nghi ngờ nhất về việc lựa chọn giữa các lớp. Mặc dù nó tốt hơn so với phương pháp đầu tiên, nó vẫn không xét xác suất của lớp thứ ba trở xuống. Do đó, dựa trên entropy cách tiếp cận (xem phương trình.1.5) đã được đề xuất bởi cùng các nhà nghiên cứu.

$$x = \operatorname{argmax}_x \quad P(yi \mid x) \log P(yi \mid x) \tag{1.5}$$

P (yi | x) logP (yi | x) là entropy của mỗi lớp. Trong phương pháp này, các thành viên có entropy cao nhất (phương sai) đã được chọn dựa trên các mô hình hiện tại và truy vấn các trường hợp này, chúng tôi đang giảm entropy giữa các trường hợp. [2].

Giảm lỗi mong đợi

Một cách tiếp cận rất hiệu quả khác là tìm phiên bản nếu nhãn của nó đang được gán, sẽ tạo ra thay đổi lớn nhất trong Lỗi tổng quát hóa. Điều này được thực hiện bởi chỉ định từng nhãn một cho các phiên bản chưa được gắn nhãn và đào tạo toàn bộ mô hình. Rõ ràng đây là một phương pháp rất hiệu quả, nhưng quá tiêu tốn tài nguyên, điều này phương pháp này có thể được sử dụng khi các truy vấn đắt hơn nhiều so với sức mạnh. Trong trường hợp này, chúng tôi xem xét tất cả các nhãn cho tất cả các phiên bản không được gắn nhãn, hãy tìm sự khác biệt mà các nhãn tạo ra và chọn phiên bản có nhãn sẽ tạo ra sửa đổi lớn nhất. [12].

Giảm phương sai

Lỗi Tổng quát hóa (E) của một mô hình học máy là sự kết hợp của phương sai và độ chệch của mô hình đó là E = Bias2 + Phương sai [5]. Khi biết thực tế này, có thể để giảm Lỗi tổng quát hóa bằng cách giảm phương sai. Trước đây trong [11] đã hiển thị rằng việc tìm kiếm các thành viên để giảm bớt phương sai ít tiêu tốn hơn nhiều so với việc tìm kiếm thành viên làm giảm sai số bởi vì khi giảm phương sai, chúng ta chỉ cần tính

phương sai của tập dữ liệu cho từng trường hợp và nhãn thay vì đào tạo toàn bộ
mô hình như cách tiếp cận giảm thiểu lỗi [12].

Thay đổi mô hình dự kiến

Trong các mô hình đang hoạt động dựa trên Gradient Descent hoặc tối ưu hóa quen thuộc
các thuật toán như Neural Networks và Gradient Bo boost, chúng ta có thể dựa vào số lượng
thay đổi mỗi phiên bản sẽ thực hiện trên các gradient nếu nhãn của nó được đưa ra. Nó đang làm việc
như Giảm lỗi mong đợi trừ khi không cần đào tạo toàn bộ mô hình từ
cào. Ví dụ: trong Mạng thần kinh, thay vào đó chúng ta chỉ cần thực hiện một bước chuyển tiếp
của một số đường chuyền về phía trước và phía sau. Vì vậy, nếu chúng tôi chọn các phiên bản có kỳ vọng cao nhất
thay đổi mô hình, chúng tôi sẽ đạt được chất lượng đáng kể với rất ít tính toán [2].

Dựa trên mật độ

Trong các phương pháp dựa trên độ không đảm bảo, thực tế là mỗi trường hợp có thể là một ngoại lệ không bao giờ
đang được xem xét. Nó có nghĩa là tất cả các trường hợp được đánh giá như nhau. Nhưng về mật độ
các phương pháp tiếp cận dựa trên, các khu vực dày đặc nhất được nhắm mục tiêu và truy vấn được chọn từ đây
vùng đất. Điều này có thể được thực hiện bằng cách tính toán Ước tính mật độ nhân (KDE) [5] hoặc
bất kỳ phương pháp tính toán mật độ tương tự. Sau đó, phiên bản không được gắn nhãn gần nhất
vùng mật độ cao nhất được chọn [2].

Dựa trên cụm

Cách tiếp cận rất giống với phương pháp dựa trên mật độ đã được đề xuất bởi Hsu et al. [14]. Ở trong
cách tiếp cận này tập hợp cá thể không gắn nhãn được nhóm thành n cụm với n là số
nhãn mác. Sau đó, các thành viên gần trung tâm của các cụm được chọn. Cái này
phương pháp tiếp cận phụ thuộc nhiều vào chất lượng phân nhóm và phân bố không gian
trong tổng số các trường hợp.

Truy vấn bởi Ủy ban hoặc Bất đồng

Tồn tại hai cách khác để tìm ra truy vấn tối ưu trong tài liệu: đầu tiên là đào tạo
một số mô hình và cố gắng tìm ra trường hợp mà tất cả các mô hình đều không chắc chắn về nó
(Truy vấn theo Ủy ban), và thứ hai là chọn các trường hợp mà mô hình có
bất đồng về chúng (Truy vấn bằng cách Bất đồng). Sau khi lựa chọn quan trọng này
ví dụ, nhãn của nó được truy vấn bởi oracle. Cách tiếp cận này đã được đề xuất bởi Seug et
al. [15] và đã được điều chỉnh cho các vấn đề khác nhau. Mô hình đề xuất của chúng tôi Đồng đào tạo
Học tập chủ động nâng cao (CEAL) sẽ dựa trên phương pháp này.

1.1.3 Kết hợp học tập tích cực và đồng đào tạo

Như đã nêu trong văn bản này, đã có rất nhiều công việc được tiến hành về Hoạt động
Học tập và Đồng đào tạo riêng biệt. Hai phương pháp này cũng đã được sử dụng cùng nhau
trong một số nghiên cứu, chẳng hạn như [16] [17] [18] [19]. Các tài liệu tham khảo đặc biệt cho luận án này
sẽ là hai nghiên cứu gần đây [20] [21]. Phương pháp đầu tiên đề xuất một phương pháp gọi là SSLCA để
tìm các vùng có mật độ cao và các trường hợp truy vấn được chọn từ các vùng này. Zhang
et. al. [20] đã sử dụng song song Co-training và Active Learning. Phương pháp của họ
liên quan đến việc lấy tập dữ liệu ở một chế độ xem và tách nó thành hai chế độ xem khác nhau, và
mỗi bước thực hiện cả Đồng đào tạo và Học tập tích cực. Trong Đồng đào tạo, họ thêm
các trường hợp có độ chắc chắn cao trong một chế độ xem và độ chắc chắn thấp nhất trong một chế độ xem khác
vào tập hợp có nhãn với nhãn thu được từ một chế độ xem nhất định. Đồng thời,
Bước Active Learning cũng diễn ra, cố gắng tìm ra phiên bản có mức thấp nhất
độ chắc chắn tổng thể và nhãn của nó bằng cách truy vấn tiên tri và thêm nó vào tập dữ liệu được gắn nhãn.

Yuce et. al. [21] đã đề xuất một phương pháp dựa trên sự bất đồng có tên là Co-Active
Học tập. Co-Active Learning cũng hoạt động với các tập dữ liệu được phân tách tự nhiên, như
Đồng đào tạo. Nó hoạt động dựa trên phân nhóm các trường hợp không chắc chắn và sau đó gửi
trung tâm cụm đến một truy vấn từ tiên tri. Giả định rằng dữ liệu sẽ
trải rộng trong n cụm không gian, với n là số lớp. Họ cho rằng
cá thể nhiều thông tin nhất tồn tại ở giữa cụm và trường hợp này sẽ
đại diện cho số lượng cá thể lớn nhất; vì vậy, bằng cách tìm nhãn của phiên bản này,
số lượng lớn hơn các phiên bản sẽ bị ảnh hưởng. Trung tâm cụm này là thành viên gần nhất

ở giữa các cụm. Hạn chế của phương pháp này là nó chỉ hoạt động với
các bộ dữ liệu được phân tách tự nhiên và phụ thuộc nhiều vào chất lượng phân nhóm và
phân phối dữ liệu.

2. ĐẶT VẤN ĐỀ VÀ PHƯƠNG PHÁP NGHIÊN CỨU.

2.1 Định nghĩa vấn đề

Trong hầu hết các trường hợp trong các vấn đề trong thế giới thực, việc thu thập dữ liệu thô rẻ hơn và dễ dàng hơn rất nhiều hơn là lấy nhãn của họ. Trong nhiều trường hợp, nhãn là kết quả của tác nhân của con người công việc. Tác nhân con người này cần phải là một chuyên gia (ví dụ: nhiễm sắc thể sai sót [1] có thể được phát hiện sau một cuộc điều tra dài bởi một chuyên gia được đào tạo chuyên sâu, hoặc động đất-phát hiện thiệt hại-dễ bị tổn thương-tòa nhà là kết quả của một cuộc điều tra dài bởi các kỹ sư). Việc tìm kiếm một chuyên gia về con người rất khó và tốn kém. Tuy nhiên, dữ liệu thô đang trở nên rẻ hơn và phong phú hơn mỗi ngày. Terabyte của con người dữ liệu bộ gen và thông tin địa lý được thu thập bởi các vệ tinh.

Để phát triển một thuật toán học máy thích hợp, một lượng lớn dữ liệu được gắn nhãn là cần thiết. Ý tưởng đào tạo một mô hình phức tạp với ít trường hợp được gắn nhãn nhất hấp dẫn, và các phương pháp khác nhau đã được đề xuất để thực hiện mục tiêu này. Hai trong số những ý tưởng này là Học tập tích cực [2] và Đồng đào tạo [3], cả hai đều thuộc về thuật toán học máy bán giám sát. Trong Học tập tích cực, chúng tôi cố gắng tìm số lượng tối thiểu các trường hợp đại diện cho toàn bộ dữ liệu theo cách tốt nhất và tránh ghi nhãn các trường hợp thừa; sau đó, chúng tôi yêu cầu nhà tiên tri (tức là một đặc vụ chuyên gia về con người) chỉ gắn nhãn các trường hợp tối thiểu này.

Trong Đồng đào tạo, một số lượng lớn các trường hợp không được gắn nhãn được khai thác. Điều này chủ yếu là có thể trong tập dữ liệu có hai hoặc nhiều chế độ xem vốn có, trong đó tất cả các chế độ xem này tự túc và độc lập. Điều này có nghĩa là cần phải có nhiều hơn một loại dữ liệu cho từng trường hợp để có thể phát hiện ra nhãn thực của những các trường hợp chỉ với một trong các loại này (chế độ xem).

Mục đích chính của nghiên cứu này là thiết kế một phương pháp có thể sử dụng sức mạnh của cả Học tập tích cực và Đồng đào tạo khi chỉ có một chế độ xem. Điều này dẫn chúng tôi tạo ra một mô hình có thể tạo ra các mô hình chất lượng tối ưu với số lượng tối thiểu của nhãn.

2.2 Phương pháp

Trong luận án này, hai phương pháp đã được đề xuất để giải quyết vấn đề này, CEAL và một phiên bản được phân vùng của Co-Active [21] và kết quả sẽ được so sánh với thực hiện SSLCA [20].

2.2.1 Đồng đào tạo Học tập tích cực nâng cao (CEAL)

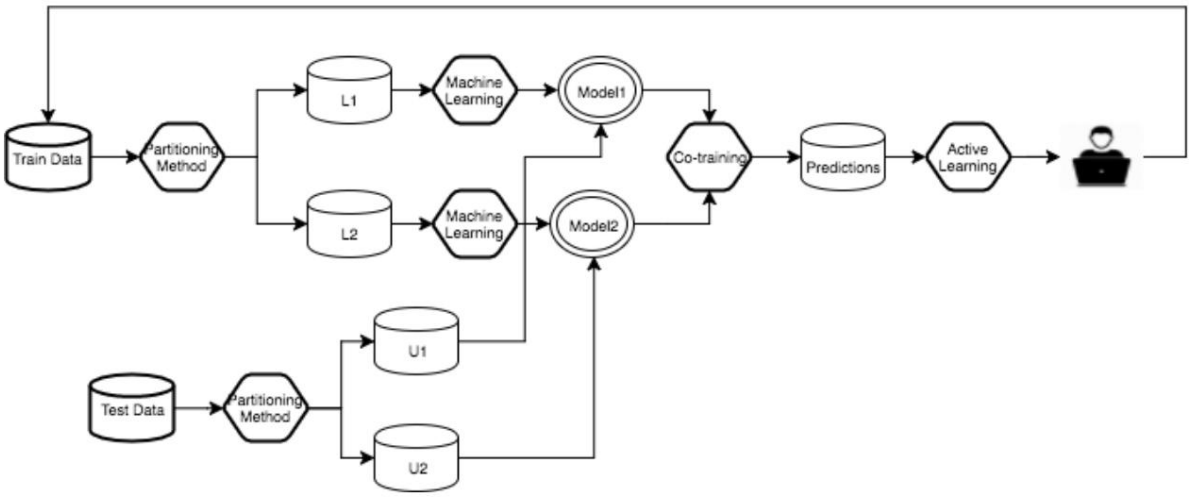
CEAL bao gồm bốn bước:

1. Phân vùng dữ liệu của một khung nhìn thành hai khung nhìn riêng biệt, độc lập. Chúng tôi cố gắng để phân vùng hai quan điểm này một cách công bằng từ góc độ thông tin. Như vậy, hai bộ dữ liệu được gắn nhãn L1 và L2 và dữ liệu không được gắn nhãn U1 và U2 sẽ được || Ư || >> || L || (|| U || và || L || là số lượng cá thể trong tập hợp không được gắn nhãn và tập hợp có nhãn).

2. Sử dụng một thuật toán từ tập hợp các thuật toán của chúng tôi để đào tạo một mô hình cho dữ liệu được gắn nhãn và có được các mô hình m1 và m2 (mỗi mô hình được huấn luyện từ một khung nhìn).

3. Triển khai các mô hình đã thu được ở bước trước trong dữ liệu không được gắn nhãn và thu được xác suất của mỗi cá thể thuộc bất kỳ lớp nào trong số các lớp P (Ci | Uj). Ci là lớp tôi đại diện trong đó Uj là j trường hợp không được gắn nhãn thứ .

4. Sử dụng các xác suất đã tính từ bước trước để tìm ra điều không chắc chắn nhất và gửi nó đến oracle để tìm nhãn của nó và thêm thể hiện được truy vấn với nhãn của nó cho các tập hợp có nhãn.



Hình 2.1 : Sơ đồ khối thuật toán CEAL.

Sau đó, quá trình này tiếp tục cho một số lần lặp nhất định. Quá trình này được minh họa tại Hình 2.1. Ở mỗi bước, các mô hình đã học đang được thử nghiệm trên các mô hình hoàn toàn riêng biệt kiểm tra tập dữ liệu và sử dụng tổng hợp các xác suất cho hai quan điểm cho thấy argmax xác suất như lớp dự đoán.

$$\hat{y} = \operatorname{argmax}_{j \in \{1,2\}} P(C_j \mid X) \tag{2.1}$$

trong đó j là chỉ số của lớp và P (Cj | X) là xác suất của X thuộc lớp j.

2.2.1.1 Phân vùng tính năng

Tài liệu đồng đào tạo gốc [3] đưa ra giả định làm việc với dữ liệu tách biệt, độc lập và tự túc. Tuy nhiên, trong các bộ thử nghiệm của mình, chúng tôi chỉ có dữ liệu xem một lần. Để triển khai CEAL, Co-Active hoặc SSLCA, có cần phải phân vùng các tập dữ liệu của chúng tôi như trong Đồng đào tạo. Do đó, chúng tôi cố gắng phân vùng bộ dữ liệu trên các tính năng theo cách mà mỗi chế độ xem (phân vùng) đều có giá trị hợp lý (gần với bằng nhau) lượng thông tin và giả định một cách ngây thơ rằng các tính năng này là độc lập và tự túc. Để đạt được sự công bằng này Thu thập thông tin, Phân tích phương sai (ANOVA), Chi-Square và ReliefF đã được sử dụng để lựa chọn tính năng. Này thuật toán hoạt động dựa trên ước tính thông tin mà mỗi tính năng có thể giữ.

Chắc chắn, một số thông tin có giá trị tồn tại trong sự kết hợp của các tính năng sẽ bị bỏ sót khi chia các tính năng thành hai tập hợp con, nhưng nó sẽ được minh họa rõ ràng

được bù đắp bằng các khía cạnh khác của thuật toán của chúng tôi.

Thông tin thu được

Mức tăng thông tin là số liệu để cho biết mức độ không chắc chắn sẽ được giải quyết trong trường hợp người ta nhận thức được điều kiện của một đối tượng địa lý [5]. Công thức của nó cho một tập hợp tính năng D được đưa ra dưới đây:

$$\text{Trong fo (D)} = -\sum_{i=1}^{m} \text{pi log2 (pi)} \quad (2.2)$$

$$\text{Trong f oA (D)} = \sum_{j=1}^{v} \frac{\text{Trong fo (Dj)}}{D} \quad (2.3)$$

trong đó m là số nhãn có thể có và pi là xác suất thuộc về thứ i nhãn cho một ví dụ. pi được tính bằng cách tìm tỷ lệ xuất hiện của lớp đó trong toàn bộ tập dữ liệu. pi log2 (pi) là entropy khi đưa nhãn thứ i cho đặc trưng. j là chỉ báo của từng tính năng mà Information Gain đang cố gắng tìm ra ảnh hưởng của nó đối với entropy của mô hình. Vì vậy, trong f oA (D) là lượng entropi giảm theo khẳng định của nhãn của từng tính năng của nó và giá trị này được gọi là Mức tăng thông tin. Có thể thấy, nó dựa trên Lý thuyết thông tin và đã được sử dụng thành công trong Cây quyết định trong việc hình thành các nhánh. Trong luận án này, Mức tăng thông tin được sử dụng để phân vùng đơn xem các tập dữ liệu thành hai chế độ xem riêng biệt và khá phân tán. Thứ nhất, Thông tin thu được cho mỗi tính năng được tính toán và các tính năng được sắp xếp theo giá trị khuếch đại của chúng. Sau đó, chỉ định các tính năng đầu tiên, thứ ba, thứ năm, v.v. vào chế độ xem số một và chỉ định thứ hai, thứ tư, thứ sáu, v.v. để xem số hai. Bằng cách này, người ta giả định rằng thu được hai chế độ xem khác biệt có Mức tăng thông tin gần nhau.

Chi-Square

Chi-Square (χ^2) thống kê tính toán sự độc lập của hai biến ngẫu nhiên [22]. Trong khi tính toán giữa một đối tượng địa lý và một nhãn phiên bản, Chi-Square đã cho chúng tôi thấy lượng tương quan của từng đối tượng địa lý với nhãn. Chúng tôi có thể xem xét tính năng có mối tương quan cao nhất với nhãn là quan trọng nhất. Phương pháp này là thường được sử dụng để lựa chọn tính năng. Do đó, các tính năng được sắp xếp như trong Thông tin Đạt được, nhưng lần này dựa trên thống kê Chi-Square và được phân thành hai chế độ xem. Các

công thức của thống kê Chi-Square như sau.

$$2\chi = \sum_{i=1}^{n} \frac{(Oi - Ni)^2}{Ei} \quad (2.4)$$

trong đó Ei là phương sai, Oi là sự quan sát đến từ mỗi trường hợp và Ni là nhãn
chẳng hạn. Điều này được tính toán cho từng tính năng và chúng tôi có thể nhận được danh sách χ^2 rằng chúng tôi
có thể sắp xếp dựa trên mức độ ảnh hưởng của chúng đối với nhãn.

ANOVA

ANOVA (Phân tích phương sai) [23] là một phương pháp được sử dụng để lựa chọn đối tượng địa lý. ANOVA
tính toán tác động của từng tính năng đối với nhãn và bất kỳ tính năng nào có nhiều nhất
tác động lên nhãn là tính năng cung cấp thông tin nhiều nhất. Do đó, như trong các phương pháp trước,
chúng tôi sắp xếp các đối tượng địa lý dựa trên giá trị ANOVA và gán chúng cho hai chế độ xem. Giá trị
được tính cho ANOVA là hiệp phương sai của mỗi đối tượng có nhãn của nó. Điều này có nghĩa là
khi một tính năng thay đổi, nó tạo ra sự khác biệt như thế nào đối với nhãn dẫn đến kiến thức
về cách chúng tôi có thể dự đoán nhãn chỉ bằng cách sử dụng tính năng này.

Cứu trợF

ReliefF là một phương pháp lặp lại để lựa chọn tính năng [24]. Nó hoạt động dựa trên
thuật toán láng giềng. Trong bước đầu tiên, một vectơ trọng lượng được tạo bởi độ dài bằng
với số lượng các tính năng. Sau đó, ở mỗi lần lặp lại, một phiên bản ngẫu nhiên được chọn và
các trường hợp được sắp xếp dựa trên khoảng cách Manhattan (định mức L-1) trong hai nhóm. Các
nhóm đầu tiên chứa các thể hiện của lớp đối diện (bỏ sót) và nhóm thứ hai
chứa các cá thể thuộc cùng một lớp (lượt truy cập). Sau đó, ví dụ gần nhất
của mỗi nhóm đến một phiên bản được chọn ngẫu nhiên, tương ứng, được gọi là nhóm gần như bỏ lỡ hoặc
đánh gần, được chọn. Sau đó, dựa trên sự khác biệt giữa giá trị của các tính năng trong
hai trường hợp này và trường hợp được chọn ngẫu nhiên của chúng tôi, vectơ trọng lượng dựa trên
phương trình dưới đây được cập nhật.

$$\omega_{ij} = \omega_{ij} + |x_i - \text{nearhiti}| - |x_i - \text{near ủy}| \quad (2,5)$$

j đâu ω tôi cho thấy trọng số tương ứng với đặc điểm i ở lần lặp j và xi là giá trị của

các trường hợp được chọn ngẫu nhiên tính năng thứ i cũng như nearhiti và nearmiss .

Trong phiên bản gốc của Cứu trợ khoảng cách Euclide (định mức L-2) thay vì

Do đó, khoảng cách Manhattan được sử dụng, công thức cập nhật như sau:

$$\omega_{ij}^{1} = \omega_{ij} + (xi - nearhiti)^{2} - (xi - near\ ủy)^{2} \quad (2,6)$$

Vào cuối lần lặp thứ m, vectơ trọng lượng sẽ đại diện cho ảnh hưởng của mỗi

tính năng trên nhãn. Nó có thể được coi là lượng thông tin của mỗi tính năng

duy trì. Chúng tôi chia các đối tượng địa lý thành hai nhóm như chúng tôi đã làm trong các phương pháp trước.

2.2.1.2 Đào tạo bộ phân loại cơ sở

Trong luận án này, một số bộ phân loại đã được sử dụng làm bộ phân loại cơ sở. Gaussian

Naive Bayes được sử dụng làm đại diện cho các mô hình tổng hợp thống kê, tạo ra

xác suất của các dự đoán vốn có. Xác suất này rất quan trọng để tính toán

sự tự tin trong giai đoạn lựa chọn truy vấn. Ngoài ra, K-NN và cây quyết định

bộ phân loại được sử dụng, cả hai đều là thuật toán đơn giản và đào tạo nhanh. Hơn nữa,

một bộ phân loại rừng ngẫu nhiên được sử dụng như một mô hình mạnh mẽ dựa trên việc tăng cường. Cuối cùng

thuật toán đã được sử dụng là bộ phân loại Multilayer Perceptron (MLP) thuộc

đối với họ mạng nơ-ron, đây cũng là một thuật toán mạnh mẽ. Để đào tạo

MLP, lan truyền ngược và trình tối ưu hóa Ước tính Moment Thích ứng (Adam) [25]

đã được sử dụng. Chúng ta sẽ thảo luận chi tiết về các thuật toán trong phần kết quả.

Naive Bayes và các mô hình MLP đã được đào tạo theo cách lặp đi lặp lại, như trên mạng

học tập. Điều này có nghĩa là tại mỗi lần lặp lại khi một phiên bản mới xuất hiện, mô hình sẽ

không được đào tạo từ đầu, nhưng mô hình cũ chỉ cập nhật trọng lượng và thông số

đang xem xét trường hợp mới. Điều này dẫn đến một cải tiến lớn về tính toán trong

sử dụng hai thuật toán này. Tuy nhiên, đặc điểm của các thuật toán khác không

cho phép sử dụng thủ thuật này và toàn bộ mô hình đào tạo là cần thiết ở mỗi lần lặp.

Các thuật toán của chúng tôi đang hoạt động dựa trên xác suất của mỗi trường hợp thuộc về các lớp khác nhau Pj (Ci | X); Ci là tôi ở đâu thứ tự lớp và j {1,2} là chỉ số của mô hình được đào tạo dựa trên chế độ xem. Bằng cách sử dụng các xác suất này, cả nhãn cuối cùng dự đoán và trường hợp tiếp theo được xác định. Như đã được đề cập, các mô hình thống kê tự nhiên đã cho chúng tôi xác suất của mỗi dự đoán (dự đoán được thực hiện bằng cách sử dụng xác suất), nhưng các mô hình phi thống kê như rừng ngẫu nhiên chỉ cho chúng ta một giá trị cho hồi quy và bằng cách hiệu chỉnh kết quả hồi quy này (chuẩn hóa đầu ra hồi quy để ràng buộc nó trong khoảng từ 0 đến 1), một giá trị có thể được coi là xác suất cho mô hình sẽ thu được với một giá trị gần đúng tốt.

2.2.1.3 Lựa chọn truy vấn

Lựa chọn truy vấn là phần quan trọng nhất trong thuật toán của chúng tôi. Truy vấn tốt nhất là truy vấn qua một phiên bản trong đó nếu nó được thêm vào tập dữ liệu, nó sẽ tạo ra sự cải tiến. Do đó, bằng cách này, một mô hình có thể được huấn luyện với ít truy vấn nhất.

Các phương pháp khác nhau đã được triển khai cho Học tập tích cực, đã đã thảo luận trong chương trước. Phương pháp phổ biến nhất cho bộ phân loại nhị phân là chọn các trường hợp có độ chắc chắn (xác suất dự đoán) thấp (gần 0,5 với bộ phân loại nhị phân). Trong những trường hợp này, mô hình được đào tạo không thể phát hiện ra lớp có độ tin cậy cao.

Trong nghiên cứu này, sự chắc chắn của những người hàng xóm cũng được xem xét trong việc tìm ra những người kém tự tin các thành viên. Chúng tôi gọi đây là mức độ đóng góp, là thước đo để tính toán độ chắc chắn cao của một thành viên và độ chắc chắn thấp của những người hàng xóm gần nhất. Trong một trường hợp trong đó một ví dụ có độ chắc chắn cao, nhưng những người hàng xóm gần nhất của nó có độ chắc chắn thấp, điều này có nghĩa là chúng tôi có một ý tưởng rất mơ hồ về khu vực. Trong trường hợp này, truy vấn nó sẽ không chỉ tìm thấy nhãn của phiên bản, nhưng cũng sẽ cho phép người ta nắm được những gì đang xảy ra trong những người hàng xóm của nó và có thể giúp gắn nhãn một nhóm lớn các trường hợp một cách chính xác. Các giá trị đóng góp được tính như sau:

$$\text{Đóng góp (Con f, xi)} = \frac{1}{K \quad \text{Con f (xi , c))}} \times N(xi) \quad [\text{Con f (xi , c)} \quad \text{Con f (x, c)}] \quad (2,7)$$

Con f (xi , c) là xác suất của trường hợp xi có nhãn là c và K là số

của những người hàng xóm gần nhất mà chúng tôi đã chọn để đánh giá. Chúng tôi đã đặt K thành năm trong thời gian này

tìm kiếm.

2.2.2 Thuật toán học cùng hoạt động

Yuce và cộng sự. [21] đã đề xuất một phương pháp gọi là Học tập cùng chủ động. Họ đã làm việc với

tập dữ liệu Định dạng dữ liệu châu Âu về chế độ ngủ (EDF) và cố gắng dự đoán các giai đoạn ngủ.

Họ đã sử dụng Fast Fourier Transform (FFT) và Empirical Mode Decomposition

(EMD) để trích xuất hai chế độ xem riêng biệt từ một tập dữ liệu. Trong luận điểm này, một chút khác biệt

cách tiếp cận dựa trên Co-Active đã được đề xuất cho vấn đề này. Sự khác biệt chính

là chúng ta có một tập dữ liệu chỉ với một chế độ xem.

Trong việc triển khai các tập dữ liệu thuật toán Co-Active Learning của chúng tôi, đã

được phân chia thành hai chế độ xem, như với CEAL. Sau đó, bước đào tạo diễn ra và

xác suất được trích xuất. Trong bước thứ ba, các thành viên không chắc chắn được nhóm lại thành k

cụm sử dụng k-mean với k là số lớp. Sau đó, ví dụ gần nhất

trung bình của mỗi cụm được coi là phiên bản không được gắn nhãn nhiều thông tin nhất.

Khoảng cách Euclide được sử dụng để tính toán khoảng cách của các cá thể. Trong bước này,

giả định rằng các cá thể được phân phối công bằng về mặt không gian được thực hiện. Trong bước tiếp theo,

nhãn của trường hợp này được truy vấn. Bằng cách này, mô hình trình bày tốt nhất

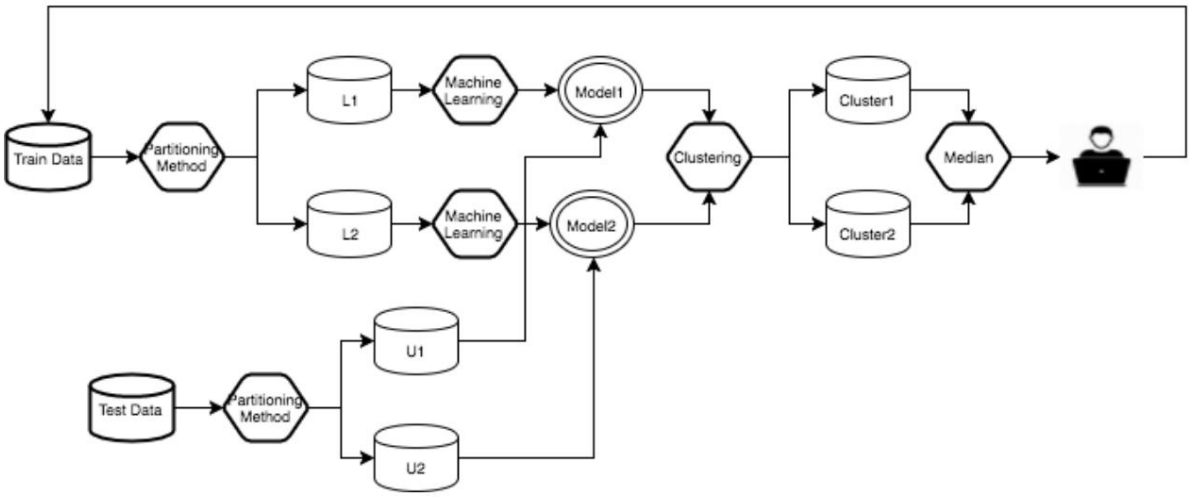
với ít thành viên nhất có thể đạt được Hình 2.2.

2.2.2.1 Phép đo độ không chắc chắn

Để chọn các thành viên không chắc chắn, xác suất của mỗi mô hình dự đoán từ mỗi

lượt xem được tính cho tất cả các trường hợp. Các thành viên được chọn này được sử dụng để thành lập

các cụm. Sau đó, một ngưỡng, θ được sử dụng với công thức này:



Hình 2.2 : Sơ đồ khối thuật toán Co-Active Learning.

$$| P (y = + | x) \quad P (y = - | x) | < \theta \tag{2,8}$$

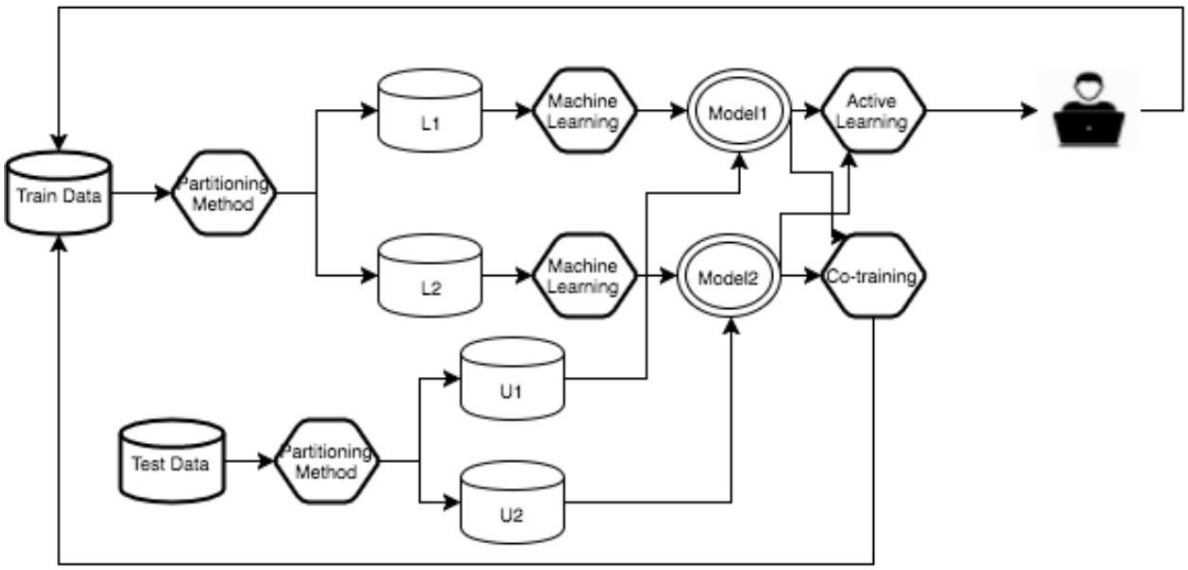
trong đó | P (y = + | x) - P (y = - | x) | là sự khác biệt giữa xác suất của trường hợp x thuộc các lớp + và -. θ là hằng số được sử dụng làm giới hạn thấp hơn cho sự khác biệt giữa các xác suất để có thể được chọn để phân cụm.

2.2.2.2 Cài đặt ngưỡng

Trong phương pháp này, giá trị θ là một tham số, ví dụ: đặt θ thành 0 nghĩa là chọn các trường hợp mà bộ phân loại của chúng tôi không đồng ý với nó mà không xem xét từng bí mật. Việc chọn θ gần bằng 0 có nghĩa là xem xét các trường hợp mà cả hai bộ phân loại thiếu tự tin khi dự đoán nó. Và cũng có θ gần bằng 0, nhưng nhỏ hơn 0 có nghĩa là những trường hợp mà cả hai bộ phân loại đều ít tin tưởng hơn về chúng và cũng không đồng ý về nhãn của họ.

2.2.2.3 Hạn chế của thuật toán đồng hoạt động

Vấn đề chính của Học tập cùng chủ động là nó giả định rằng các điểm dữ liệu được trải rộng mạch lạc về mặt không gian. Nó có nghĩa là những cá thể thuộc cùng một lớp là gần nhất bằng phép đo euclide và có thể được phân cụm dễ dàng bằng cách sử dụng phương tiện k. Và trong trường hợp, điều đó các thành viên cùng lớp sẽ ở trong các cụm không gian khác nhau, nó sẽ không hoạt động tốt.



Hình 2.3 : Sơ đồ khối thuật toán SSLCA.

2.2.3 Học tập bán giám sát kết hợp Đồng đào tạo với Học tập tích cực

(SSLCA)

Là phương pháp thứ ba, việc triển khai chính xác SSLCA do Zhang et al đề xuất. có
đã được thực hiện [20]. Phương pháp này đã được thực hiện để đảm bảo một tài liệu tham khảo
điểm. Rất tiếc, chúng tôi không thể lấy mã của tác giả hoặc kết quả thực nghiệm và
trong bài báo của họ, họ chỉ bao gồm một nhóm các số liệu. Hơn nữa, họ không cho
chi tiết về các thử nghiệm của họ, vì vậy cài đặt tạo ra kết quả gần nhất đã
được thực hiện. Sơ đồ khối của thuật toán SSLCA được cho trong Hình 2.3.

Sự khác biệt lớn nhất của SSLCA với phương pháp của chúng tôi là trong SSLCA họ đã sử dụng
Đồng đào tạo và Học tập tích cực trong một kiến trúc song song. Nó có nghĩa là ở mỗi lần lặp lại
hai bước Học tập tích cực và Đồng đào tạo diễn ra. Trong bước Học tập tích cực,
cá thể không chắc chắn nhất được chọn để gửi tới oracle để lấy nhãn. Trong Đồng đào tạo
bước, trường hợp mâu thuẫn nhất mà một trong các mô hình chắc chắn về nhãn và
khác không chắc chắn về nhãn của nó được chọn. Nhãn nhất định được chỉ định làm phiên bản của
nhãn mác. Sau đó, hai trường hợp này với nhãn được chỉ định của chúng được thêm vào tập dữ liệu chưa được gắn nhãn.
Trong thuật toán CEAL, chúng tôi đang nhận các giá trị từ Đồng đào tạo nhưng không sử dụng chúng trong
các trường hợp ghi nhãn.

Một đổi mới khác trong nghiên cứu của chúng tôi là sử dụng Chi-Square, ANOVA và ReliefF để phân chia các tính năng thành hai chế độ xem, ngoài Mức tăng thông tin.

3. Kết quả thực nghiệm

Trong phần này, trước tiên chúng tôi sẽ cung cấp thông tin chi tiết về các bộ dữ liệu. Tiếp theo, chúng tôi sẽ đề cập đến thiết lập thí nghiệm và sau đó kết luận với kết quả phân loại.

3.1 Bộ dữ liệu

Tám bộ dữ liệu khác nhau đã được sử dụng để kiểm tra giả thuyết được đề xuất. Các bộ dữ liệu này là được xuất bản trong Kho lưu trữ Máy học của Đại học California, Irvine (UCI) [26]. Kho lưu trữ Học máy UCI là một kho lưu trữ các tập dữ liệu được thu thập bởi Đại học California, Irvine. Các tập dữ liệu đã được sử dụng trong Bảng 3.1.

3.2 Thiết lập thử nghiệm

Các thử nghiệm trên tất cả tám tập dữ liệu được thực hiện theo kiểu xác nhận chéo 10 lần. Tại mỗi gấp lại, 10% tập dữ liệu được dành cho thử nghiệm và 90% còn lại được sử dụng cho đào tạo và xác nhận. Tất cả các kết quả được báo cáo bằng cách sử dụng thử nghiệm 10% này và thực hiện trung bình của 10 lần gấp.

Từ 90% dành riêng cho đào tạo, 20% trường hợp được sử dụng để đào tạo như các phiên bản được gắn nhãn và 80% còn lại là nhóm không được gắn nhãn. Mô hình thu được từ quá trình đào tạo đã được kiểm tra trên tập hợp không được gắn nhãn và dựa trên kết quả dự đoán và xác suất ở mỗi lần lặp, sử dụng ba thuật toán khác nhau, một trường hợp là được chọn để truy vấn. Kết quả của các thuật toán này sẽ được thảo luận trong phần tiếp theo trong chi tiết. Phiên bản mới được gắn nhãn này đã được thêm vào tập huấn luyện và các bộ phân loại của nó đã được đào tạo lại.

Cài đặt này được thực hiện cho tám bộ dữ liệu, ba phương pháp (CEAL, Co-Active, SSLCA), bốn phương pháp phân vùng (Tăng thông tin, ANOVA, Chi-Square và ReliefF)

Bảng 3.1 : Thuộc tính của bộ dữ liệu được sử dụng trong kết quả thực nghiệm.

	số phiên bản	số tính năng	số lớp
ung thư	698	10	2
cờ vua	3195	37	2
sonar	207	61	2
tầng	389	16	2
điện ly tín dụng	350	35	2
QUẢNG CÁO	3278	1558	2
cử tri	434	17	2
đốt sống	309	7	2

và sử dụng năm thuật toán học máy khác nhau (Naive Bayes, cây quyết định,
K-hàng xóm gần nhất, rừng ngẫu nhiên và MLPs).

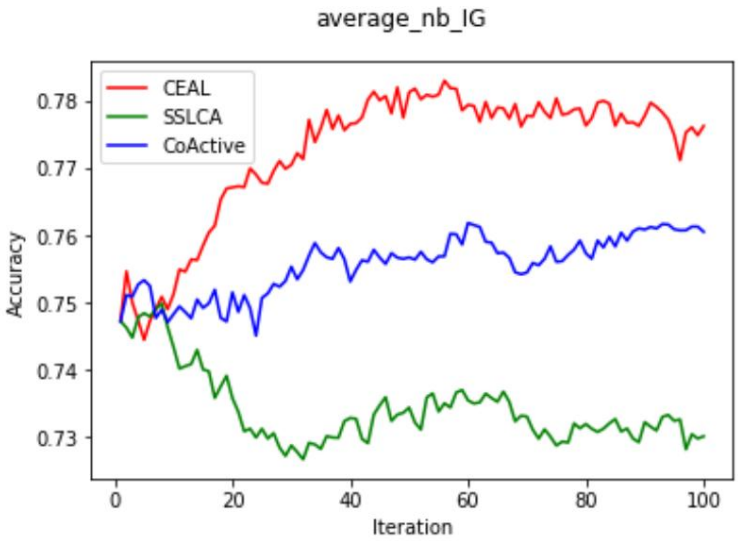
3.2.1 So sánh CEAL với Đồng đào tạo và Thuật toán học tập tích cực

Kết quả của CEAL lần đầu tiên được so sánh với các thuật toán cơ bản cho Đồng đào tạo và
Học tập tích cực. Naive Bayes và Information Gain được sử dụng cho máy mặc định
thuật toán học và phương pháp phân vùng, tương ứng. Đối với Học tập tích cực, điều cơ bản
thuật toán dựa trên bất đồng đã được triển khai và Đồng đào tạo đã được triển khai
dựa trên [10]. Tất cả các thuật toán đều chạy trong 100 lần lặp.

Kết quả được báo cáo trong Bảng 3.2. Như chúng ta có thể thấy, CEAL vượt qua cả Active
Học tập và Đồng đào tạo trong hầu hết các trường hợp. Đồng đào tạo là phương pháp yếu nhất trong số
ba thuật toán và trong hai trường hợp, nó làm xấu kết quả sau 100 lần lặp.
Active Learning đứng sau CEAL trong mọi trường hợp ngoại trừ tập dữ liệu cờ vua. Mỗi
ví dụ trong tập dữ liệu cờ vua là một tập hợp các nước đi trong một trò chơi cờ vua. Kể từ khi di chuyển
có liên quan nhiều đến nhau, chia chúng thành hai quan điểm khiến CEAL hoạt động
tệ hơn so với Học tập tích cực với một chế độ xem duy nhất đã làm. Đồng đào tạo hoạt động mà không có
bất kỳ sự giám sát hoặc trợ giúp nào từ bất kỳ tác nhân nào của con người / máy móc; do đó, chất lượng thấp hơn của nó là
bình thường, và thậm chí là ấn tượng.

Bảng 3.2 : Cải thiện độ chính xác phân loại cho các thuật toán sau khi không gắn nhãn sử dụng dữ liệu. (Bộ phân loại cơ sở: Naive Bayes, Tách tính năng: Tăng thông tin).

	Đồng đào tạo	Học tập tích cực CEAL	
ung thư	-0,8	2,2 3,26	
cờ vua	12.1	37,7	30,87
sonar	7,8	28,4	36,78
tầng	6,5	14,1	25.47
điện ly tín dụng	-1,7	4,01	9.16
QUẢNG CÁO	10.4	33.3	38.42
cử tri	9.2	23	45.31
đốt sống	6.9	11.1	19,22



Hình 3.1 : Độ chính xác phân loại trung bình trên các bộ dữ liệu sử dụng Naive Bayes Bộ phân loại và Tách thông tin thu được.

3.3 So sánh các thuật toán CEAL, SSLCA và Co-Active

Thuật toán CEAL được đề xuất đã được so sánh với các thuật toán Co-Active và SSLCA. Co-Active đã được đề xuất trong [21] và đã được sửa đổi cho các bộ vấn đề của chúng tôi, và SSLCA đã được giới thiệu trong [20] và đã được thực hiện như đã được báo cáo.

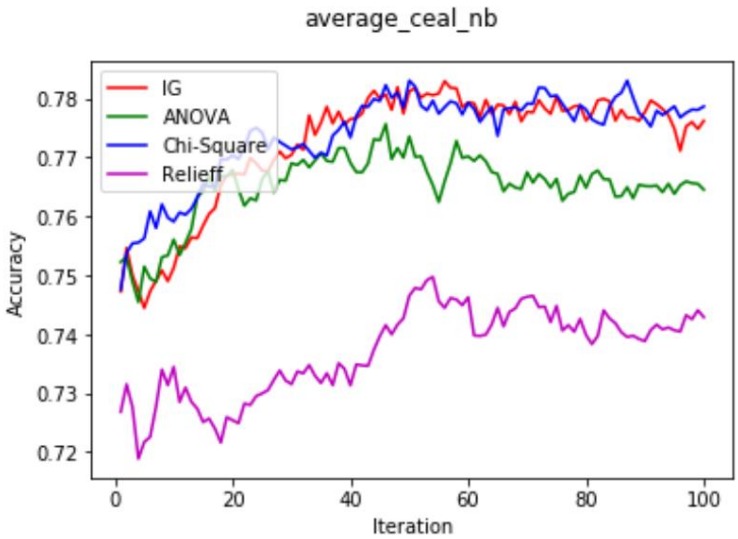
Naive Bayes được coi là thuật toán mặc định và Mức tăng thông tin là phương pháp phân vùng mặc định. Độ chính xác phân loại trung bình cho tám bộ dữ liệu được báo cáo trong Hình 3.1. Lý do chọn Naive Bayes và Information Gain như thuật toán mặc định là có thể so sánh với [20] đang hoạt động chỉ với các thuật toán.

Kết quả cho thấy rằng trong hầu hết các trường hợp, SSLCA làm giảm chất lượng của mô hình trong lần đầu tiên
sự lặp lại. Điều này chủ yếu là do bước Đồng đào tạo mà nó có. Như chúng ta có thể thấy,
kết quả tốt nhất nhận được từ CEAL, và mức trung bình tốt nhất là từ Co-Active.

Trong Bảng 3.3, chúng ta có thể thấy sự cải tiến của mỗi thuật toán được thực hiện trong 100 lần lặp. Này
kết quả được báo cáo theo tỷ lệ phần trăm. Trong số tám bộ dữ liệu, CEAL vượt trội hơn so với các bộ dữ liệu khác
phương pháp trong sáu bộ dữ liệu và SSLCA trong hai trong số đó.

3.4 Kết quả phân vùng tính năng

Bốn phương pháp khác nhau đã được sử dụng để chia bộ dữ liệu thành hai dạng xem riêng biệt.
Các phương pháp này là Tăng thông tin, Chi-Square, ANOVA và ReliefF và là
chủ yếu được sử dụng để lựa chọn tính năng và chia sẻ tính năng một cách công bằng giữa hai tập dữ liệu.



Hình 3.2 : Độ chính xác phân loại trung bình của bộ dữ liệu sử dụng thuật toán CEAL và NB làm bộ phân loại.

Như được hiển thị trong phần phụ trước, thuật toán CEAL hoạt động tốt hơn các thuật toán tương ứng của nó xét về độ chính xác của phân loại. Do đó, trong phần này, CEAL được đặt làm mặc định vì nó là phương pháp tốt nhất trong phần trước và Naive Bayes là đặt làm thuật toán học máy mặc định. Kết quả trung bình cho tám tập dữ liệu được minh họa trong Hình 3.2. Bảng 3.3 minh họa rằng các phương pháp phân vùng không tạo ra sự khác biệt hữu hình, ngoại trừ trong một trường hợp duy nhất, ReliefF tập dữ liệu tín dụng, trong đó chúng vượt quá các phương pháp khác đáng kể.

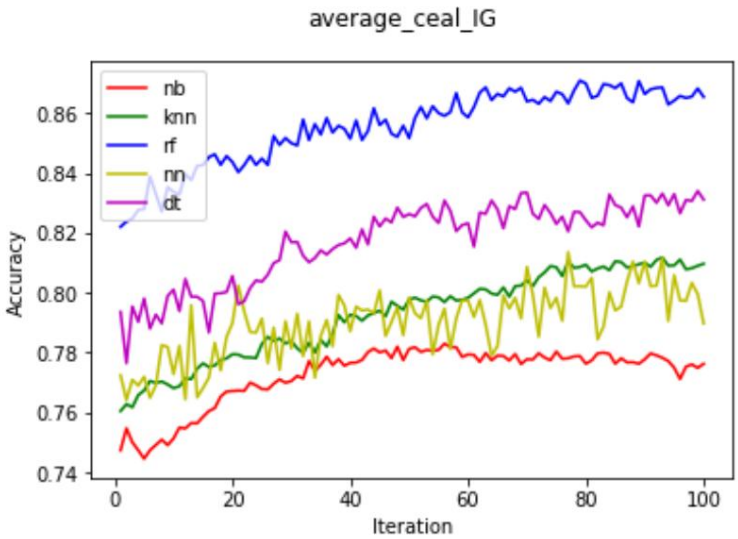
hai phương pháp tạo ra kết quả tốt nhất và hoạt động rất giống nhau.

Information Gain và Chi-Square là hai phương pháp có kết quả tốt nhất, chúng đang hành động rất thân thiết với nhau.

3.5 So sánh các bộ phân loại cơ bản

Các so sánh cuối cùng đã được thu được cho các bộ phân loại cơ sở. Trong thử nghiệm này, bộ phân loại cơ sở của các mô hình đã được thay đổi và độ chính xác phân loại cuối cùng đã được so sánh.

Với mục đích này, Naive Bayes, KNN, mạng nơ-ron, cây quyết định và rừng ngẫu nhiên bộ phân loại đã được so sánh. K trong thuật toán KNN được đặt thành 5. Trong thuật toán mạng nơron kiến trúc tối ưu tổng thể là 2 lớp ẩn và 200 nút cho mỗi lớp. Vì



Hình 3.3 : Độ chính xác phân loại trung bình của bộ dữ liệu sử dụng thuật toán CEAL và IG như thuật toán phân tách tính năng.

rừng ngẫu nhiên 100 cây quyết định được huấn luyện. Đối với thuật toán SSLCA ban đầu, chúng tôi đã sử dụng Mức tăng thông tin làm phương pháp phân vùng cho tất cả các thuật toán. Trận chung kết độ chính xác phân loại trung bình của tám bộ dữ liệu được đưa ra trong Hình 3.3. Bằng được hiển thị trong hình, việc tăng số lượng mẫu dữ liệu không được gắn nhãn sẽ làm tăng độ chính xác phân loại trung bình. Mặt khác, các khu rừng ngẫu nhiên có độ chính xác phân loại. Kết quả tốt nhất được lấy từ rừng ngẫu nhiên và kết quả tồi tệ nhất kết quả được sản xuất bởi Naive Bayes.

3.5.1 Cải thiện độ chính xác phân loại của các thuật toán sau khi không được gắn nhãn sử dụng dữ liệu

Trong Bảng 3.3, tiến trình của mỗi phương pháp kết hợp được báo cáo dựa trên trên Naive Bayes. Mỗi số trong bảng hiển thị sự kết hợp tương ứng của phương pháp luận và cải tiến hơn 100 lần lặp lại theo tỷ lệ phần trăm.

Như chúng ta có thể thấy trong hầu hết các bộ dữ liệu, CEAL hoạt động hiệu quả hơn các thuật toán khác. Ở trong chỉ có hai tập dữ liệu, SSLCA tốt hơn một chút so với CEAL và chúng ta có thể thấy rằng điều này chủ yếu là kết quả của thuật toán tách đặc trưng.

4.KẾT LUẬN

Gần đây, trong kỷ nguyên dữ liệu lớn, số lượng mẫu dữ liệu đã tăng lên rất nhiều.
Tuy nhiên, rất khó để có được các ví dụ được gắn nhãn cho hầu hết các vấn đề. Do đó, trong
trong hầu hết các trường hợp, có thể dễ dàng lấy được một vài mẫu có nhãn và một lượng lớn các mẫu không được dán nhãn
các ví dụ. Trong luận án này, thuật toán CEAL đã được đề xuất để sử dụng dữ liệu không được gắn nhãn
mẫu. Thuật toán được đề xuất là một thuật toán lặp đi lặp lại tính toán độ không đảm bảo
giá trị cho mỗi phiên bản không được gắn nhãn để lựa chọn truy vấn. Các trường hợp đã chọn sau đó
được gắn nhãn bởi một tiên tri và được thêm vào tập huấn luyện.

Thuật toán đề xuất được so sánh với SSLCA được đề xuất gần đây và
Các phương pháp Co-Active và kết quả được sử dụng để so sánh các phương pháp phân vùng,
các chiến lược lựa chọn truy vấn và phương pháp học máy.
trên cùng một tập dữ liệu bằng cách sử dụng cùng một số lần lặp lại phiên bản không được gắn nhãn. ll trong số
các bộ thử nghiệm khác nhau minh họa tính ưu việt của CEAL so với SSLCA và Co-Active
như đã báo cáo trong [27]. CEAL sử dụng sức mạnh của Đồng đào tạo để tăng cường Hoạt động
Học tập và vượt trội so với SSLCA, sử dụng Đồng đào tạo và Học tập tích cực trong
song song. Có một yếu tố Đồng đào tạo trong SSLCA có thể làm giảm hiệu suất của
thuật toán tổng thể khi các mô hình có một tập hợp con dữ liệu được gắn nhãn rất nhỏ được huấn luyện.
Nếu bộ phân loại được đào tạo với ít ví dụ và có độ phân loại rất thấp
độ chính xác, thì thuật toán Đồng đào tạo có thể không cải thiện hiệu suất của
SSLCA. Zheng và cộng sự. đã đào tạo mô hình của họ với một phần dữ liệu đào tạo lớn hơn nhiều,
mặc dù họ không báo cáo cài đặt chính xác của mình, nhưng chúng tôi đã thu được kết quả tương tự trong
kiểm tra khi SSLCA được đào tạo với 60% dữ liệu được gắn nhãn, con số này nhiều hơn những gì
chúng tôi đang sử dụng trong nghiên cứu của mình (5%). Mặt khác, thuật toán Active Learning
không phụ thuộc vào số lượng các ví dụ được gắn nhãn và cũng có thể hoạt động mà không cần bất kỳ
tập dữ liệu được gắn nhãn. Trong kết quả thử nghiệm, các phương pháp phân tách đối tượng địa lý khác nhau là
so.

Sử dụng các thuật toán phân vùng khác nhau không tạo ra sự khác biệt lớn trong bất kỳ tập dữ liệu nào ngoại trừ tập dữ liệu Tín dụng. Trong tập dữ liệu Tín dụng, ReliefF hoạt động tốt hơn ba phần còn lại thuật toán của một lợi nhuận cao. Tuy nhiên, trong các trường hợp khác, tất cả các kết quả đều rất gần nhau, mặc dù Information Gain và Chi-Square nhỉnh hơn một chút những người khác. Chi-Square là một thuật toán rất hiệu quả về mặt tính toán so với sang ReliefF hoặc Information Gain.

Tìm kiếm thuật toán học máy tốt nhất không phải là mục tiêu của nghiên cứu này, nhưng các thử nghiệm đã được triển khai bằng các thuật toán khác nhau và người ta thấy rằng, không có gì ngạc nhiên khi rừng ngẫu nhiên là thuật toán tốt nhất trong hầu hết các trường hợp. Do họ đặc điểm tăng cường, các khu rừng ngẫu nhiên dễ điều chỉnh và dễ tránh trùng lắp, vì vậy nó thu được kết quả hợp lý trong hầu hết các trường hợp. Gaussian Naive Bayes hay hơn hơn mong đợi, và ngoại trừ trường hợp của Tập dữ liệu quảng cáo, nó có kết quả nổi bật. Qua ngược lại, mạng nơ-ron không tốt như mong đợi, có thể do việc sử dụng của cùng một kiến trúc, bao gồm số lớp và số nút, trong các lớp của tất cả các tập dữ liệu. Cây quyết định cũng có kết quả khá tốt do độ phức tạp của nó so với rừng ngẫu nhiên. Việc sử dụng các thuật toán khác nhau đã cho chúng tôi thấy rằng độ chính xác 100% có thể trong một số trường hợp, thậm chí với ít hơn 100 lần lặp.

NGƯỜI GIỚI THIỆU

[1] Haque, MM, Holder, LB, Skinner, MK và Cook, DJ (2013). Học tích cực dựa trên truy vấn tổng quát để xác định các vùng bị methyl hóa khác nhau trong DNA, Giao dịch IEEE / ACM trên Sinh học tính toán và Tin sinh học, 10 (3), 632–644.

[2] Giải quyết, B. (2010). Khảo sát về văn học học tập tích cực, Đại học Wisconsin, Madison, 52 tuổi (55-66), 11 tuổi.

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

[4] Nigam, K. và Ghani, R. (2000). Tìm hiểu hành vi của đồng đào tạo, Kỷ yếu hội thảo KDD-2000 về khai thác văn bản, trang.15–17.

[5] Alpaydin, E. (2014). Giới thiệu về máy học, báo chí MIT.

[6] Nigam, K. và Ghani, R. (2000). Phân tích hiệu quả và khả năng áp dụng của đồng đào tạo, Kỷ yếu hội nghị quốc tế lần thứ chín về quản lý thông tin và tri thức, ACM, trang 86–93.

[7] Kiritchenko, S. và Matwin, S. (2011). Phân loại email với đồng đào tạo, Kỷ yếu Hội nghị năm 2011 của Trung tâm Nghiên cứu Nâng cao về Nghiên cứu Hợp tác, Tập đoàn IBM, tr.301–312.

[8] Pierce, D. và Cardie, C. (2001). Hạn chế của việc đồng đào tạo đối với việc học ngôn ngữ tự nhiên từ các bộ dữ liệu lớn, Kỷ yếu của Hội nghị năm 2001 về các phương pháp thực nghiệm trong xử lý ngôn ngữ tự nhiên, tr.1–9.

[9] Bishop, CM (2006). Nhận dạng mẫu, Học máy, 128, 1–58.

[10] Goldman, S. và Zhou, Y. (2000). Tăng cường học tập có giám sát với không gắn nhãn dữ liệu, ICML, trang.327–334.

[11] Cohn, D., Atlas, L. và Ladner, R. (1994). Cải thiện khả năng tổng quát hóa với hoạt động học tập, Máy học, 15 (2), 201–221.

[12] Balcan, MF và Urner, R. (2016). Học tập tích cực - lý thuyết học tập hiện đại, Bách khoa toàn thư về thuật toán, 8–13.

[13] Giải quyết, B. (2012). Học tập tích cực, Bài giảng tổng hợp về Trí tuệ nhân tạo và Máy học, 6 (1), 1–114.

[14] Dasgupta, S. và Hsu, D. (2008). Lấy mẫu phân cấp cho học tập tích cực, Kỷ yếu hội nghị quốc tế lần thứ 25 về Học máy, ACM, tr.208–215.

[15] Seung, HS, Opper, M. và Sompolinsky, H. (1992). Truy vấn của ủy ban, Kỷ yếu hội thảo thường niên lần thứ năm về Lý thuyết học tập tính toán, ACM, trang.287–294.

[16] Mao, CH, Lee, HM, Parikh, D., Chen, T. và Huang, SY (2009). Phương pháp tiếp cận dựa trên học tập tích cực và đồng đào tạo bán giám sát để phát hiện xâm nhập từ nhiều chế độ xem, Kỷ yếu hội thảo ACM 2009 về Máy tính Ứng dụng, ACM, tr.2042–2048.

[17] Muslea, I., Minton, S. và Knoblock, CA (2000). Chọn mẫu chọn lọc với quan điểm dư thừa, AAAI / IAAI, trang.621–626.

[18] Muslea, I., Minton, S. và Knoblock, CA (2002). Học tập tích cực + bán giám sát = học qua nhiều chế độ xem mạnh mẽ, ICML, tập 2, trang.435–442.

[19] Cheng, J. và Wang, K. (2007). Học tập tích cực để truy xuất hình ảnh với Co-SVM, Nhận dạng mẫu, 40 (1), 330–334.

[20] Zhang, Y., Wen, J., Wang, X. và Jiang, Z. (2014). Học tập bán giám sát kết hợp đồng đào tạo với học tập tích cực, Hệ thống chuyên gia với Ứng dụng, 41 (5), 2372–2378.

[21] Yüce, AB và Yaslan, Y. (2016). Một phương pháp học tập đồng tích cực dựa trên sự bất đồng để phân loại giai đoạn ngủ, Hệ thống, Tín hiệu và Xử lý hình ảnh (IWSSIP), Hội nghị quốc tế 2016 về, IEEE, tr.1–4.

[22] Zheng, Z., Wu, X. và Srihari, R. (2004). Lựa chọn tính năng để phân loại văn bản về dữ liệu không cân bằng, Bản tin Khám phá ACM Sigkdd, 6 (1), 80–89.

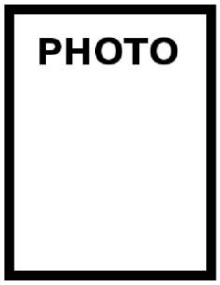
[23] Guyon, I. và Elisseeff, A. (2003). Giới thiệu về lựa chọn biến và tính năng, Tạp chí nghiên cứu máy học, 3 (Mar), 1157–1182.

[24] Kononenko, I. (1994). Các thuộc tính ước tính: phân tích và các phần mở rộng của RELIEF, Hội nghị châu Âu về máy học, Springer, trang.171–182.

[25] Kingma, D. và Ba, J. (2014). Adam: Một phương pháp để tối ưu hóa ngẫu nhiên, arXiv preprint arXiv: 1412.6980.

[26] Kho lưu trữ, UML, https://archive.ics.uci.edu/ml/datasets.html.

[27] Azad, PV và Yaslan, Y. (2017). Sử dụng đồng đào tạo để trao quyền cho học tập tích cực, Hội nghị Ứng dụng Truyền thông và Xử lý Tín hiệu (SIU), lần thứ 25 năm 2017, IEEE, trang 1–4.



SƠ YẾU LÝ LỊCH

Tên Họ: Payam VAKIL ZADEH AZAD

Nơi và Ngày sinh: 06.08.85 Tabriz, Iran

E-Mail: vakil@itu.edu.tr

GIÁO DỤC:

- B.Sc: 2009, Đại học Tabriz, Khoa Kỹ thuật Điện và Máy tính

CÔNG BỐ, TRÌ NH BÀY VÀ CÁC BỐ CỤC VỀ LUẬN ÁN:

- Azad, Payam V. và Yusuf Yaslan. "Sử dụng đồng đào tạo để trao quyền cho việc học tập tích cực." Hội nghị Ứng dụng Truyền thông và Xử lý Tín hiệu (SIU), lần thứ 25 năm 2017. IEEE, 2017.