Phân loại văn bản bán giám sát thông qua tự thử nghiệm

1.Giới thiệu:

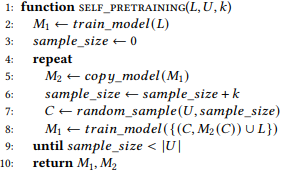
Thành công do khả năng tổng quát hóa của mạng nơ-ron. Các bộ phân loại hiện đại dựa vào các tập huấn luyện lớn, giới thiệu các phép nhúng từ ngữ theo ngữ cảnh và đào tạo trước mô hình ngôn ngữ làm giảm đáng kể nhu cầu chú thích dữ liệu thủ công. Mô hình thần kinh hiện đại vẫn có xu hướng bị quá tải, đặc biệt là ở những khu vực có mô hình ngôn ngữ chuyên biệt và thưa thớt. Các lĩnh vực này bao gồm rất nhiều nhưng không giới hạn ở: lĩnh vực pháp lý, lĩnh vực y tế và lĩnh vực truyền thông xã hội. Giải pháp để giải quyết vấn đề này là tự động xây dựng một tập dữ liệu lớn. Tuy nhiên, điều này không phải lúc nào cũng khả thi. Một cách tiếp cận phương pháp hơn là sử dụng các kỹ thuật cải thiện khả năng tổng quát hóa. Các kỹ thuật này bao gồm khai thác các phép nhúng từ thần kinh, tăng dữ liệu, và thích ứng miền. Nghiên cứu này bổ sung vào phần nội dung tài liệu về học bán giám sát bằng cách sử dụng các thuộc tính của mạng nơ-ron và đề xuất một cách mới để sử dụng dữ liệu không được gắn nhãn. Tập trung vào một trong những lĩnh vực được báo cáo là thiếu đủ dữ liệu đào tạo, tức là khai thác mạng xã hội. Thuật toán được gọi là tự tiền thử nghiệm(lấy cảm hứng từ mô hình tự đào tạo). Thuật toán của chúng tôi là lặp đi lặp lại và trong mỗi lần lặp lại chọn một tập hợp các tài liệu chưa được gắn nhãn để gắn nhãn. Khác với đào tạo, thuật toán không có ngưỡng => không xếp hạng các tài liệu không được gắn nhãn dựa trên những dự đoán của chính nó => thuật toán phù hợp với các mô hình mạng nơ-ron do các đầu ra được hiệu chỉnh kém. Ngoài ra, thuật toán có thể đối phó với vấn đề trôi ngữ nghĩa(khả năng chống lại tiếng ồn trong nhãn giả khi số lần lặp lại tăng lên và tỷ lệ lỗi của trình phân loại bên dưới tăng lên). Tự tiền đào tạo có khả năng sửa đổi nhãn của các tài liệu đã được dán nhãn trước đó. Mô hình sử dụng quy trình chưng cất lặp đi lặp lại. Trong mỗi lần lặp, thông tin thu được trong các lần lặp trước đó được chắt lọc vào bộ phân loại. Nó chuyển một giả thuyết qua các lần lặp lại và sử dụng mô hình học hai giai đoạn. Trong tập hợp các nhãn giả được sử dụng để khởi tạo trình phân loại và tập hợp các tài liệu có nhãn được sử dụng để tinh chỉnh trình phân loại. Tự tiền thử nghiệm điều chỉnh một lịch trình tốc độ học tập mới để tích hợp hiệu quả hai bộ ví dụ đào tạo ồn ào và không gây tiếng ồn. Cuối cùng, để giảm thiểu tác động của các nhãn giả nhiễu(noisy pseudo-labels) trong mỗi lần lặp lại. Những đóng góp trong nghiên cứu của chúng tôi như sau: 1.Đề xuất một khung học tập bán giám sát mới có tên là tự tiền thử nghiệm. 2.Đề xuất một lịch trình tốc độ học tập mới để tích hợp hiệu quả quy trình tối ưu hóa với quy trình học tập bán giám sát hai giai đoạn. 3.Giảm thiểu vấn đề trôi ngữ nghĩa, mô hình phân phối lớp của các nhãn giả(một quá trình ngẫu nhiên qua các lần lặp khởi động và đề xuất một cách tiếp cận mới để chuyển đổi các phân phối lớp). 4.Thực hiện một bộ thử nghiệm toàn diện trên ba tập dữ liệu Twitter công khai và chứng minh mô hình hoạt động tốt hơn một số đường cơ sở hiện đại.

2.Công việc liên quan:

Dữ liệu chưa được gắn nhãn trong học tập bán giám sát. Dữ liệu không được gắn nhãn có thể được khai thác theo nhiều cách. Nó có thể được sử dụng như một siêu nguồn thông tin, nó có thể được sử dụng như một bộ điều chỉnh, hoặc nó có thể được sử dụng trong cài đặt điều chỉnh miền để tạo mối tương quan giữa nguồn và dữ liệu đích. Tự đào tạo là phương pháp học tập bán giám sát lâu đời nhất ra đời từ năm 1965. Ý tưởng tái xuất hiện trong công việc chính của Yarowsky cho các nhiệm vụ NLP vào năm 1995, và một lần nữa trong cộng đồng thị giác máy tính vào năm 2013 như dán nhãn giả. Thuật toán này là một trình bao bọc liên tục sử dụng một thuật toán được giám sát làm mô hình cơ bản. Trong vài năm gần đây, các nghiên cứu đã khám phá hiệu quả của mạng nơ-ron như một mô hình dự đoán cơ bản trong quá trình tự đào tạo. Một biến thể mạng nơ-ron của tri-training với bất đồng được trình bày và cho thấy rằng sự kết hợp là một cơ sở mạnh mẽ đáng ngạc nhiên trong bối cảnh thích ứng miền. Các tác giả trong đề xuất sử dụng điểm phân vị thay vì điểm tin cậy để chọn các nhãn giả tốt nhất; và số khác sử dụng mạng nơ-ron Bayes để chọn các nhãn giả ít tin cậy nhất trong mỗi lần lặp lại. Tự tiền thử nghiệm là mô hình đầu tiên sử dụng quá trình chưng cất mô hình cùng với một giả thuyết để chuyển thông tin qua các lần lặp lại, cho phép nó có khả năng sửa đổi các nhãn giả. Tích hợp mô hình đào tạo trước và tinh chỉnh với tự đào tạo, sử dụng quy trình tối ưu hóa hiệu quả cùng với kỹ thuật nhiễu loạn để giảm thiểu tác động tiêu cực của các nhãn giả nhiễu. Có các nghiên cứu liên quan chặt chẽ khác. Ngoài các nghiên cứu trên, tự tiền thử nghiệm cũng liên quan đến các nghiên cứu về chưng cất mô hình và tổ hợp thời gian. Chưng cất mô hình đã được đề xuất để chuyển kiến thức từ mô hình này sang mô hình khác.

3.Học nửa giám sát thông qua tự tiền thử nghiệm:

Trong thuật toán tự đào tạo, một tập hợp nhỏ các tài liệu được gắn nhãn và một tập hợp lớn các tài liệu không được gắn nhãn có sẵn để đào tạo. Thuật toán lặp đi lặp lại và trong mỗi lần lặp lại mô hình dự đoán được đào tạo trên bộ hiện tại và được sử dụng để gắn nhãn tập hợp hiện tại một cách xác suất. Có ba nhược điểm của thuật toán này: 1.Vấn đề trôi ngữ nghĩa. Tác động tiêu cực ngày càng tăng của các nhãn giả nhiễu làm lu mờ lợi ích của việc kết hợp dữ liệu không được gắn nhãn. 2.Phụ thuộc vào hiệu chuẩn mô hình. Nếu bộ phân loại bên dưới không thể lập mô hình chính xác các phân phối lớp, thì nó sẽ không thể xếp hạng đúng các tài liệu ứng viên. 3.Không thể sửa đổi các nhãn giả sau khi chúng được gán cho các tài liệu chưa được gắn nhãn và được bổ sung thêm với tập hợp các tài liệu được gắn nhãn.

 Thuật toán 1 minh họa quá trình tự tiền thử nghiệm ở dạng cơ bản của nó. Ban đầu, được sử dụng để đào tạo M1(Dòng 2), sau đó là các tham số của M1 được sao chép vào M2(Dòng 5). Tập hợp các tài liệu không được gắn nhãn được lấy ngẫu nhiên(Dòng 7). Tập hợp này được gắn nhãn bởi M2 và được sử dụng L để đào tạo lại M1 (Dòng 8). Vai trò của hai M được đảo ngược trong lần lặp tiếp theo. Trong mỗi lần lặp lại, kích thước mẫu được tăng lên k(Dòng 6) và thuật toán dừng khi tập hợp mẫu bao gồm toàn bộ tập hợp. Thuật toán có hai ưu điểm: 1.Để chọn các nhãn giả, phân phối lớp không được tính đến => không có ràng buộc về khả năng của bộ phân loại trong việc xếp hạng các tài liệu không được gắn nhãn. Điều này ngăn không cho mô hình lặp lại chọn một tập hợp cố định gồm các tài liệu không được gắn nhãn trong mỗi lần lặp lại(tập hợp các nhãn giả có độ tin cậy cao) 2.Thông tin được chuyển qua các lần lặp ở dạng giả thuyết hơn là một tập hợp các nhãn giả cố định.

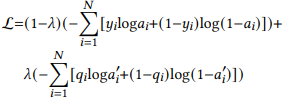
3.1.Chuyển giao giả thuyết và chưng cất lặp lại:

Tự tiền thử chuyển một giả thuyết một hàm đã học từ lần lặp này sang lần lặp tiếp theo. Trong mỗi lần lặp lại, giả thuyết này được sử dụng để tạo một giả thuyết mới bằng cách tạo một tập hợp các nhãn giả và bổ sung chúng với tập hợp các tài liệu được gắn nhãn. Thuật toán gọi là chưng cất mô hình để chuyển kiến thức từ một mô hình lớn (được gọi là giáo viên) sang một mô hình nhỏ (được gọi là học sinh). Tác giả sử dụng phương pháp chưng cất mô hình để chuyển kiến thức từ mạng này sang mạng khác bằng cách sửa đổi lớp softmax như sau:

là lớp cuối của phần tử i, j là số lớp, là dự đoán lớp. Siêu tham số được gọi là nhiệt độ và được sử dụng để làm trơn các dự đoán của lớp. Nhiệt độ cao hơn dẫn đến độ hỗn loạn cao hơn trong các dự đoán. Điều này đặc biệt mong muốn, vì mạng nơ-ron được biết là có độ hỗn loạn thấp trong các dự đoán của chúng. Tự tiền thử nghiệm và chắt lọc hiệu quả các lần lặp trước đó vào mạng sinh viên M1. Trong mỗi lần lặp, thay vì sử dụng giáo viên M2, chúng tôi sử dụng các dự đoán mềm cùng với tập hợp để đào tạo mạng sinh viên-Thuật toán 1(Dòng 8).

3.2.Học tập bán giám sát hai giai đoạn:

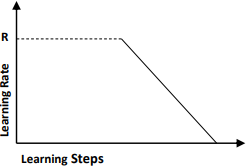
Tự đào tạo gặp phải vấn đề trôi ngữ nghĩa. Sự cố xảy ra khi các lỗi chủ yếu do nhãn giả gây ra tích tụ qua các lần lặp lại và cuối cùng làm sai lệch ranh giới của bộ phân loại. Để giảm thiểu, giải pháp là xử lý tập hợp các nhãn giả và tách thông tin mâu thuẫn với thông tin được lưu trữ trong tập hợp.

Sử dụng hàm mục tiêu trong khi đào tạo M1 sử dụng L: N là số lượng tài liệu trong L, là nhãn thực của tài liệu , là lớp dự đoán của M1 cho , là lớp dự đoán của M1 cho với nhiệt độ cao như được mô tả trong 3.1 và là lớp dự đoán của M2 cho với cùng nhiệt độ M1. λ là một siêu tham số để chi phối trọng số tương đối của hai số hạng (0≤λ≤1). Từ các bậc của số hạng thứ hai trong phương trình 2 ẩn chia tỷ lệ bằng , theo thứ tự để cân bằng tác động của hai thuật ngữ trong lan truyền ngược, nhân các độ dốc này với -phương trình 1. Số hạng đầu tiên trong phương trình 2 ẩn là độ hỗn loạn chéo giữa các nhãn thực cơ bản và xác suất lớp của M1. Số hạng thứ hai là độ hỗn loạn chéo giữa các xác suất lớp của M2 và M1. Chức năng mục tiêu này là một nỗ lực để giữ cân bằng giữa thông tin được chuyển từ các lần lặp trước và thông tin được trích xuất từ tập hợp các tài liệu được gắn nhãn.

3.3.Tỷ lệ học tập hình thang phải:

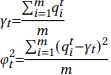
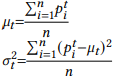
Chiến lược hai giai đoạn này tạo ra một cơ hội thích hợp để nâng cao quá trình tối ưu hóa. Vì các nhãn giả có khả năng bị nhiễu => đề xuất sử dụng tập hợp để khám phá không gian giả thuyết và phát hiện vùng có chứa tối ưu địa phương(localoptima) tốt hơn.

Sử dụng tỷ lệ học tập hình thang phải. t biểu thị bước thời gian hiện tại, là lô tài liệu hiện tại đang được xử lý. là tỷ lệ học tập hiện tại, R là tỷ lệ học tập ban đầu, C là tập hợp các nhãn giả, L là tập hợp các tài liệu được dán nhãn, là số lô nhãn giả và là số lô được dán nhãn. Tỷ lệ học tập bao gồm hai giai đoạn: 1.Tỷ lệ học tập cố định-nơi nhãn giả được sử dụng để đào tạo M1(Thuật toán 1). Giai đoạn này, các tham số mạng có thể tự do cập nhật => có thể khám phá không gian giả thuyết. 2.Tốc độ học tập giảm dần-nơi các tài liệu được gắn nhãn được sử dụng để đào tạo thêm. Giai đoạn này, trình tối ưu hóa ổn định => sử dụng các tài liệu được dán nhãn không gây nhiễu vì một sự xáo trộn nhỏ trong dữ liệu cũng có thể gây ra mất mát đáng kể. Tỷ lệ học tập hai giai đoạn tích hợp hữu cơ với quy trình học tập bán giám sát hai giai đoạn => tốc độ học giảm dần, ngăn mục tiêu của nhiệm vụ thứ hai xóa hoàn toàn kiến thức đã chuyển từ các lần lặp trước.

3.4. Phân phối lớp quán tính:

Lịch trình tỷ lệ học tập tự thử nghiệm. Đường gạch ngang là tốc độ học của mạng trong quá trình đào tạo theo nhãn giả và đường nghiêng là tốc độ học trong quá trình đào tạo bởi các tài liệu được gắn nhãn.

Giả sử quá trình chỉ bao gồm một họ của hai hàm gauss ngẫu nhiên biến là và . là trạng thái của các nhãn giả dương, là trạng thái của các nhãn giả âm. Giá trị trung bình mẫu và phương sai của trong vòng lặp lại t(tức )được đưa ra:

n là số lượng nhãn giả dương trong lần lặp lại t, là xác suất của i-th thuộc về nhãn giả dương vào lớp dương-0.5≤ bởi vì mẫu là dương. Tương ứng, giá trị trung bình mẫu và phương sai của trong vòng lặp t(tức )được đưa ra: m là số lượng nhãn giả âm trong lần lặp lại t và là xác suất của i-th thuộc về nhãn giả âm của lớp âm-0.5≤ và mọi nhãn giả +=1. Trong vòng lặp t+1, các phân phối mẫu của ngẫu nhiên biến và biến thành ∼ N(, ) và ∼ N(, ). Những cập nhật này có thể là do sự ngẫu nhiên trong khởi tạo mô hình, tính ngẫu nhiên trong tập hợp các tài liệu không được gắn nhãn đã chọn trong lần lặp t, hoặc một phần do các nhãn giả nhiễu được giới thiệu trong lần lặp lại t. Sự phân loại sai của M2 trong lần lặp lại t(Thuật toán 1)được sử dụng để đào tạo trước M1, và cuối cùng làm sai lệch phân phối lớp của tập hợp các nhãn giả trong lần lặp t+1. Để giảm bớt tác động của sự nhiễu, sử dụng 2 hàm gauss và như 1 sự kết hợp tuyến tính của phân phối lớp trong các phần lặp t và t+1 và nhãn giả trong thành và các nhãn giả trong thành . Sử dụng cơ chế này sẽ tránh được những thay đổi đột ngột trong việc phân phối nhãn giả.

Thực hiện bước này sau khi tạo các nhãn giả M2 và trước khi sử dụng bộ này để đào tạo trước M1(Thuật toán 1, dòng 8).

4.Thiểt lập thực nghiệm:

Phần này mô tả các tập dữ liệu đã sử dụng, sau đó cung cấp tổng quan ngắn gọn về các mô hình cơ sở và cuối cùng là xem xét chi tiết các thử nghiệm.

4.1.Bộ dữ liệu: