Họ và tên: Phạm Tuấn Anh 15110010

Lỷ Say Hội 15110053

**ĐỒ ÁN 3: MEACHINE LEARNING**

**Đề tài: Aggressive Sampling for Multi-class to Binary Reduction with Applications to Text Classification**

1. **Giới thiệu**

Vấn đề các lớp có quy mô lớn hoặc extreme classification liên quan vấn đề giải quyết số lượng cực lớn các lớp xuất hiện trong các kho lưu trữ văn bản như Wikipedia, Yahoo!, … Và gần đây vấn đề này đã phát triển thành một nhánh học máy phổ biến với nhiều ứng dụng trong gắn thẻ, đề xuất và xuất hạng. Cách tiếp cận để giải quyết bài toán trong bài toán này là One-Versus-All (OVA), trong đó một lớp nhị phân độc lập được học trên mỗi lớp. Mặc dù vô cùng đơn giản, cách tiếp cận này có hai hạn chế chính. Đầu tiên, nó trở nên khó tính toán khi số lượng các lớp phát triển lớn hơn, ảnh hưởng cùng lúc với dự đoán. Thứ hai, nó bị vấn đề mất cân đối.

Gần đây, hai phương pháp tiếp cận chính đã được nghiên cứu để giải quyết với những hạn chế này. Vấn đề đầu tiên, được phân chia rộng rãi trong các phương pháp tree-based và embedding-based, và đã được đề xuất với mục đích giảm không gian hiệu quả của nhãn để kiểm soát sự phức tạp của vấn đề học. Phương pháp tree-based dựa vào các nhánh cây nhị phân mà mỗi lá tương ứng với một lớp và suy luận được thực hiện bằng cách đi qua cây từ trên xuống dưới; một lớp nhị phân được sử dụng ở mỗi nút để xác định nút con cần phát triển. Các phương thức này có phức tạp về thời gian lôgarít với nhược điểm là một nhiệm vụ khó khăn để tìm ra một cấu trúc cây cân bằng có thể phân chia các nhãn lớp. Hơn nữa, mặc dù các chẩn đoán khác nhau đã được phát triển để giải quyết vấn đề không cân bằng, nhưng các phương pháp này phải chịu một số quyết định trước khi đạt được một thể loại cuối cùng, dẫn đến lỗi truyền nhánh và do đó giảm độ chính xác. Mặt khác, embedding-based là phương pháp tiếp cận dự án đầu tiên label-matrixinto một không gian con tuyến tính chiều thấp và sau đó sử dụng một lớp OVA. Tuy nhiên, giả định cấp thấp của ma trận nhãn thường bị vượt qua trong cài đặt extreme classification và các phương pháp này thường dẫn đến lỗi dự đoán cao.

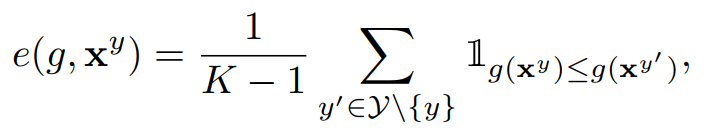
Phương pháp tiếp cận thứ 2 nhằm giảm vấn đề đa lớp thành vấn đề nhị phân, đầu tiên mở rộng tập train băng cách chiếu các cặp observations và classes và một không gian dyadic chiều thấp và sau đó học một trình phân loại đơn để phân chia giữa các cặp được cấu thành với các ví dụ, các class và các cặp constituted của chúng với các ví dụ với các lớp khác. Mặc dù sự dự đoán trong không gian mới tương đối nhanh, việc xây dựng của các dyadic training observations là tương đối tốn thời gian và chiếm ưu thế trong thời gian training và dự đoán.

1. **Chiến lược giảm chiều từ multi-class sang binary**

* Chúng tôi giải quyết vấn đề về phân lớp nhãn đơn multi-class được xác định trên không gian chung

với là input space và là output space, K là số class. là biểu thị như . Hơn nữa, ta giả sử tập huấn luyện là được tạo bởi các ví dụ hoặc các cặp class phân phối theo phân bố xác suất cố định nhưng không xác định D, và ta xem xét hàm dự đoán của lớp

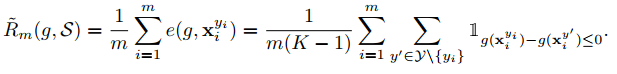
* Ta xác định mất mát của dự đoán trong ví dụ

 (1)

* Trong đó là hàm chỉ định bằng 1 nếu dự đoán là đúng, nếu không sẽ bằng 0. So với lỗi classial multi-class, , theo hàm mất mát (1) ước tính số trung bình của class, cho bất kỳ dữ liệu đầu vào nào , có được điểm cao hơn so với class đúng. Hàm mất mát (1) là một tiêu chuẩn xếp hạng, và SVM nhiều lớp va AdaBoost. MR tối ưu hóa các ham thay thế lồi của sự mất mát. Nó cũng được sử dụng trong xếp hạng label. Mục tiêu của ta là tìn một hàm số với rủi ro nhỏ nhất

 (2)

* Bằng cách giảm thiểu sai số thực nghiệm được xác định là số lượng trung bình của các ví dụ training trong đó, trung bình được ghi điểm thấp hơn cho

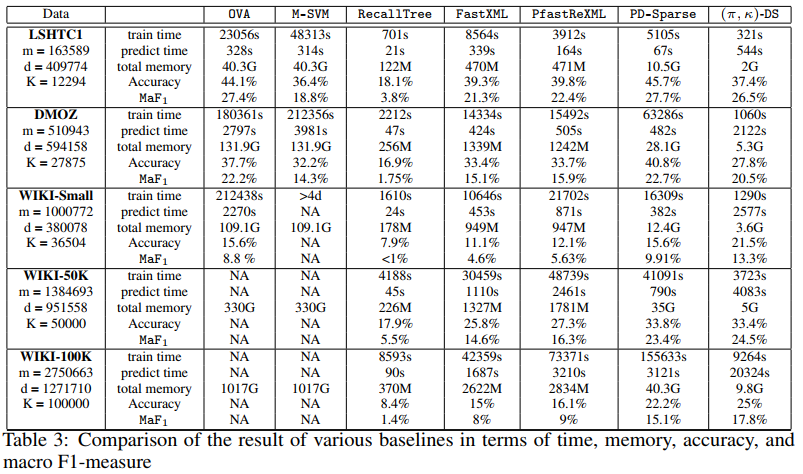
(3)

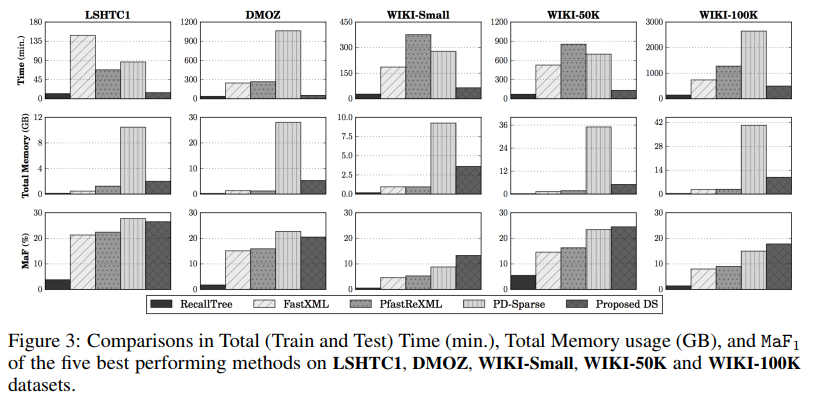
1. **Thử nghiệm**

* Trong phần này, ta cung cấp một đánh giá thực nghiệm về cách tiếp cận giảm được đề xuất với chiến lược lấy mẫu để phân loại nhiều loại tài liệu quy mô lớn. Đầu tiên, chúng tôi trình bày ánh xạ . Sau đó, ta cung cấp so sánh thống kê và tính toán phương pháp của chúng tôi với các phương pháp tiếp cận state-of-the-art large-scale trên các tập dữ liệu phổ biến

1. **Kết quả**

* Các thông số của các tập dữ liệu cùng với kết quả cho các phương pháp so sánh được thể hiện trong Bảng 3. Các kết quả được cung cấp trong khi train và khi dự đoán, tổng dung lượng bộ nhớ và hiệu suất dự đoán được đo bằng độ chính xác và Macro F1 (MaF1). Để hiển thị và so sánh tốt hơn, chúng tôi vẽ các kết quả tương tự như các ô thanh trong Hình 3, chỉ giữ 5 phương pháp tốt nhất trong khi so sánh tổng thời gian chạy và sử dụng bộ nhớ.
* Đầu tiên, chúng tôi quan sát rằng các phương pháp tiếp cận dựa trên cây (FastXML, PfastReXML và RecallTree) có hiệu suất dự đoán tồi tệ hơn so với các phương pháp khác. Điều này là do thực tế là lỗi dự đoán được thực hiện ở top-level của cây không thể được sửa chữa ở các mức thấp hơn, còn được gọi là hiệu ứng xếp tầng. Mặc dù chúng có thời gian chạy và sử dụng bộ nhớ thấp hơn, chúng bị tác dụng phụ này.
* Đối với các tập có quy mô lớn (WIKI-Small, WIKI-50K và WIKI-100K), các giải quyết có hiệu suất dự đoán cạnh tranh là OVA, M-SVM, PD-Sparse và . Tuy nhiên, tiêu chuẩn OVA và M-SVM có độ phức tạp tăng trưởng tuyến tính với K do đó tổng thời gian chạy và sử dụng bộ nhớ rất lớn trên các bộ dữ liệu đó, khiến chúng trở thành không thể. Ví dụ, trên Wiki mẫu tập dữ liệu lớn 100K class, mức tiêu thụ bộ nhớ của cả hai cách tiếp cận vượt quá Terabyte và chúng mất vài ngày để hoàn thành việc training. Hơn nữa, trên bộ dữ liệu này và bộ sưu tập Wikipedia lớn thứ hai (WIKI-50K và WIKI-100K), phương pháp đề xuất có tính cạnh tranh cao về Thời gian, Tổng bộ nhớ và cả hai biện pháp thực hiện tương ứng với tất cả các phương pháp tiếp cận khác. Những kết quả này cho thấy rằng phương pháp ít bị ảnh hưởng nhất bởi các phân bố long-tail class là (π; κ) -DS, chủ yếu





* Vì hai lý do: thứ nhất, lấy mẫu có xu hướng làm cho tập huấn luyện được cân bằng và thứ hai, tập dữ liệu nhị phân đã giảm chứa số lượng tương tự các ví dụ dương và âm. Do đó, đối với phương pháp đề xuất, có sự cải thiện cả về độ chính xác và các biện pháp của MaF1.
* Phương pháp PD-Sparse gần đây cũng có hiệu suất dự đoán cạnh tranh nhưng nó đòi hỏi phải lưu trữ các vectơ trọng số trung gian trong quá trình tối ưu hóa, điều này ngăn cản việc mở rộng quy mô tốt. Bộ giải PD-Sparse cung cấp một tùy chọn cho băm dẫn đến việc sử dụng bộ nhớ ít hơn trong quá trình training mà chúng tôi đã sử dụng trong các thử nghiệm; Tuy nhiên, việc sử dụng bộ nhớ vẫn cao đáng kể cho các tập dữ liệu lớn và đồng thời tùy chọn này làm chậm quá trình training đáng kể.
* Nói chung, trong số các phương pháp có hiệu suất dự đoán cạnh tranh, (π; κ) -DS dường như trình bày cách sử dụng bộ nhớ và thời gian chạy tốt nhất; thời gian chạy của nó thậm chí còn cạnh tranh với hầu hết các phương pháp dựa trên cây, dẫn đầu nó để cung cấp sự thỏa hiệp tốt nhất trong số các phương pháp so sánh trong ba lần, bộ nhớ và các biện pháp hiệu suất.

1. **Tổng kết**

* Chúng tôi đã trình bày một phương pháp mới để giảm vấn đề phân loại đa lớp thành phân loại nhị phân. Chúng tôi sử dụng tính năng tương tự đại hiện cho lớp và ví dụ và lược đồ ngẫu nhiên lấy mẫu kép cho quá trình giảm class. Mặc dù lược đồ lấy mẫu thay đổi sự phân phối của các lớp và việc giảm vấn đề ban đầu đối với một vấn đề phân loại nhị phân mang lại sự phụ thuộc lẫn nhau giữa các ví dụ dyadic; chúng tôi cung cấp các giới hạn lỗi tổng quát cho thấy nguyên tắc Giảm thiểu rủi ro thực nghiệm đối với việc chuyển đổi tập huấn được lấy mẫu vẫn còn phù hợp. Hơn nữa, các đặc tính của thuật toán đóng góp cho hiệu suất tuyệt vời của nó trong việc sử dụng bộ nhớ và tổng thời gian chạy và làm cho cách tiếp cận được đề xuất rất phù hợp với kịch bản lớp lớn.

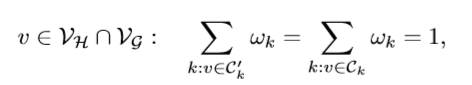
1. **Phần lý thuyết:**

**A.1 Kỹ thuật Lemmas**

**Bổ đề 1**. Phân số chẵn lẻ là đơn điệu trong bao hàm đồ thị: Nếu nghĩa là và , ta có .

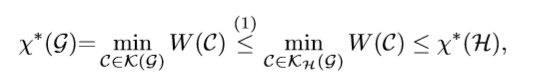
Bằng chứng, ta xem xét bất kỳ phân đoạn chính xác phù hợp của biểu đồ cho một số chỉ mục . Bằng cách loại bỏ mỗi đỉnh thuộc về và các cạnh ngẫu nhiên chúng ta có được sự bao phủ từ đồ thị . Khi cho một nhất định giữ ta loại bỏ nó khỏi , mà về cơ bản nó giống như gán .

Sự bao phủ là một phần phân đoạn chính xác của vì số lượng kết nối giữa các đỉnh trong là một tập hợp con của các giá trị trong với mọi . Sự bao phủ cũng chính xác vì với mọi :



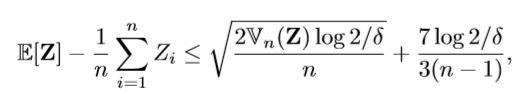
Trong đó: là một phần phân đoạn chính xác của đồ thị . Nghĩa là mỗi độ chính xác phân đoạn phù hợp của đồ thị có thể được chuyển đổi thành một phần chính xác phân đoạn phù hợp của đồ thị mà không làm tăng độ bao phủ . Chỉ rõ tập hợp tất cả các lớp phủ phân đoạn chính xác thích hợp của đồ thị dưới dạng và các lớp phủ thu được bằng cách cắt tỉa như trên thông qua .

Bằng cách xác định số chromatic phân số chúng ta có:



Trong đó: (1) đưa vào .

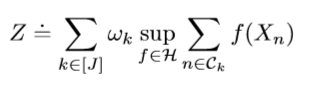
**Bổ đề 2** (Bất đẳng thức Bennet). Cho phép là biến với các giá trị và để . Sau đó, với xác suất ít nhất trong chúng ta có:



Trong đó: là phương sai mẫu.

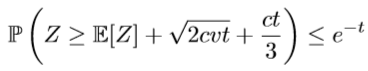
**Bổ đề 3** (Nồng độ các hàm phụ phụ phân số). Gọi là một tập hợp các hàm từ đến và giả sử các hàm trong có thể đo lường được, có thể tích phân và thõa mãn và . Giả sử rằng là một lớp của đồ thị phụ thuộc của và cho .

Ta định nghĩa:



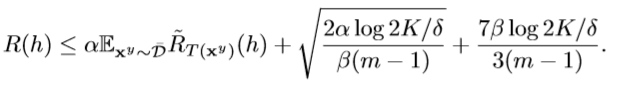
Gọi để 

Sau, với mọi



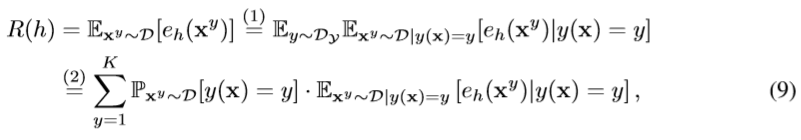
Cho dưới là một thước đo xác suất trên . Nó có thể phân tách thành một sản phầm trực tiếp của với phân phối biên trên và với điều kiện trên . Cho là thước đo được chuẩn hóa lại theo thuật toán, ví dụ: . Trong đó, .

**Bổ đề 4.** Cho là một tập dữ liệu của các ví dụ của được vẽ theo thang đo tính khả thi trên và bộ chuyển đổi thu được bởi hàm chuyển đổi được xác định trong phương trình 5. Cho là thước đo trên được sử dụng trong thuật toán . Với lớp của hàm và với mọi cho tất cả với xác suất ít nhất , ta có:

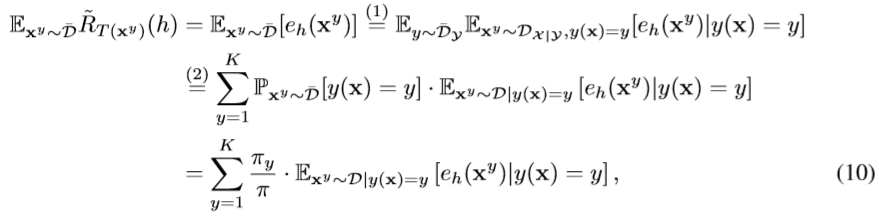


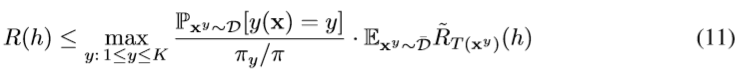
Giữ cho tất cả các , trong đó: là mất, và và là tỷ lệ lớp trong tập huấn luyện.

Bằng chứng. Đầu tiên, phân hủy rủi ro dự kiến R (h) như một tổng các rủi ro có điều kiện trong các lớp

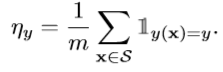


Trong đó (1) và (2) là do luật của tổng số kỳ vọng. Tương tự, xem xét sự mất mát dự kiến :

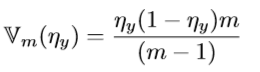


Trong đó (1) và (2) cũng là do luật của tổng số kỳ vọng. Từ (9) và (10) ta kết luận

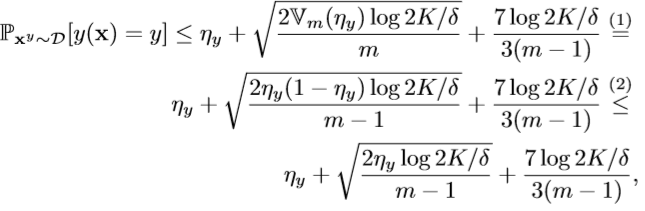
Cuối cùng, ta cần ràng buộc hệ số trước trong (11). Biểu thị qua xác suất thực nghiệm của lớp



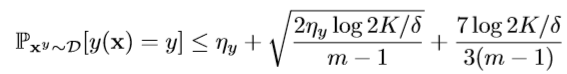
Lưu ý phương sai thực nghiệm phù hợp với bổ đề 2 là



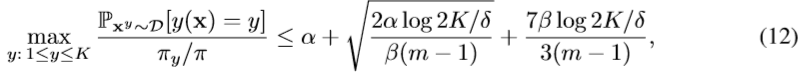
Với mọi ta có xác suất ít nhất theo bổ đề 2:



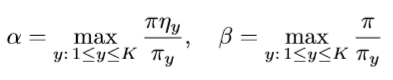
Trong đó (1) là sự thay thế của bởi giá trị rõ ràng của nó. (2) là do thực tế là . Sau đó, đồng thời cho tất cả chúng ta có xác suất ít nhất .



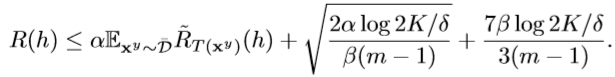
Như vậy với xác suất



Với



Từ phương trình (11) và (12) và thực tế , ta có xác suất ít nhất



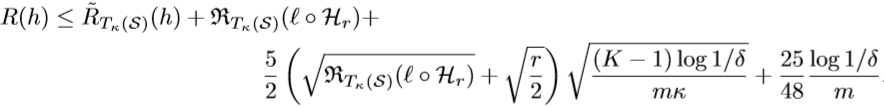
**A.2. Bằng chứng**

Kết quả của phần trước đòi hỏi bổ đề sau đây.

**Bổ đề 5.** Cho là tập dữ liệu của các ví dụ của được vẽ theo phân bố xác suất theo đường trên và bộ chuyển đổi thu được như trong phương trình (5) và vẽ mẫu đối nghịch bằng thuật toán . Vời lớp của hàm và , xét lớp được tham số hóa với được xác định là:



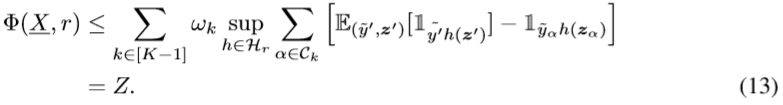
Trong đó, biểu thị phương sai. Sau đó, đối với bất kỳ và mất : , với xác suất ít nhất thì tổng quát hóa sau đây được giữ cho tất cả :



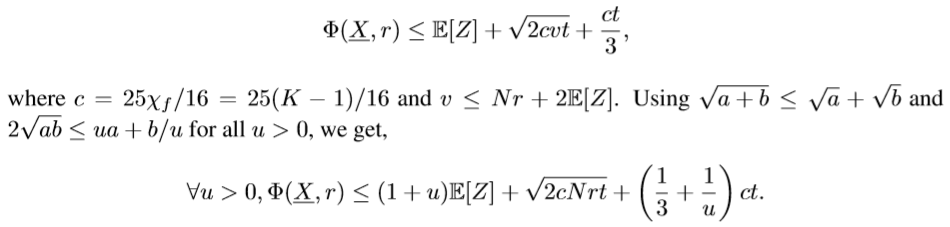
Bằng chứng, xem như hàm được định nghĩa như:

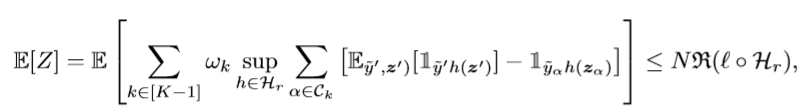


Trong đó, là bản sao của và trong đó ta sử dụng ký hiệu cho để làm rõ sự phụ thuộc vào chuỗi các biến phụ thuộc . Nó rất dễ dàng để thấy rằng:



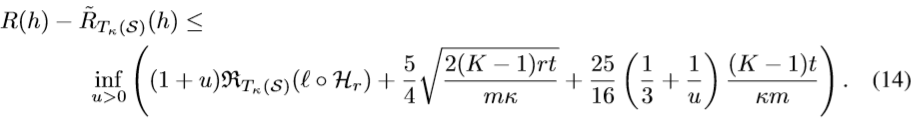
Bổ đề 3 dễ dàng áp dụng cho giới hạn trên bên phải của (13). Do đó, đối với giữ xác suất ít nhất :



Hơn nữa, với một đối số đối xứng đơn giản, ta có: 

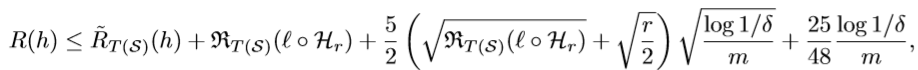
Với cho tất cả từ số chromatic phân số của biểu đồ phụ thuộc tương ứng với mẫu bằng và viết tắt của lớp bao phủ được xác định bởi phương trình (4), với đơn vị trọng lượng .

Hơn nữa, như và số chromatic phân số của , với xác suất ít nhất , ta có cho tất cả

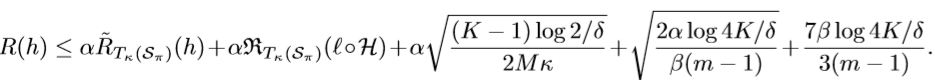


Sự tối thiểu của phía bên tay phải của bất đẳng thức đạt được đối với , gắn vào bộ giảm tối thiểu và giải quyết cho cho kết quả

Bằng chứng về định lý 1. Định lý 1 của [16] nói rằng số chromatic phân số bị giới hạn từ trên bởi . Sau đó bằng bổ đề 5 ta có xác suất ít nhất



Định lý 2(a). Cho là tập dữ liệu của các ví dụ được vẽ theo một thước đo xác suất trên và bộ chuyển đổi thu được bởi hàm chuyển đổi được xác định trong phương trình 5. Cho và là một tập huấn luyện bắt nguồn từ và tương ứng bằng cách sử dụng thuật toán với các tham số và . Với lớp của hàm và , ta có ràng buộc về rủi ro dự kiến của classification:

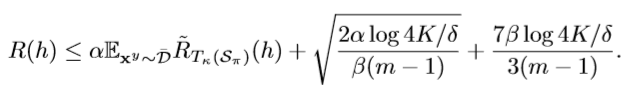


Giữ với xác suất ít nhất là đối với mọi cho tất cả , là mất và

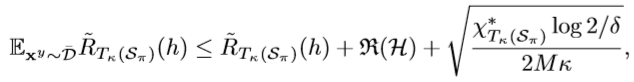


Và ηy là xác suất thực nghiệm dương tính của lớp y trên S.

Bằng chứng, bởi bổ đề 4 ta có cho với xác suất ít nhất



Theo định lý 4, ta có xác suất ít nhất



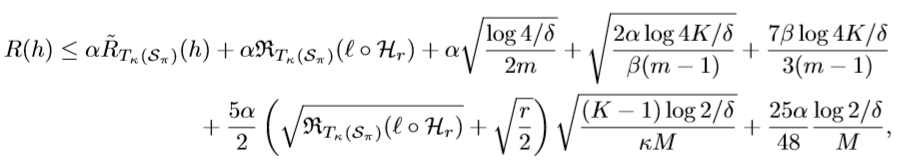
Trong đó có đồ thị phụ thuộc cho mẫu phụ là biểu đồ con của biểu đồ phụ thuộc cho toàn bộ mẫu .

Sau đó bởi bổ đề 1 ta có , cuối cùng do định lý 1. Trong đó và đại diện cho cho số chromatic phân số của biểu đồ phụ thuộc cho độ phản xạ và . Tập hợp hai phương trình cuối cùng ta chứng minh được định lý.

Định lý 2 (b). Cho là tập dữ liệu của các ví dụ được vẽ theo một thước đo xác suất trên và bộ chuyển đổi thu được bởi hàm chuyển đổi được xác định trong phương trình 5. Cho và là một tập huấn luyện bắt nguồn từ và tương ứng bằng cách sử dụng thuật toán với các tham số và . Với lớp của hàm và , xem như họ được tham số hóa, trong đó, đối với được xác định là:

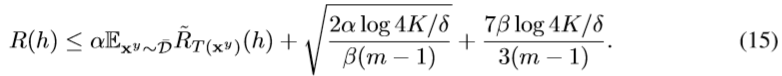


Trong đó: biểu thị phương sai. Sau đó, với mọi nào có xác suất ít nhất thì tổng quát hóa sau đây được giữ cho tất cả

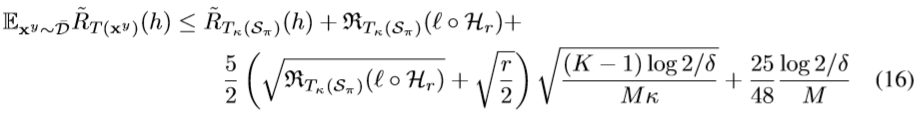


Trong đó: : , là mất và một chuỗi biến độc lập Rademacher sao cho và và và là xác suất thực nghiệm của lớp trên .

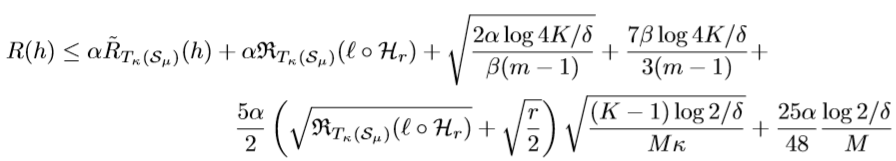
Bằng chứng về định lý cơ bản kết hợp các kết quả của định lý 1 và bổ đề 4. Bổ đề 4 ra có xác suất ít nhất là :



Bổ đề 2 (a) áp dụng cho xác suất ít nhất là :



Thay thế (15) trong (16) ta được:



Bằng chứng về định lý 2. Tuyên bố của định lý 2 trong bài báo về cơ bản là một sự kết hợp của các phát biểu của định lý 2 (a) và định lý 2 (b) đã được chứng minh ở trên.

**B. Phần thực nghiệm:**

Bảng 4 tóm tắt các thông số được điều chỉnh cho từng phương pháp. Trong các thí nghiệm của chúng ta đã sử dụng các bộ giải được cung cấp cho mỗi phương pháp cát điều chỉnh các thông số quan trọng nhất. Các giá trị được hiển thị trong Bảng 4 là các giá trị được sử dụng trong các kết quả thử nghiệm được báo cáo trong bài báo, dẫn đến hiệu suất dự đoán tốt nhất trong tập dữ liệu được lưu giữ.

