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.