ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

ĐẶNG ĐÌNH TUYẾN

PHÂN LỚP VĂN BẢN NHỜ MÁY VÉC – TƠ HỖ TRỢ VỚI HÀM STRING KERNEL

Chuyên ngành: Khoa học máy tính

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Người hướng dẫn khoa học: PGS.TS.Nguyễn Tân Ân

THÁI NGUYÊN - 2016

LỜI CẢM ƠN

Luận văn được hoàn thành tại trường Đại học Công nghệ Thông tin và Truyền thông Thái Nguyên.

Tác giả luận văn xin bày tỏ lòng biết ơn sâu sắc tới thầy hướng dẫn khoa học: PGS.TS Nguyễn Tân Ân đã tận tình hướng dẫn, giúp đỡ và tạo mọi điều kiện để tác giả thực hiện luận văn này. Tác giả cũng xin chân thành cảm ơn tập thể các thầy cô giáo trong khoa CNTT, phòng quản lý sau đại học Trường Đại học Công nghệ Thông tin và Truyên thông Thái Nguyên đã tạo mọi điều kiện giúp đỡ cho tác giả nghiên cứu, học tập và hoàn thành luận văn.

Xin cảm ơn gia đình, bạn bè, đồng nghiệp đã tạo điều kiện thuận lợi về tinh thần và vật chất cho tác giả hoàn thành luận văn này. Xin cảm ơn tất cả!

Thái Nguyên, tháng 6 năm 2016

Tác giả luận văn

Đặng Đình Tuyến

LỜI CAM ĐOAN

Tôi là Đặng Đình Tuyến, học viên cao học K13, chuyên ngành Khoa học máy tính, khoá 2014-2016. Tôi xin cam đoan luận văn thạc sĩ "Phân lớp văn bản nhờ Máy Véc-tơ hỗ trợ (SVM) với hàm string kernel" là công trình nghiên cứu của riêng tôi cùng với sự hướng dẫn của PGS.TS Nguyễn Tân Ân. Các số liệu, kết quả nêu trong Luận văn là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác.

Tất cả những tham khảo từ các nghiên cứu liên quan đều được nêu nguồn gốc một cách rõ ràng từ danh mục tài liệu tham khảo của luận văn. Trong luận văn, không có việc sao chép tài liệu, công trình nghiên cứu của người khác mà không chỉ rõ về tài liệu tham khảo.

Thái Nguyên, tháng 6 năm 2016

Tác giả

Đặng Đình Tuyến

MỤC LỤC

DANH MỤC HÌNH ẢNH

DANH MỤC BẢNG BIỂU

CHƯƠNG 1: BÀI TOÁN PHÂN LỚP

1.1. Nội dung bài toán phân lớp

Phân lớp (classification) là một tiến trình xử lý nhằm xếp các mẫu dữ liệu hay các đối tượng vào một trong các lớp đã được định nghĩa trước. Các mẫu dữ liệu hay các đối tượng được xếp về các lớp dựa vào giá trị của các thuộc tính (attributes) cho một mẫu dữ liệu hay đối tượng. Sau khi đã xếp tất cả các đối tượng đã biết trước vào các lớp tương ứng, lúc này mỗi lớp được đặc trưng bởi tập các thuộc tính của các đối tượng chứa trong lớp đó. Ví dụ: phân lớp văn bản, tế bào để xác định tế bào ung thư.

Phân lớp còn được gọi là phân lớp có giám sát (supervised classification), là một trong những lĩnh vực phổ biến nhất của học máy (machine learning) và khai thác dữ liệu (data mining). Nó giải quyết việc xác định những quy tắc giữa số lượng biến số độc lập và kết quả đạt được hay một biến số xác định phụ thuộc trong tập dữ liệu được đưa ra. Tổng quát, đưa ra một tập mẫu học $\left(x_{i_1}, x_{i_2}, ..., x_{i_k}, y_i\right)$, i=1,….,N, nhiệm vụ là phải ước lượng được một bộ phân lớp hay một mô hình xấp xỉ một hàm y = f(x) chưa biết mà phân lớp chính xác cho bất kỳ mẫu nào thuộc tập các mẫu học. Có nhiều cách để biểu diễn một mô hình phân lớp và có rất nhiều thuật toán giải quyết nó. Các thuật toán phân lớp tiêu biểu bao gồm như mạng neural, cây quyết định, suy luận quy nạp, mạng Beyesian, Support Vector Machine…. Tất cả các cách tiếp cập này xây dựng những mô hình đều có khả năng phân lớp cho một mẫu mới chưa biết dựa vào những mẫu tương tự đã được học.

Bài toán phân lớp có thể xử lý thông tin được thu thập từ mọi lĩnh vực hoạt động của con người và thế giới tự nhiên được biểu diễn dưới dạng các bảng. Bảng này bao gồm các đối tượng và các thuộc tính. Các phần tử trong bảng là các giá trị xác định các thuộc tính (attributes hay features) của các đối tượng. Trong đó số cột chính là số thuộc tính của các đối tượng, mỗi cột là một thuộc tính và số dòng chính là số đối tượng chứa trong dữ liệu này. Mọi dữ liệu được biểu diễn dưới các dạng khác có thể được chuyển thành dạng bảng như trên để thực hiện quá trình phân lớp.

1.2. Các phương pháp phân lớp

1.2.1. Phương pháp Naïve Bayes (NB)

NB là phương pháp phân loại dựa vào xác suất được sử dụng rộng rãi trong lĩnh vực máy học (Mitchell trình bày năm 1996, Joachims trình bày năm 1997 và Jason năm 2001) được sử dụng lần đầu tiên trong lĩnh vực phân loại bởi Maron vào năm 1961, sau đó trở nên phổ biến dùng trong nhiều lĩnh vực như trong các công cụ tìm kiếm (được mô tả năm 1970 bởi Rijsbergen), các bộ lọc mail (mô tả năm 1998 bởi Sahami)...

* Ý tưởng

Ý tưởng cơ bản của cách tiếp cận Naïve Bayes là sử dụng xác suất có điều kiện giữa từ và chủ đề để dự đoán xác suất chủ đề của một văn bản cần phân loại. Điểm quan trọng của phương pháp này chính là ở chỗ giả định rằng sự xuất hiện của tất cả các từ trong văn bản đều độc lập với nhau. Với giả định này NB không sử dụng sự phụ thuộc của nhiều từ vào một chủ đề, không sử dụng việc kết hợp các từ để đưa ra phán đoán chủ đề và do đó việc tính toán NB chạy nhanh hơn các phương pháp khác với độ phức tạp theo hàm số mũ.

* Công thức

Mục đích chính là tính được xác suất *Pr(Cj,d')*, xác suất để văn bản *d'* nằm trong lớp *Cj*. Theo luật Bayes, văn bản *d'* sẽ được gán vào lớp *Cj* nào có xác suất *Pr(Cj, d')* cao nhất. Công thức sau dùng để tính *Pr(Cj,d')* (do Joachims đề xuất năm 1997)

$$H_{BAYES} = \arg\max_{C_j^{'} \in C} \left(\frac{\Pr(C_j).\prod_{i=1}^{d^{'}} \Pr(w_i \mid C_j)}{\sum_{C^{'} \in C} \Pr(C^{'}).\prod_{i=1}^{d^{'}} \Pr(w_i \mid C^{'})} \right) = \arg\max_{C_j \in C} \left(\frac{\Pr(C_j).\prod_{i=1}^{d^{'}} \Pr(w_i \mid C_j)^{IF(w,d^{'})}}{\sum_{C^{'} \in C} \Pr(C^{'}).\prod_{i=1}^{d^{'}} \Pr(w_i \mid C^{'})^{IF(w,d^{'})}} \right)$$

Với:

- *(TF,d')* là số lần xuất hiện của từ w_i trong văn bản *d'*
- |*d'*| là số lượng các từ trong văn bản d'
- w_i là một từ trong không gian đặc trưng F với số chiều là |F|

Pr(Cj) được tính dựa trên tỷ lệ phần trăm của số văn bản mỗi lớp

$\Pr(C_j) = \frac{\| C_j \|}{\| C \|} = \frac{\| C_j \|}{\sum_{C' \in C} \| C' \|}$ tương ứng với tập dữ liệu huấn luyện.

- Pr(w_i|C_j) được tính sử dụng phép ước lượng Laplace (do Naplik trình bày năm 1982)

$$\Pr(w_i \mid C_j) = \frac{1 + TF(w_i, C_j)}{|F| + \sum_{w' \in |F|} TF(w', C_j)}$$

Ngoài ra còn có các phương pháp NB khác có thể kể ra như sau ML Naive Bayes, MAP Naive Bayes, Expected Naive Bayes, Bayesian Naive Bayes (Jason mô tả năm 2001). Naive Bayes là một công cụ rất hiệu quả trong một số trường hợp. Kết quả có thể rất tồi nếu dữ liệu huấn luyện nghèo nàn và các tham số dự đoán (như không gian đặc trưng) có chất lượng kém. Nhìn chung đây là một thuật toán phân loại tuyến tính thích hợp trong phân loại văn bản nhiều chủ đề. NB có ưu điểm là cài đặt đơn giản, tốc độ nhanh, dễ dàng cập nhật dữ liệu huấn luyện mới và có tính độc lập cao với tập huấn luyện, có thể sử dụng kết hợp nhiều tập huấn luyện khác nhau. Tuy nhiên NB ngoài giả định tính độc lập giữa các từ còn phải cần đến một ngưỡng tối ưu để cho kết quả khả quan. Nhằm mục đích cải thiện hiệu năng của NB, các phương pháp như multiclass- boosting, ECOC (do Berger trình bày năm 1999 và Ghani mô tả lại năm 2000) có thể được dùng kết hợp.

1.2.2. Phương pháp K–Nearest Neighbor (kNN)

Đây là phương pháp truyền thống khá nổi tiếng về hướng tiếp cận dựa trên thống kê đã được nghiên cứu trong nhận dạng mẫu hơn bốn thập kỷ qua (theo tài liệu của Dasarathy năm 1991). kNN được đánh giá là một trong những phương pháp tốt nhất (áp dụng trên tập dữ liệu Reuters phiên bản 21450), được sử dụng từ những thời kỳ đầu của việc phân loại văn bản (được trình bày bởi Marsand năm 1992, Yang năm 1994, Iwayama năm 1995)