Xây dựng mô hình Machine Learning dự báo cháy rừng ở các tỉnh Tây Nguyên dựa vào dữ liệu lịch sử thời tiết.

Nguyễn Đại Kỳ 19521731
Văn Viết Hiếu Anh 19521225
Lê Văn Phước

Ngày 1 tháng 8 năm 2021

Môn học:	CS114 - Máy học
Giảng viên hướng dẫn:	Lê Đình Duy
	Phạm Nguyễn Trường An

Mục lục

1 Tổng quan **2**
1.1 Mô tả bài toán . . . 2
1.2 Mô tả dữ liệu . . . 2
1.2.1 Weather Data . . . 2
1.2.2 Fire Data . . . 3

2 Introduction **3**

3 Xây dựng bộ dữ liệu **3**
3.1 Imputation of Data . . . 4
3.2 Creation of Dataset . . . 4

4 Methods **4**
4.1 Convolutional neural network . . . 4
4.2 Fully-connected neural network . . . 4

5 Conclusion **4**

1 Tổng quan

Bài viết là về quá trình thực nghiệm nghiên cứu các model Machine Learning với mục đích chọn ra mô hình tối ưu để dự đoán mức độ cháy rừng dựa vào dữ liệu thời tiết trong lịch sử của từng địa phương. Với mục đích hổ trợ trong việc dự đoán để phục vụ trong công tác phòng chống cháy rừng ở nước ta. Vì bài viết là ghi chép của quá trình thực nghiệm nên sẽ có nhiều phương pháp được đưa ra sử dụng.

1.1 Mô tả bài toán

1.2 Mô tả dữ liệu

Cả 3 nguồn dữ liệu gồm 2 nguồn dữ liệu thời tiết và nguồn dữ liệu về điểm cháy đều không có API cung cấp một cách đại chúng cho việc khai thác. Tuy nhiên vì các website này được xây dựng dựa trên cấu trúc Asynchronous (ASP.NET, Jquery hoặc ReactJS), nên sau khi phân tích và ghi lại các reqest, ta hoàn toàn có thể tìm được các cổng API và phương thức giao tiếp với server. Từ đó dùng vào việc khai thác dữ liệu tự động.

1.2.1 Weather Data

Có thể nói việc lấy dữ liệu thời tiết là công đoạn gây ra nhiều khó khăn nhất. Đa phần dữ liệu lịch sử thời tiết là rất lớn và các công ty hay tập đoàn công nghệ đều dùng để bán chứ không public trên website của họ. Ngay cả trên giao diện chính của weather.com của IBM cũng chỉ hiển thị dữ liệu thời tiết trong 2 năm trở lại (tức là 2021 và 2020). Tuy nhiên sau khi phân tích và tìm ra API và thử thay đổi các thông số, ta hoàn toàn có thể nhận được dữ liệu trong thời gian xa hơn.

weather.com Như đã nói ở trên, sau khi phân tích các request mà website này gửi về server của họ, chúng em đã tìm ra cổng API cung cấp dữ liệu thời tiết. Sau khi thay đổi các thông số, cổng API này chấp nhận cung cấp dữ liệu đến 1/1/2014.

Dữ liệu từ website này cung cấp có độ chính xác đến từng tọa độ, có nghĩa là chỉ cần cung cấp tọa độ (làm tròn đến 2 chữ số thập phân) thì server sẽ trả về thời tiết tại điểm đó tùy vào thời gian mà ta muốn. Tuy nhiên điểm yếu của dữ liệu này là chỉ cung cấp các đặc tính cơ bản nhất của thời tiết tại địa điểm đó gồm: nhiệt độ cao nhất và thấp nhất trong ngày, thời gian mặt trời mọc và lặn, lượng mưa (tích lũy trong 7 ngày, trong 1 tháng hoặc chỉ ngày hôm đó).

Sau khoảng nhiều ngày khai thác và xử lý, nhóm đã lấy được dữ liệu của 5 tỉnh Tây Nguyên vào từng xã từng ngày kéo dài từ 1/1/2014 đến 8/6/2021. Dữ liệu gồm các trường cơ bản sau:

- ward: Mã của xã, phường, thị trấn, thị xã. Vì sẽ có những địa điểm trùng tên nên nhóm sử dụng mã để phân biệt các địa phương (mã này cung cấp bởi API của firewatchvn.kiemlam.org.vn [1])
- date: là ngày mà record được ghi lại.
- max/min: là nhiệt độ cao nhất và thấp nhất được ghi nhận trong ngày(celcius).
- sunrise/sunset: là thời gian mặt trời mọc và lặn.
- 7_rain: lượng mưa tổng tính từ ngày chủ nhật gần nhất trước đó (cm)
- m_rain: lượng mưa tổng tính từ ngày 1 của tháng đó(cm)
- 24_rain: lượng mưa ghi nhận trong ngày(cm)

Dưới đây là một đoạn mẫu trong dữ liệu.

ward	date	max	min	sunrise	sunset	7_rain	m_rain	24_rain
24727.0	2016-10-17	30	25	05:36 AM	05:28 PM	5.94	30.91	1.82
24727.0	2016-10-18	30	25	05:36 AM	05:28 PM	7.54	32.66	1.75
24727.0	2016-10-19	30	25	05:36 AM	05:27 PM	7.54	32.96	0.3
24761.0	2019-10-14	32	24	05:36 AM	05:30 PM	0.02	0.07	0.0
24761.0	2019-10-15	32	25	05:36 AM	05:30 PM	0.02	0.07	0.0
24761.0	2019-10-16	32	25	05:36 AM	05:29 PM	0.02	0.07	0.02
24761.0	2019-10-17	32	25	05:36 AM	05:29 PM	0.02	0.07	0.0
24761.0	2019-10-18	32	25	05:36 AM	05:28 PM	0.02	0.07	0.0
24761.0	2019-10-19	32	26	05:36 AM	05:28 PM	0.02	0.07	0.0
24761.0	2019-10-20	32	26	05:36 AM	05:27 PM	0.02	0.1	0.0
24761.0	2019-10-21	33	25	05:36 AM	05:27 PM	0.02	0.1	0.0
24761.0	2019-10-22	32	25	05:37 AM	05:26 PM	0.02	0.1	0.0
24761.0	2019-10-23	33	25	05:37 AM	05:26 PM	0.0	0.1	0.0
24761.0	2019-10-24	33	25	05:37 AM	05:25 PM	0.0	0.1	0.0

worldweatheronline.com

1.2.2 Fire Data

firewatchvn.kiemlam.org.vn

2 Introduction

Ở nước ta có 3 thảm họa lớn nhất, gây thiệt hại lớn hàng năm về cả người và của. Cùng với lũ lụt và hạn hán, cháy rừng là một thảm họa gây thiệt hại không chỉ về kinh tế mà còn cả con người và hệ sinh thái. Theo thống kê của Cục Kiểm lâm từ năm 1992 đến 2006, trung bình mỗi năm xảy ra 1254 vụ cháy rừng gây thiệt hại khoảng 6646 ha rừng, trong đó có 2854 ha là rừng tự nhiên và 3791 ha là rừng trồng. Bên cạnh việc nâng cao năng lực phòng cháy chữa cháy rừng (PCCCR) cho lực lượng kiểm lâm như đầu tư trang thiết bị, cơ sở vật chất, xây dựng cơ chế điều hành phối hợp và tuyên truyền nâng cao nhận thức trách nhiệm của chủ rừng và người dân, công tác cảnh báo nguy cơ cháy rừng cũng như tổ chức phát hiện sớm và thông báo kịp thời điểm cháy rừng là rất cần thiết.

Từ đầu năm 2007, Cục Kiểm lâm (Bộ Nông nghiệp và Phát triển Nông thôn) đã lắp đặt và vận hành trạm thu ảnh viễn thám MODIS tại Hà Nội với mục đích chính là phát hiện sớm các điểm cháy rừng (hotspots) trên toàn lãnh thổ Việt Nam. Hệ thống trạm thu của TeraScan đã tự động thu nhận, xử lý và sao lưu dữ liệu ảnh MODIS hàng ngày từ 2 vệ tinh TERRA và AQUA với mô-đun Vulcan tự động xử lý và tạo ra dữ liệu các điểm cháy sử dụng thuật toán ATBD-MOD14.

Hệ thống này cung cấp dữ liệu về điểm cháy ghi nhận được từ vệ tinh và lưu lại thời gian và tọa độ cháy. Từ khi bắt đầu lắp đặt đến nay hệ thống dữ liệu cháy của cục kiểm lâm được ghi lại được gần 1 triệu điểm cháy. Nhờ lượng dữ liệu này việc xây dựng một hệ thống tự động phân tích mức độ cháy rừng dựa vào các đặc trưng cơ bản của dữ liệu khí tượng thủy văn là hoàn toàn có cơ sở và khả quan.

3 Xây dựng bộ dữ liệu

Để xây dựng dược mô hình, việc đầu tiên sau khi khai thác dữ liệu là phải thực hiện phân tích và làm sạch dữ liệu. Từ đó chọn ra những đặc trưng có ảnh hưởng nhiều đến output của bài toán để tạo ra được dataset phù hợp.

3.1 Imputation of Data

3.2 Creation of Dataset

4 Methods

4.1 Convolutional neural network

4.2 Fully-connected neural network

5 Conclusion

Tài liệu

[1] http://firewatchvn.kiemlam.org.vn/.