Tìm lỗi quan sát nhãn dữ liệu

Vo Chi Cong

8th July 2022

Mở bài

Có một bộ dữ liệu bảng số, được gắn nhãn để phân loại, ví dụ nhãn dương tính và âm tính. Giả sử các nhãn dương tính có độ tin cậy cao, còn các nhãn âm tính có độ tin cậy thấp hơn, có thể xem như chứa cả dữ liệu dương tính chưa bộc phát. Lấy ví dụ với dữ liệu đánh giá tín dụng thì những ca vỡ nợ sẽ có nhãn dương tính. Với dữ liệu khám nghiệm ung thư thì những ca đã phát bệnh là dương tính.

Giả sử phân bố nhãn trong bộ dữ liệu trên có tương quan đủ mạnh với phân bố nhãn tiềm ẩn thật sự. Trong phân bố nhãn có tiềm ẩn tính đa dạng, bất định mang tính bản chất. Ví dụ có 10 người có các thuộc tính xấu gần giống nhau, nhưng sẽ trong đó sẽ chỉ có 8 người ngẫu nhiên nào đó vỡ nợ, hoặc bị ung thư.

Bộ dữ liệu trên có một số lượng nhất định các nhãn bị gắn sai. Có thể nào xác định được ranh giới để phân biệt được lỗi gắn nhãn với tính bất định bản chất của dữ liệu hay không?

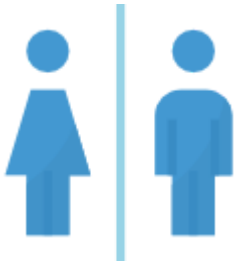
Giả sử rằng suất nhãn bị gán sai không phụ thuộc vào từng ca dữ liệu cụ thể mà chỉ phụ thuộc vào đặc tính của các lớp nhãn dữ liệu. Ví dụ trường hợp phân loại 3 loài vật là chó, mèo và chuột, thì xác suất nhầm chó với mèo cao hơn là nhầm mèo với chuột hoặc chó với chuột.

Có một mô hình dự đoán xác suất dương tính đối với bộ dữ liệu nêu trên. *Giả sử xác suất do mô hình đưa ra có tương quan đủ mạnh đối với phân bố thật sự của nhãn*.

Với những giả sử nêu trên, ta có thể ước lượng được xác suất nhãn gắn trên một ca dữ liệu là thật sự đúng hay không

Lỗi quan sát nhãn là gì?

Với một đối tượng khảo sát, quan sát viên sẽ quan sát, xem xét, nghiên cứu rồi gán một nhãn nhất định cho dữ liệu đó. Trong môi trường lý tưởng thì ta sẽ nhận định và gán đúng nhãn “chân lý" cho đối tượng.



Restroom icons created by Freepik — Flaticon

Ví dụ ta có thể quan sát ngực, bụng, mông của một người nào đó và nhận định giới tính. Trong thực tế thì có thể xảy ra nhầm lẫn ở một bước nào đó trong quá trình từ khi bắt đầu quan sát cho đến khi gắn xong nhãn. Nhầm lẫn đó có thể dẫn tới gán nhầm nhãn “Nam” cho đối tượng vốn là “Nữ”, hoặc ngược lại. Chúng ta gọi "lỗi quan sát nhãn" và "lỗi gắn nhãn" với cùng một ý nghĩa.

Có thể có một số nam giới và nữ giới có số đo 3 vòng khá giống nhau, nhưng “đương nhiên" họ có 2 giới tính khác nhau, tức là các nhãn “chân lý" của họ là khác nhau về bản chất, chứ không nhất thiết có liên quan đến việc gắn nhãn có lỗi hay không.

Nói cách khác từ số đo 3 vòng ta có thể không suy đoán được chắc chắn 100% nhưng có thể tính được xác suất giới tính Nam/Nữ của đối tượng. Quy tắc hay mô hình suy đoán có thể học được từ một tập dữ liệu có số đo 3 vòng và giới tính tương ứng của nhiều mẫu người khác nhau. Nếu trong tập dữ liệu này có những nhãn giới tính bị gắn sai thì việc học xác suất “chân lý” sẽ bị lệch lạc.

Định nghĩa và ký hiệu

Quy trình nhiễu theo lớp

Giả sử có một bộ dữ liệu số được gắn nhãn phân loại thành m lớp khác nhau $[m] \coloneqq 1, 2, \ldots, m$. Giả sử đối với mỗi mẫu dữ liệu ta có một nhãn "tiềm ẩn" thật là y^*. Trước khi quan sát được nhãn $\tilde{y}$, giả sử có một quy trình gây nhiễu biến $y^* = j$ thành $\tilde{y} = i$ với xác suất $p(\tilde{y} = i, y^* = j)$ chỉ phụ thuộc vào $i, j \in [m]$ và độc lập với các mẫu dữ liệu cụ thể,

$$p(\tilde{y}|y^*; \boldsymbol{x}) = p(\tilde{y}|y^*) \forall \boldsymbol{x}.$$

Ví dụ khi phân loại 3 loài vật là chó, mèo và chuột, thì xác suất nhầm chó với mèo cao hơn là nhầm mèo với chuột hoặc chó với chuột, và xác suất đó không phụ thuộc vào từng con cho, con mèo hoặc con chuột cụ thể. Giả sử này là hợp lý và thường được sử dụng trong các nghiên cứu về xử lý nhiễu (Goldberger and Ben-Reuven, 2017; Sukhbaatar et al., 2015).

Ma trận nhiễu theo lớp

$$\boldsymbol{Q}_{\tilde{y},y^*} \coloneqq \begin{bmatrix} p(\tilde{y}=1,y^*=1) & \dots & p(\tilde{y}=1,y^*=m) \\ \vdots & p(\tilde{y}=i,y^*=j) & \vdots \\ p(\tilde{y}=m,y^*=1) & \dots & p(\tilde{y}=m,y^*=m) \end{bmatrix}$$

là ma trận kích thước $m \times m$ thể hiện phân phối xác suất đồng thời cho $\tilde{y}$ và y^*.

Độ thưa là tỷ lệ số 0 chiếm lĩnh các vị trí ngoại trừ đường chéo của ma trận $\boldsymbol{Q}_{\tilde{y},y^*}$: độ thưa bằng 0 nói rằng mọi tỷ lệ nhiễu $p_{\tilde{y},y^*}$ đều khác 0, còn độ thưa 1 thể hiện tình trạng lý tưởng, hoàn toàn không có nhiễu trong nhãn.

Gọi $\boldsymbol{X}_{\tilde{y}=i}$ là tập hợp các mẫu $\boldsymbol{x}$ đã được gán nhãn $\tilde{y}=i$. **Độ tự tin** $\hat{p}(\tilde{y}=i;\boldsymbol{x}\in\boldsymbol{X}_{\tilde{y}=i},\boldsymbol{\theta})$ là xác suất mô hình $\boldsymbol{\theta}$ đưa ra đối với mẫu $\boldsymbol{x}$, dự đoán nó có label đúng như label $\tilde{y}$ đã được gán. *Độ tự tin thấp là một dấu hiệu của khả năng nhãn có lỗi.*

Phương pháp học tự tin

Confident Learning Method

Đầu vào

1. Các nhãn $\tilde{y}_k$ đã quan sát được đối với các mẫu $\boldsymbol{x}_k \in \boldsymbol{X}$
2. Xác suất $\hat{p}(\tilde{y}=i;\boldsymbol{x}_k\in\boldsymbol{X})$ dự đoán mẫu $\boldsymbol{x}_k \in \boldsymbol{X}$ có nhãn $i \in [m]$

Các bước

1. Tính t_i, độ tự tin trung bình trong từng lớp $i \in [m]$
2. Ước lượng phân bố xác suất đồng thời $\hat{\boldsymbol{Q}}_{\tilde{y},y^*}$ cho nhãn quan sát và nhãn thật
3. *Lọc và xếp hạng các mẫu theo mức độ khả nghi nhãn bị lỗi*
4. Loại bỏ các mẫu khả nghi nhất là nhãn bị lỗi
5. Đặt trọng số cho các mẫu trong từng lớp $i \in [m]$ để học lại mô hình $\boldsymbol{\theta}$

Chỉ tiêu tự tin

Với mỗi lớp $i \in [m]$ ta có thể chọn một chỉ tiêu tự tin $t_j \in (0,1)$. Chỉ tiêu này sẽ được dùng để thiết lập ma trận tự tin (1). Một cách chọn chỉ tiêu tự tin là dùng độ tự tin trung bình (2). Đối với từng mẫu $\boldsymbol{x}$ và từng nhãn i, giá trị xác suất dự đoán $\hat{p}(\tilde{y}=i;\boldsymbol{x},\boldsymbol{\theta})$ đưa ra bởi mô hình $\boldsymbol{\theta}$, nếu không nhỏ chỉ tiêu t_i thì ta cho rằng có khả năng nhãn i đúng cho mẫu $\boldsymbol{x}$. Tập hợp các nhãn i's có thể đúng với mẫu $\boldsymbol{x}$ là

$$\{l \in [m] : \hat{p}(\tilde{y}=l;\boldsymbol{x},\boldsymbol{\theta}) \geq t_l\} \neq \emptyset;$$

Trong tập đó thì nhãn j có xác suất dự đoán lớn nhất có vẻ là lớp "thật" của $\boldsymbol{x}$.

Ma trận đếm cặp nhãn

Gọi $\boldsymbol{X}_{\tilde{y}=i,y^*=j}$ là tập (không tường minh) các mẫu có nhãn quan sát là i và nhãn thật là j, ta ước lượng nó như sau bằng cách sử dụng các chỉ tiêu tự tin t_j cho từng lớp $j \in [m]$:

$$\hat{\boldsymbol{X}}_{\tilde{y}=i,y^*=j} \coloneqq \left\{ \boldsymbol{x} \in \boldsymbol{X}_{\tilde{y}=i} : \underset{l \in [m]: \hat{p}(\tilde{y}=l;\boldsymbol{x},\boldsymbol{\theta}) \geq t_l}{\arg\max} \hat{p}(\tilde{y}=l;\boldsymbol{x},\boldsymbol{\theta}) \equiv j \right\} \tag{1}$$

Ma trận đếm cặp nhãn $\hat{\boldsymbol{C}}_{\tilde{y},y^*}$ kích thước $m \times m$ lưu số phần tử của các tập $\hat{\boldsymbol{X}}_{\tilde{y}=i,y^*=j}$,

$$\hat{\boldsymbol{C}}_{\tilde{y}=i,y^*=j} \coloneqq |\hat{\boldsymbol{X}}_{\tilde{y}=i,y^*=j}|.$$

Số lượng mẫu được quan sát có nhãn $\tilde{y}=i$ là $\boldsymbol{C}_{\tilde{y}=i} \coloneqq |\boldsymbol{X}_{\tilde{y}=i}|$. Vì (1) chỉ là ước lượng của $\boldsymbol{X}_{\tilde{y}=i,y^*=j}$ cho nên nếu đặt $\hat{\boldsymbol{C}}_{\tilde{y}=i} \coloneqq \sum_{j \in [m]} \hat{\boldsymbol{C}}_{\tilde{y}=i,y^*=j}$ thì có thể $\hat{\boldsymbol{C}}_{\tilde{y}=i} \neq \boldsymbol{C}_{\tilde{y}=i}$.

Ví dụ $\hat{\boldsymbol{C}}_{\tilde{y}=3,y^*=1} = 10$ có nghĩa là, ta đếm được 10 mẫu được gán nhãn 3 nhưng "thật ra" nên có nhãn 1.

Độ tự tin trung bình

Độ tự tin trung bình trong lớp $i \in [m]$ là

$$t_i = \frac{1}{\hat{\boldsymbol{C}}_{\tilde{y}=i}} \sum_{\boldsymbol{x} \in \boldsymbol{X}_{\tilde{y}=i}} \hat{p}(\tilde{y}=i;\boldsymbol{x},\boldsymbol{\theta}) \tag{2}$$

Ước lượng ma trận nhiễu

Hiệu chỉnh ma trận đếm cặp nhãn qua hai bước

1. Hiệu chỉnh từng dòng theo số mẫu của từng lớp đã quan sát $i \in [m]$

$$\check{Q}_{\tilde{y}=i,y^*=j} = \hat{\boldsymbol{C}}_{\tilde{y}=i,y^*=j} \frac{\boldsymbol{C}_{\tilde{y}=i}}{\hat{\boldsymbol{C}}_{\tilde{y}=i}} \tag{3a}$$

2. Chia đều toàn bộ để tổng số các yếu tố trở thành 1

$$\hat{Q}_{\tilde{y}=i,y^*=j} = \frac{\check{Q}_{\tilde{y}=i,y^*=j}}{\sum_{i \in [m], j \in [m]} \check{Q}_{\tilde{y}=i,y^*=j}} \tag{3b}$$

Lọc và xếp hạng nhãn lỗi

Với phương pháp đơn giản nhất, các mẫu $\left\{ \boldsymbol{x} \in \hat{\boldsymbol{X}}_{\tilde{y}=i,y^*=j} : i \neq j \right\}$ nằm ngoài đường chéo của ma trận $\hat{\boldsymbol{X}}_{\tilde{y},y^*}$ bị tình nghi là có nhãn lỗi. Các mẫu đó được xếp hạng mức độ khả nghi theo dựa theo xác suất do mô hình $\boldsymbol{\theta}$ dự đoán

$$\hat{e}(\boldsymbol{x} \in \boldsymbol{X}_{\tilde{y}=i}, \boldsymbol{\theta}) \coloneqq \max_{j \neq i} \hat{p}(\tilde{y}=j;\boldsymbol{x},\boldsymbol{\theta}) - \hat{p}(\tilde{y}=i;\boldsymbol{x},\boldsymbol{\theta})$$

theo cách làm trong CleanLab của Curtis et al.'s (2021), và đảo dấu so với Wei et al.'s (2018).

Curtis et al.'s (2021) trình bày một số phương pháp khác dùng ma trận nhiễu $(3b)$ để chọn lọc và xếp hạng nhãn khả nghi có lỗi.

Tương lai

Một số hướng nghiên cứu tương lai

- Tối ưu hóa giá trị chỉ tiêu tự tin
- Xử lý với bài toán hồi quy
- Tương tác qua lại giữa việc học mô hình và việc khử lỗi

Tham khảo

- Curtis G. Northcutt and Lu Jiang and Isaac L. Chuang (2021). Confident Learning: Estimating Uncertainty in Dataset Labels. Journal of Artificial Intelligence Research (JAIR)
- An Introduction to Confident Learning: Finding and Learning with Label Errors in Datasets (curtisnorthcutt.com)
- cleanlab/cleanlab: The standard data-centric AI package for data quality and machine learning with messy, real-world data and labels. (github.com)
- Are Label Errors Imperative? Is Confident Learning Useful? | by Suneeta Mall | May, 2022 | Towards Data Science (medium.com)
- Wei, C., Lee, J. D., Liu, Q., and Ma, T. (2018). On the margin theory of feedforward neural networks. Computing Research Repository (CoRR)

Published from denoise.jmd using Weave.jl v0.10.9 on 2022-07-08.