Tạp chí Khoa học Trường Đại học Cần Thơ
website: sj.ctu.edu.vn

NÂNG CAO ĐỘ CHÍNH XÁC PHÂN LOẠI LỚP ÍT MẪU TỪ TẬP DỮ LIỆU MẤT CÂN BẰNG

Bùi Minh Quân[1], Phạm Xuân Hiền[1] và Huỳnh Xuân Hiệp[1]
[1] Khoa Công nghệ Thông tin & Truyền thông, Trường Đại học Cần Thơ

Thông tin chung:
Ngày nhận: 03/09/2013
Ngày chấp nhận: 21/10/2013

Title:
Improving prediction of the minority class in an imbalance dataset

Từ khóa:
Học với chi phí nhạy cảm, tập đa lớp, dữ liệu mất cân bằng

Keywords:
Cost-sensitive learning, multi-class, imbalanced data

ABSTRACT

A dataset is called imbalance if it has some classes containing more instances than others. In this case, accurately classifying samples in small classes is very difficult. The higher the imbalanced ratio, the more difficult getting a good solution. Cost-sensitive learning is an effective solution for the imbalanced problem. In this paper, we present a decision system with misclassification cost. The system improves the degree of precision in the minor classes which are interested in imbalanced dataset. The system is based on the study of methods of classifying on the imbalanced dataset by cost-sensitive. This system is applied in medical diagnostic. The experimental results show that the accuracy of the diagnostic system is improved.

TÓM TẮT

Vấn đề mất cân bằng dữ liệu xảy ra khi trong tập dữ liệu có lớp chứa số mẫu nhiều hơn các lớp khác. Phân loại chính xác cho mẫu thuộc lớp nhỏ trong tập mất cân bằng là khó khăn. Khi tỷ lệ mất cân bằng của tập dữ liệu càng cao thì việc phát hiện được mẫu của lớp nhỏ càng khó. Học với chi phí nhạy cảm là giải pháp hiệu quả để giải quyết vấn đề mất cân bằng. Trong bài báo này, chúng tôi trình bày một hệ thống gọi là hệ thống quyết định với chi phí, hệ thống giúp cải thiện khả năng phân loại chính xác của lớp nhỏ trong tập dữ liệu mất cân bằng, lớp dữ liệu rất được quan tâm. Hệ thống được xây dựng dựa vào kết quả nghiên cứu giải pháp phân loại trên dữ liệu mất cân bằng tiếp cận với chi phí nhạy cảm. Hệ thống được áp dụng vào lĩnh vực chẩn đoán y học, kết quả thực nghiệm cho thấy khả năng phát hiện chính xác bệnh nhân của hệ thống chẩn đoán được cải thiện.

1 GIỚI THIỆU

Dữ liệu thu được trong các ứng dụng thực tế thường là các tập dữ liệu mất cân bằng. Tập mất cân bằng thường xuất hiện trong các lĩnh vực như chẩn đoán y tế [24], giám sát hệ thống mạng, phát hiện xâm nhập hệ thống [10]... Thông thường trong những lĩnh vực này lớp cần quan tâm lại có rất ít mẫu (lớp nhỏ) [1][9] so với các lớp khác trong tập dữ liệu. Lớp bệnh nhân có rất ít mẫu so với các lớp khác trong ứng dụng y học, giao dịch tấn công có rất ít mẫu so với các lớp giao dịch khác của hệ thống mạng. Việc chẩn đoán đúng nhãn của mẫu thuộc lớp nhỏ là cần thiết và quan trọng. Nếu mẫu thuộc lớp nhỏ chẩn đoán nhãn sai thì giá phải trả là cao hơn nhiều so với chẩn đoán sai nhãn cho mẫu thuộc lớp lớn [5][11]. Các mẫu thuộc các lớp khác nhau khi chẩn đoán sai thì giá phải trả cũng không giống nhau. Chúng tôi gọi đây là bài toán phân loại (chẩn đoán nhãn) có đặc điểm là lớp có số mẫu nhỏ gắn với chi phí chẩn đoán sai lớn và lớp có số mẫu lớn gắn với chi phí chẩn đoán sai nhỏ.

Tỷ lệ mất cân bằng của tập dữ liệu ảnh hưởng rất lớn đến kết quả gán nhãn của lớp ít mẫu. Tỷ lệ mất cân bằng được xem là tỷ lệ so sánh sự phân phối mẫu của các lớp trong tập dữ liệu với lớp có số mẫu nhỏ nhất [6]. Các giải thuật phân loại truyền thống luôn cố gắng cực đại hóa độ chính xác. Các giải thuật này có xu hướng gán nhãn cho mẫu chưa xác định là thuộc lớp chiếm số mẫu lớn và bỏ qua lớp nhỏ [4]. Nếu dựa vào nguyên tắc số đông gán nhãn cho mẫu trong tập mất cân bằng thì độ chính xác khi phân loại trên tập dữ liệu dễ dàng đạt tới ≈ 99% trong khi độ chính xác của lớp nhỏ là ≈ 0%.

Để cải thiện độ chính xác phân loại trên lớp ít mẫu và giữ được độ chính xác trên toàn tập ở mức chấp nhận, bài báo đề xuất xây dựng hệ thống quyết định với chi phí (Decision System with Misclassification Cost). Hệ thống được xây dựng dựa trên nền tảng nghiên cứu thành tựu của phương pháp học với chi phí nhạy cảm trên tập dữ liệu mất cân bằng [1][7][15][26]. Hệ thống cho phép mở rộng áp dụng vào nhiều lĩnh vực giải quyết vấn đề mất cân bằng của ứng dụng cụ thể.

Nội dung phần 2, chúng tôi sẽ trình bày các nghiên cứu liên quan và lý thuyết xây dựng hệ thống quyết định với chi phí. Tiếp theo trong phần 3, chúng tôi mô tả các chức năng trong hệ thống quyết định với chi phí. Kết quả thực nghiệm sẽ được trình bày trong phần 4, kết luận trong phần 5.

2 CÁC NGHIÊN CỨU LIÊN QUAN

Trong những năm gần đây, học với chi phí đã thu hút được nhiều quan tâm của máy học và cộng đồng khai mỏ dữ liệu [26]. Đã có nhiều nghiên cứu học với chi phí nhạy cảm được thực hiện. Những nghiên cứu học tiếp cận với chi phí chia làm hai dạng. Dạng thứ nhất, chi phí gắn theo mẫu, mỗi mẫu có một chi phí đính kèm [1]. Dạng thứ hai là chi phí theo lớp, mỗi lớp có một chi phí, các mẫu trong cùng lớp có cùng chi phí [4][8][9]. Chi phí gắn với giải thuật học có thể chia làm nhiều loại: chi phí kiểm tra [14], chi phí huấn luyện [24], chi phí phân loại sai [26],... Trong phạm vi bài báo này chúng tôi nghiên cứu chi phí phân loại sai theo lớp.

Tiếp cận với chi phí là giải pháp hiệu quả để giải quyết vấn đề mất cân bằng [15]. Để giải quyết vấn đề mất cân bằng, điều chỉnh tỷ lệ là cách tiếp cận phổ biến của phương pháp học với chi phí. Tuy nhiên, điều chỉnh mất cân bằng theo phương pháp truyền thống chỉ mang lại hiệu quả cao khi áp dụng cho tập hai lớp [26]. Trên tập đa lớp có thể chia làm điều chỉnh trực tiếp (đồng thời) và điều chỉnh gián tiếp. Điều chỉnh đồng thời mất cân bằng trên tập đa lớp chỉ thực hiện khi tìm được bộ chi phí phân loại sai thích hợp giữa các lớp. Nếu không tìm được thì phải điều chỉnh gián tiếp tập đa lớp thông qua các tập con hai lớp [2][26]. Giải thuật học dùng chi phí như là một đại lượng (trọng số) điều chỉnh tỷ lệ mất cân bằng của tập. Trọng số được gán cho mẫu và đưa vào giải thuật học cây quyết định (C4.5) huấn luyện mô hình phân loại [20][21], mô hình sẽ quan tâm đến các mẫu có trọng số cao trong quá trình phân loại [21].

3 PHÂN LOẠI DỮ LIỆU TIẾP CẬN VỚI CHI PHÍ NHẠY CẢM

3.1 Điều chỉnh mất cân bằng tập hai lớp

Trong tập dữ liệu hai lớp, các giải thuật điều chỉnh mất cân bằng với chi phí thì mẫu được phân loại thuộc về lớp có tổng chi phí tổn thất là thấp nhất [8][9].

Ví dụ 1: với tập dữ liệu hai lớp, p là xác suất phân loại mẫu thuộc lớp 1 và 1-p là xác suất mẫu thuộc lớp 2. Các giá trị cm_{ij} của ma trận chi phí là chi phí phân loại sai mẫu thuộc lớp i vào lớp j (i≠j), $i, j \in \{1..2\}$, $cm_{ij}=0$ (với i=j) là chi phí phân loại đúng.

$$\mathrm{cm} = \begin{bmatrix} 0 & cm_{12} \\ cm_{21} & 0 \end{bmatrix}$$

Mẫu được gán vào lớp 1 khi và chỉ khi tổng chi phí phân loại mẫu vào lớp 1 nhỏ hơn tổng chi phí phân loại mẫu vào lớp 2 [26], công thức tính tổng chi phí như sau:

$$p \times cm_{11} + (1-p) \times \mathrm{cm}_{21} \le \mathrm{p} \times \mathrm{cm}_{12} + (1-\mathrm{p}) \times \mathrm{cm}_{22} \quad (1)$$

Từ công thức (1) cho thấy việc gán nhãn mẫu phụ thuộc vào hai đại lượng: xác suất và chi phí. Khi chi phí thay đổi nhãn gán cho mẫu sẽ thay đổi. Chứng tỏ rằng việc gán nhãn cho mẫu nhạy cảm với chi phí. Chi phí của lớp là đại lượng điều chỉnh mất cân bằng các lớp trong tập. Chi phí cao được gán cho các lớp ít mẫu, chi phí thấp được gán cho các lớp nhiều mẫu. Chi phí giúp cân bằng so với số mẫu. Trọng số gán cho mẫu trong tập huấn luyện được tính từ chi phí [21][26]. Trong quá trình phân loại, mô hình sẽ quan tâm đến các mẫu có trọng số cao, nghĩa là quan tâm đến các lớp ít mẫu. Nếu gọi cm_i là chi phí của lớp thứ i thì cm_i được tính như sau:

$$cm_i = \sum_{j=1}^{k} cm_{ij} \quad (2)$$

Với k=2 (k là số lớp của tập dữ liệu), thì $cm_1=cm_{12}$, $cm_2=cm_{21}$. Gọi w_i là trọng số gán mẫu thuộc lớp thứ i, w_i được tính như sau:

$$w_i = \frac{\left(num \times \mathrm{c}\, m_i\right)}{\sum_{j=1}^{k}\left(num_j \times \mathrm{c}\, m_j\right)} \quad (3)$$

Với num: số mẫu tập huấn luyện, num_j: số mẫu của lớp thứ j trong tập huấn luyện.

3.2 Điều chỉnh mất cân bằng trên tập đa lớp

Theo kết quả của nghiên cứu điều chỉnh mất cân bằng trên tập đa lớp đã chỉ rõ: phương pháp điều chỉnh truyền thống chỉ cho kết quả tốt trên tập hai lớp, trên tập đa lớp độ chính xác phân loại chưa tốt [26]. Nguyên nhân là chưa tìm ra được bộ trọng số thích hợp cho phép điều chỉnh mất cân bằng trên tập đa lớp.

Nghiên cứu đã chỉ ra rằng, điều chỉnh tỷ lệ trên tập dữ liệu k lớp chỉ thực hiện trực tiếp khi tìm được bộ trọng số w[w_1,w_2,..,w_k] thích hợp, với w_i là trọng số của lớp thứ i [26]. Giả sử cần tìm bộ trọng số w cho tập dữ liệu k lớp $i, j \in \{1..\mathrm{k}\}$, ma trận chi phí cm như sau:

$$\mathrm{cm}=\begin{bmatrix} 0 & cm_{12} & cm_{13} & .... & cm_{1k} \\ cm_{21} & 0 & cm_{23} & .... & cm_{2k} \\ .... & .... & .... & .... & .... \\ cm_{k1} & cm_{k2} & cm_{k3} & .... & 0 \end{bmatrix}$$

Giải phương trình hệ số (4) với k biến, nếu tìm được nghiệm của phương trình (4) $w\_=[w_1,w_2,..,w_k]^T$ thì các lớp có thể điều chỉnh mất cân bằng đồng thời [26]. Bộ nghiệm được xem dùng như là bộ trọng số điều chỉnh mất cân bằng đồng thời các lớp khi chi phí phân loại sai mẫu của các lớp là bằng nhau (w=w_). Trong thực tế chi phí phân loại sai mẫu của các lớp là không bằng nhau. Bộ nghiệm w_ tìm được từ phương trình (4) được xem là bộ chi phí phân loại sai của lớp (cm=w_). Bộ trọng số của lớp w=[w_1,w_2,..,w_k] được tính từ công thức (3) (tính từ chi phí của lớp, số mẫu của lớp và số mẫu tập huấn luyện) [26].

$$\begin{cases} w_1 \times \mathrm{c}m_{21} - w_2 \times \mathrm{c}m_{12} + w_3 \times 0 + \ldots + w_k \times 0 = 0 \\ w_1 \times \mathrm{c}m_{31} + w_2 \times 0 - w_3 \times \mathrm{c}m_{13} + \ldots + w_k \times 0 = 0 \\ \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots = 0 \\ w_1 \times \mathrm{c}m_{k1} + w_2 \times 0 + w_3 \times 0 + \ldots - w_k \times \mathrm{c}m_{1k} = 0 \\ w_1 \times 0 + w_2 \times \mathrm{c}m_{32} - w_3 \times \mathrm{c}m_{23} + \ldots + w_k \times 0 = 0 \\ \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots = 0 \\ w_1 \times 0 + w_2 \times \mathrm{c}m_{k2} + w_3 \times 0 + \ldots - w_k \times \mathrm{c}m_{2k} = 0 \\ \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots = 0 \\ \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots = 0 \\ w_1 \times 0 + w_2 \times 0 + w_3 \times 0 + \ldots - w_k \times \mathrm{c}m_{k-1,k} = 0 \end{cases} \quad (4)$$

Bộ nghiệm chỉ tìm được khi và chỉ khi hạng của ma trận hệ số M được trình bày theo công thức (5) là nhỏ hơn k (Rank(M)<k) [26]. Điều này cũng ít khi xảy ra, vì hạng của một ma trận $\frac{k(k-1)}{2} \times k$ với (k>2) hầu như luôn bằng k.

$$\begin{bmatrix} \mathrm{c}m_{21} & -\mathrm{c}m_{12} & 0 & \ldots & 0 \\ \mathrm{c}m_{31} & 0 & -\mathrm{c}m_{13} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & 0 \\ \mathrm{c}m_{k1} & 0 & 0 & \ldots & -\mathrm{c}m_{1k} \\ 0 & \mathrm{c}m_{32} & -\mathrm{c}m_{23} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & 0 \\ 0 & \mathrm{c}m_{k2} & 0 & & -\mathrm{c}m_{2k} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & \ldots & -\mathrm{c}m_{k-1,k} \end{bmatrix} \quad (5)$$

Khi hạng của ma trận hệ số M là k (Rank(M) = k), có nghĩa là không có bộ nghiệm nào được tìm thấy, tức không tìm được bộ trọng số thích hợp áp dụng điều chỉnh trực tiếp tỷ lệ đồng thời giữa các lớp [26]. Như vậy thay vì điều chỉnh trực tiếp trên tập đa lớp, chúng ta đưa về thành điều chỉnh gián tiếp [2]. Bằng cách điều chỉnh mất cân bằng trên các tập hai lớp và ma trận chi phí của hai lớp theo phương pháp truyền thống [4][7]. Các tập hai lớp là các cặp lớp được tách ra từ tập đa lớp, số tập hai lớp bằng tổ hợp chập 2 của k lớp. Ma trận chi phí của hai lớp được tách từ ma trận chi phí tập đa lớp. Trọng số được tính từ ma trận chi phí. Trọng số được gán cho mẫu trong tập huấn luyện. Giải thuật học dựng mô hình phân loại từ tập mẫu đã gán trọng số. Kết quả phân loại mẫu là kết quả được bình chọn từ tập mô hình hai lớp [26].

Tóm tắt điều chỉnh mất cân bằng trên tập đa lớp tiếp cận với chi phí phân loại sai [26].

Đầu vào: Tập huấn luyện D (với k lớp), ma trận chi phí, giải thuật học G(C4.5/J48)

1. Tạo ma trận hệ số M từ ma trận chi phí Matrix(cm)
2. Kiểm tra hạng của M nếu Rank(M) < k
3. Tìm ra bộ vector trọng số w=[w_1,w_2,..,w_k], gán trọng số cho mẫu trong tập D, kết quả thu được tập D*. Dùng G dựng nên mô hình H từ tập dữ liệu D^* (H=G(D*)).
4. Ngược lại

5. for i = 1 to k-1 do
6. for j = i + 1 to k do
7. Từ tập D lấy ra tập D_{ij}, D_{ij} là tập dữ liệu có 2 lớp chỉ chứa các mẫu thuộc lớp i và lớp j. $cm_{ij} = \begin{bmatrix} cm_{ii} & cm_{ij} \\ cm_{ji} & cm_{jj} \end{bmatrix}$ là ma trận chi phí của lớp i và j tách ra từ ma trận cm.
8. $h_{ij} \leftarrow$ Traditional_rescaling(D_{ij},cm_{ij},G)

% áp dụng phương pháp điều chỉnh tỷ lệ truyền thống trên tập hai lớp D_{ij}.

9. End for
10. End for
11. $H(x) \leftarrow \arg\max_{y\in\{1,..,k\}} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} I(h_{ij}(x) = y)$

% kết quả phân loại của mẫu x là sự bình chọn nhãn từ tập mô hình phân loại h_{ij}

12. end if

Đầu ra: mô hình phân loại với chi phí.

4 HỆ THỐNG QUYẾT ĐỊNH VỚI CHI PHÍ

4.1 Xây dựng hệ thống quyết định với chi phí

Dựa trên kết quả nghiên cứu của phương pháp điều chỉnh tỷ lệ mất cân bằng trên tập đa lớp tiếp cận với chi phí. Trong phần này chúng tôi đưa ra định nghĩa cho hệ thống quyết định truyền thống (DS– Decision System) có 5 thành phần như sau:

Định nghĩa 1: S= (U, C, d, V, f)

Trong đó, U là tập không gian mẫu, không rỗng. C là tập các thuộc tính điều kiện và d là thuộc tính quyết định. Q là tập các thuộc tính, $Q = U \cup d$. $V = U_{q\in Q} V_q$ với V_q là miền giá trị của thuộc tính q. f: $U \times Q \rightarrow V$ là tổng chức năng quyết định (chức năng thông tin). Như vậy $f(x,q) \in V_q$ với $q \in Q$, $x \in U$ [12].

Dựa trên định nghĩa 1, mở rộng định nghĩa cho hệ thống quyết định với chi phí có 7 thành phần sau:

S= (U, C, d, V, f, cm, n)

Các thành phần U, C, d, V, f tương tự như định nghĩa 1. Với $k=|V_d|$ là miền giá trị của thuộc tính quyết định (số lớp của tập dữ liệu). Khi đó $n \in R^+ \cup \{0\}$ là giá trị lớn nhất có thể gán các phần tử trong ma trận chi phí cm. mc: $k \times k \rightarrow R^+ \cup \{0\}$, với $i, j \in \{1..k\}$, $0 <= cm_{ij} <= n$, $cm_{ii} = 0$ chi phí gán nhãn đúng mẫu, cm_{ij} (nếu $i \neq j$) là chi phí phải trả khi gán nhãn sai mẫu (mẫu thuộc lớp i được gán nhãn vào lớp j).

Đơn vị đo lường chi phí: thực tế để xây dựng ma trận chi phí phải dựa vào từng lĩnh vực chuyên môn. Đơn vị của chi phí có thể là tiền, thời gian tính toán, thời gian khắc phục hậu quả, sức khỏe bệnh nhân...

Ví dụ 1: trong tập dữ liệu ann [3] ghi nhận kiểm tra mắc bệnh suy giảm tuyến giáp của 7200 bệnh nhân. Mỗi bệnh nhân được kiểm tra tuyến giáp và chẩn đoán vào một trong ba nhóm: lớp 1 là có bệnh, có 166 mẫu; lớp 2 có tuyến giáp dưới mức bình thường, có 368 mẫu và lớp 3 là bình thường, có 6666 mẫu. Tập có phân phối mẫu các lớp có tỷ lệ [1:2.2:40.1], tập có tỷ lệ mất cân bằng cao [6] và ma trận chi phí của tập có dạng như sau:

$$cm = \begin{bmatrix} cm_{11} & cm_{12} & cm_{13} \\ cm_{21} & cm_{22} & cm_{23} \\ cm_{31} & cm_{32} & cm_{33} \end{bmatrix}$$

Để xây dựng được ma trận chi phí, cần xác định các giá trị của ma trận, trong đó cm_{11} là chi phí khắc phục hậu quả một người bệnh được chẩn đoán là bệnh (chi phí bằng 0); cm_{12} là chi phí khắc phục hậu quả một người bệnh được chẩn đoán là dưới bình thường; cm_{13} là chi phí khắc phục hậu quả một người bệnh được chẩn đoán là bình thường; mở rộng tính toán các giá trị còn lại. Trong thực tế việc phân loại sai mẫu vào các lớp khác nhau có chi phí khác nhau.

Ma trận chi phí được xây dựng từ việc tính toán giá phải trả khi phân loại sai mẫu của một lớp vào các lớp còn lại. Thông thường việc ước lượng này là mất nhiều công sức tính toán. Trong một số lĩnh vực ma trận chi phí được xây dựng từ tri thức chuyên gia [26].

4.2 Chỉ số đánh giá

Trên tập dữ liệu đa lớp mất cân bằng số mẫu lớp nhỏ là rất ít, tỷ lệ phân loại chính xác trên lớp ít mẫu là rất được quan tâm. Trong quá trình thực nghiệm, kết quả phân loại mẫu được trình bày tại ma trận confusion matrix [26]. Kết quả được so sánh trên ba chỉ số đo rất được quan tâm trong các ứng dụng chẩn đoán y học là Accuracy (độ chính

xác toàn tập) và $Recall_i$ (độ bao phủ lớp thứ i), $Precision_i$ độ chính xác lớp thứ i.

Bảng 1: Confusion matrix trên tập đa lớp

Lớp đúng	Lớp dự đoán			
	Lớp 1	**Lớp 2**	**…**	**Lớp k**
Lớp 1	N_{11}	N_{12}	…	N_{1k}
Lớp 2	N_{21}	N_{22}	…	N_{2k}
…	…	…	…	…
Lớp k	N_{k1}	N_{k2}	…	N_{kk}

N_{ii} số lượng mẫu thuộc lớp i phân đúng vào lớp i. N_{ij} số lượng mẫu thuộc lớp i phân loại sai vào lớp j (với $i \neq j$).

Accuracy(Acc): độ chính xác phân loại toàn tập dữ liệu là xác suất tính trên số mẫu phân loại đúng trên tổng số mẫu phân loại của tập.

$Recall_i$ (Re_i): độ bao phủ, khả năng phát hiện chính xác một mẫu thuộc lớp thứ i, là xác suất phát hiện chính xác mẫu lớp i trên tổng số mẫu phân loại thuộc lớp i.

$Precision_i$ (Pr_i): độ chính xác của lớp thứ i, là xác suất phân loại chính xác mẫu lớp i trên tổng số mẫu được phân loại thuộc về lớp i.

$$Acc = \frac{\sum_{i=1}^{|k|} N_{ii}}{\sum_{i=1}^{k}\sum_{j=1}^{k} N_{ij}} \qquad \mathrm{Re}_i = \frac{N_{ii}}{\sum_{j=1}^{|k|} N_{ij}}$$

$$\mathrm{Pr}_i = \frac{N_{ii}}{\sum_{j=1}^{|k|} N_{ji}}$$

5 XÂY DỰNG HỆ THỐNG QUYẾT ĐỊNH VỚI CHI PHÍ

Chúng tôi đã xây dựng hệ thống quyết định với chi phí dựa trên phương pháp phân loại dữ liệu mất cân bằng tiếp cận với chi phí. Hệ thống được áp dụng vào chẩn đoán bệnh suy giảm tuyến giáp ở người (ann tập dữ liệu có tỷ lệ mất cân bằng >40) [6]. Tập dữ liệu được tiến hành với hai loạt thực nghiệm điều chỉnh trực tiếp và gián tiếp. Thông qua kết quả thực nghiệm kiểm chứng vai trò của chi phí trong việc cải thiện tỷ lệ phân loại chính xác trên lớp ít mẫu.

Điều kiện cần để thực thi hệ thống là tập huấn luyện, ma trận chi phí và tập kiểm thử. Vì vậy tập dữ liệu được tách ra thành hai tập con: tập huấn luyện và tập kiểm thử theo tỷ lệ 1:1. Giá trị U, C, d, V, f được xác định từ tập huấn luyện, đặt n=10.

Khởi tạo ma trận chi phí cm cho tập dữ liệu ann, chi phí chỉ mang tính chất chất thực nghiệm. Giá trị ma trận chi phí cm thỏa 3 ràng buộc. Thứ nhất, ma trận có ít nhất một giá trị 1 và giá trị của ma trận được sinh ra ngẫu nhiên và phải nhỏ hơn hoặc bằng n. Thứ hai, chi phí phân loại sai của lớp nhỏ nhất vào lớp lớn nhất là lớn nhất. Thứ ba, chi phí phân loại sai của lớp lớn nhất vào lớp nhỏ nhất là nhỏ nhất.

Chúng tôi xây dựng bộ công cụ của hệ thống quyết định với chi phí có các chức năng sau:

5.1 Sinh ma trận chi phí điều chỉnh trực tiếp mất cân bằng trên tập đa lớp

Đặt hằng n=10, khởi tạo ma trận cm(k, k) với k là số lớp của tập dữ liệu, $i, j \in \{1..k\}$, cm[i,j] $\in R^+ \cup \{0\}$, 0<=cm[i,j] <=n (cm_{ij} được sinh ngẫu nhiên). Khởi tạo vector cv với k phần tử, $i \in \{1..k\}$, cv[i] $\in R^+ \cup \{0\}$, 0<=cv[i]<=n (cv[i] được sinh ngẫu nhiên).

Kiểm tra điều kiện của ma trận chi phí cm thỏa ba điều kiện đã trình bày ở trên. Kiểm tra cv có phải là nghiệm của ma trận hệ số từ ma trận chi phí cm theo công thức (4), hạng của ma trận hệ số theo công thức (5) là nhỏ hơn k. Nếu đúng thì vector cv được dùng tính trọng số w, trọng số w dùng điều chỉnh mất cân bằng trực tiếp trên tập đa lớp. Chi phí này gọi là chi phí nhất quán. Giá trị vector cv được kiểm tra thỏa 3 ràng buộc như ma trận chi phí trước khi gán cho lớp [26].

5.2 Sinh ma trận chi phí điều chỉnh gián tiếp mất cân bằng trên tập đa lớp

Đặt hằng n=10, khởi tạo ma trận cm(k,k) với k là số lớp của tập dữ liệu, $i, j \in \{1..k\}$, các giá trị của ma trận được sinh ra ngẫu nhiên tương tự như ma trận chi phí điều chỉnh trực tiếp (0<=cm[i,j] <=n). Kiểm tra điều kiện của ma trận chi phí cm thỏa ba điều kiện ràng buộc được trình bày như trên, ngoài ra hạng (Rank) của ma trận hệ số theo công thức (5) là không nhỏ hơn k. Ma trận chi phí này được dùng điều chỉnh gián tiếp trên tập đa lớp, chi phí này gọi là chi phí không nhất quán [24][26].

5.3 Dựng mô hình phân loại với chi phí

Trọng số w_i gán cho mẫu lớp i được tính từ chi phí của lớp. Khi điều chỉnh trực tiếp chi phí lớp là nghiệm của phương trình (4). Khi điều chỉnh gián tiếp chi phí tập hai lớp được tính theo công thức (2). Áp dụng công thức (3) tính trọng số w_i của lớp thứ i. Trọng số lớp dùng gán cho các mẫu thuộc lớp trong tập huấn luyện. Tập huấn luyện đã gán

trọng số đưa vào giải thuật học C4.5 (trên java C4.5 là J4.8). Kết quả thu được là mô hình phân loại với chi phí.

5.4 Chức năng phân loại

Tập kiểm thử được đưa vào mô hình phân loại với chi phí. Kết quả phân loại được trình bày qua ma trận confusion matrix và thống kê các chi số đo Acc, Re_i, Pr_i (i là lớp được quan tâm).

6 KẾT QUẢ THỰC NGHIỆM

Tập dữ liệu ann thuộc lĩnh vực chẩn đoán tuyến giáp, tập được chọn từ UCI (University of California, Irvine) của Blake *et at*., 1998 [3]. Tập dữ liệu ann có 7200 mẫu phân phối vào 3 lớp, được tài trợ bởi Randolf Werner. Lớp 1 là có bệnh, 166 mẫu; lớp 2 tuyến giáp dưới mức bình thường, 368 mẫu và lớp 3 là bình thường, 6666 mẫu [166;368;6666]. Tập dữ liệu có 21 thuộc tính (15 thuộc tính kiểu nhị phân, 6 thuộc tính kiểu liên tục).

Tập dữ liệu ann được tách ra là 2 tập con: tập huấn luyện (tập train) [83;184;3333] và tập kiểm thử (tập test) [83;184;3333]. Lớp được quan tâm đến tỷ lệ phân loại chính xác là lớp 1 (lớp bệnh).

Thực nghiệm 1: thực hiện dựng mô hình với giải thuật J4.8 và tập huấn luyện với các mẫu không gán trọng số [25]. Dùng tập test kiểm thử, kết quả ghi nhận trong thực nghiệm như sau:

Acc = 99.47 %, Re_1 = 94.00 %, Pr_1 = 95.12%

Bảng 2: Confusion matrix trên TN1

Lớp đúng	Lớp dự đoán		
	1	**2**	**3**
1	78	0	5
2	0	175	9
3	4	1	3328

Thực nghiệm 2 (TN2): Kết quả thực nghiệm khi điều chỉnh trực tiếp với chi phí nhất quán. Khởi tạo ma trận chi phí. Bộ nghiệm tìm được là cv={1.0000; 0.8686; 0.5992}. Như vậy chi phí lớp 1 là 1.000, chi phí lớp 2 là 0.8686, chi phí lớp 3 là 0.5992. Chi phí lớp được dùng tính trọng số lớp, trọng số gán cho mẫu trong tập train. Dựng mô hình từ tập train đã được gán trọng số. Mô hình được dùng kiểm thử với tập test. Kết quả ghi nhận trong thực nghiệm như sau:

Acc = 99.61 %, Re_1 = 98.80 %, Pr_1 = 96.47%

Bảng 3: Confusion matrix trên TN2

Lớp đúng	Lớp dự đoán		
	1	**2**	**3**
1	82	0	1
2	0	177	7
3	3	3	3327

Bảng 4: Thống kê sự thay đổi phân phối mẫu trên tập (điều chỉnh trực tiếp)

Lớp dữ liệu	chi phí nhất quán		không chi phí	
	Số mẫu được dự đoán	***Số mẫu đúng lớp***	***Số mẫu được dự đoán***	***Số mẫu đúng lớp***
1	85	82	82	78
2	180	177	176	175
3	3335	3327	3342	3328

Thực nghiệm 3: Kết quả thực nghiệm khi điều chỉnh gián tiếp với chi phí không nhất quán. Khởi sinh ma trận chi phí không nhất quán, tập huấn luyện được tách ra thành tổ hợp chập 2 của k các tập hai lớp, ma trận chi phí của các tập hai lớp được tách ta từ ma trận sinh ra ban đầu. Sau đó, xây dựng các mô hình từ các tập hai lớp và ma trận chi phí tương ứng. Sử dụng tập kiểm tra để phân loại dữ liệu dựa vào việc bình chọn từ tập các mô hình. Kết quả thực nghiệm như sau:

Acc = 99.56 %, Re_1 = 96.39 %, Pr_1 = 95.23%

Bảng 5 : Confusion matrix trên TN3

Lớp đúng	Lớp dự đoán		
	1	**2**	**3**
1	80	0	3
2	0	176	8
3	4	1	3328

Bảng 6: Thống kê sự thay đổi phân phối mẫu trên tập (điều chỉnh gián tiếp)

Lớp dữ liệu	chi phí điều chỉnh gián tiếp		không chi phí	
	Số mẫu được dự đoán	***Số mẫu đúng lớp***	***Số mẫu được dự đoán***	***Số mẫu đúng lớp***
1	84	80	82	78
2	177	176	176	175
3	3339	3328	3342	3328

Phân tích kết quả thực nghiệm, với TN1 số mẫu được phân loại vào lớp 1 là 82 trong đó có 78 mẫu

là chính xác thuộc về lớp 1 (78/82). Với TN2 số mẫu được phân loại vào lớp 1 là 85 trong đó có 82 mẫu là chính xác thuộc về lớp 1 (82/85). Với TN3 số mẫu được phân loại vào lớp 1 là 84 trong đó có 80 mẫu là chính xác thuộc về lớp 1 (80/84). Kết quả TN2 và TN3 đều nâng cao số lượng bệnh nhân được chẩn đoán đúng bệnh điều này mang lại cho bệnh nhân cơ hội chăm sóc, cơ hội sống nhờ phát hiện kịp thời, đúng thời điểm. Việc chẩn đoán sai người có bệnh vào các nhóm khác phải trả giá rất cao: sức khỏe, sự sống của bệnh nhân lý do là không phát hiện kịp thời, bỏ qua thời điểm vàng điều trị bệnh.

Từ kết quả thực nghiệm cho thấy đã có sự phân phối mẫu lại giữa các lớp; số lượng mẫu được phân loại có xu hướng di chuyển vào lớp được gán trọng số cao. Sự dịch chuyển này làm nâng cao số lượng mẫu phân loại vào lớp ít mẫu và số mẫu được phân loại đúng trên lớp ít mẫu.

Kết quả phân loại mẫu phụ thuộc nhiều vào giá trị ma trận chi phí. Chúng tôi tiến hành thực nghiệm 10 lần (TN2 và TN3), kết quả được trình bày dưới dạng: trung bình ± độ lệch chuẩn.

Bảng 7: Thống kê trên 3 dạng thực nghiệm

	chi phí nhất quán	chi phí không nhất quán	không chi phí
Acc	0.9962 ± 0.0004	0.9953 ± 0.0003	0.9947
Re (lớp1)	0.9856 ± 0.0069	0.9675 ±0.0108	0.9400
Pr (lớp1)	0.9635 ± 0.0037	0.9525 ±0.0005	0.9512

Đối với các tập dữ liệu mất cân bằng cao, cả chi phí nhất quán và chi phí không nhất quán đều cải thiện được tỷ lệ phân loại chính xác mẫu lớp nhỏ của tập dữ liệu trên hai độ đo Recall và Precision. Độ chính xác phân loại ổn định và không làm giảm tỷ lệ phân loại chính xác trên toàn tập (độ đo Accuracy). Thông qua kết quả thực nghiệm cho thấy được vai trò của chi phí trong quá trình phân loại dữ liệu. Việc gán nhãn cho mẫu nhạy cảm với chi phí, chi phí đã làm thay đổi sự phân phối mẫu giữa các lớp. Sự thay đổi có xu hướng di chuyển về lớp có chi phí phân loại sai lớn. Chứng tỏ việc gán nhãn cho mẫu không chỉ phụ thuộc vào nguyên tắc số đông mà còn phụ thuộc vào chi phí.

7 KẾT LUẬN

Bài báo đã trình bày các bước xây dựng hệ thống quyết định với chi phí phân loại sai. Hệ thống cải thiện được tỷ lệ phân loại chính xác trên lớp ít mẫu trong tập mất cân bằng. Nếu áp dụng hệ thống vào ứng dụng chẩn đoán y học sẽ nâng cao hiệu quả chẩn đoán, nếu áp dụng vào lĩnh vực phát hiện xâm nhập, tấn công sẽ năng cao hiệu quả giám sát hệ thống. Tuy nhiên, chúng tôi chưa đề xuất được một bộ tiêu chí xây dựng ma trận chi phí cho tập dữ liệu mất cân bằng, việc xây dựng ma trận chủ yếu dựa vào tri thức chuyên gia.

TÀI LIỆU THAM KHẢO

1. Abe, N., Zadrozny, B., Langford, J. (2004), An iterative method for multi-class cost-sensitive learning, Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, pp. 3–11.
2. Allwein, E. L., Schapire, R. E., Singer, Y. (2000), Reducing multiclass to binary: A unifying approach for margin classifiers, Journal of Machine Learning Research 1, pp. 113–141.
3. Blake, C., Keogh, E., Merz, C. J. (1998), UCI repository of machine learning databases, [http://www.ics.uci.edu/~mlearn/MLRepository.html], Department of Information and Computer Science, University of California, Irvine, CA.
4. Breiman, L., Friedman, J. H., Olsen, R. A., Stone, C. J. (1984), Classification and Regression Trees. Wadsworth, Belmont, CA.
5. Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. (2002), SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16, pp. 321–357.
6. Ding, Z. (2011). Diversified Ensemble Classifier for Highly imbalanced Data Learning and their application in Bioinformatics, Ph. D thesis, College of Arts and science, Department of Computer Science, Georgia State University,2011. Http://digitalarchive.gsu.edu/cs_diss/60
7. Domingos, P. (1999), MetaCost: A general method for making classifiers costsensitive, Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, pp. 155–164.
8. Drummond, C., Holte, R. C. (2003), C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling,

Working Notes of the ICML'03 Workshop on Learning from Imbalanced Data Sets, Washington, DC.

9. Elkan, C. (2001), The foundations of cost-senstive learning, Proceedings of the 17th International Joint Conference on Artificial Intelligence. Seattle, WA, pp. 973–978.
10. Engen Vegard. 2010. Machine Learning for Network Based Intrusion Detection. Ph. D thesis, Bounemouth University, 2010.
11. Hido, S. and Kashima, H. 2008. Roughly Balanced Bagging for Imbalanced Data. "In Proceedings of SIAM Conference on Data Mining (SDM2008), Atlanta, Georgia, USA, April, 2008.
12. Hong Zhao, Fan Min, William Zhu 2012. Minimal cost feature selection of data with normal distribution measurement errors. Lab of Granular Computing, Zhangzhou Normal University, Zhangzhou 363000, China.
13. Jeffrey P. Bradford., Clayton Kunz., Ron Kohavi., Clifford Brunk., Carla E. Brodley. (1998), Pruning Decision Trees with Misclassification Costs. ECML-98, pp.131-136.
14. Ling, C. X., Yang, Q.,Wang, J., Zhang, S. (2004), Decision trees with minimal costs, Proceedings of the 21st International Conference on Machine Learning. Banff, Canada, pp. 69–76.
15. Liu, X.-Y., Zhou, Z.-H. (2006), The influence of class imbalance on cost-sensitive learning: An empirical study, Proceedings of the 6th IEEE International Conference on Data Mining. Hong Kong, China, pp. 970–974.
16. Lozano, A. C., Abe, N., 2008. Multi-class cost-sensitive boosting with p-norm loss functions, Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, NV, pp. 506–514.
17. Maloof, M. A., 2003. Learning when data sets are imbalanced and when costs are unequal and unknown, Working Notes of the ICML'03Workshop on Learning from Imbalanced Data Sets. Washington, DC.
18. Margineantu, D. (2001), Methods for cost-sensitive learning. Ph.D. thesis, department of Computer Science, Oregon State University, Corvallis, OR.
19. Provost, F, Domingos, P. (2003), Tree induction fof probability-base ranking, Machine Learning 52 (3), 199–215.
20. Quinlan, J. R. (1993), C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, California.
21. Ting, K. M. (2002), An instance-weighting method to induce cost-sensitive trees. IEEE Transactions on Knowledge and Data Engineering 14 (3), 659–665.
22. Turney, P. D. (2000), Types of cost in inductive concept learning, Proceedings of the ICML'2000 Workshop on Cost-Sensitive Learning. Stanford, CA, pp. 15–21.
23. Witten, I. H., Frank, E. (2011), Data Mining: Practical Machine Learning Tools and Techniques, Third Edition. Morgan Kaufmann Publishers. www.mkp.com. ISBN: 978-0-12-374856-0.
24. Yang, Y. and Ma, G. 2010. Ensemble-based Active Learning for Classification Problem. J. Biomedical and Engineering, 2010, 3, pp. 1021- 1028. Published online in SciRes. Http:/www. Scrip.org/journal/jbise.
25. Zadrozny, B., Langford, J., Abe, N. (2002), A simple method for cost-sensitive learning. Tech. rep., IBM.
26. Zhou, Z.-H., Liu, X.-Y. (2006a), On multi-class cost-sensitive learning, Proceeding of the 21st National Conference on Artificial Intelligence. Boston, WA, pp. 567–572.