DỰ ĐOÁN BỆNH NHÂN MẮC BỆNH TIỂU ĐƯỜNG BẰNG THUẬT TOÁN IG-ENANSMOTE

NGUYỄN THỊ LAN ANH, NGUYỄN THỊ LAN***

**Giảng viên, **Học viên Cao học Khoa Tin học, Trường ĐH Sư Phạm, ĐH Huế*

Tóm tắt: Bệnh tiểu đường là một trong những chứng bệnh phổ biến và có nhiều tác hại nghiêm trọng đối với sức khỏe con người, có khả năng dẫn đến tử vong nếu không được điều trị kịp thời. Vì vậy, phát hiện bệnh sớm để có phương pháp can thiệp sớm phù hợp là một trong những nhu cầu cấp thiết. Trong bài báo này, chúng tôi đề xuất một phương pháp dự đoán bệnh nhân mắc bệnh tiểu đường kết hợp giữa lựa chọn đặc trưng và phương pháp sinh thêm phần tử để làm loại bỏ các thuộc tính dư thừa và làm giảm sự mất cân bằng dữ liệu. Kết quả thực nghiệm cho thấy phương pháp do chúng tôi đề xuất có thể so sánh được với các phương pháp dự đoán khác.

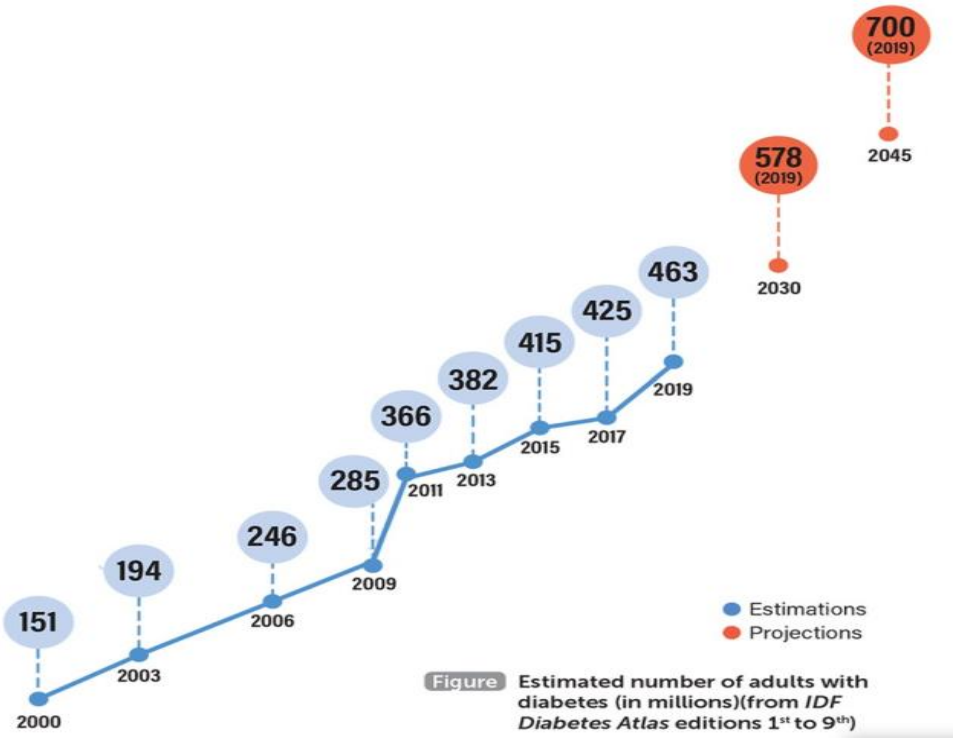
Từ khóa: Bệnh tiểu đường, Dữ liệu mất cân bằng, phương pháp sinh thêm phần tử, Lựa chọn đặc trưng

1. TỔNG QUAN

Bệnh tiểu đường là một trong những chứng bệnh phổ biến và có nhiều tác hại nghiêm trọng đối với sức khỏe con người, thuộc nhóm các bệnh chuyển hóa bất thường và mãn tính, gây ra tình trạng lượng đường trong máu tăng cao, và kéo dài [1] mà nguyên nhân là do sự rối loạn chuyển hóa không đồng nhất, có đặc điểm tăng glucose huyết do khiếm khuyết về tiết insulin, về tác động của insulin hoặc cả hai. Lượng đường trong máu cao có thể dẫn đến việc đi tiểu nhiều, khát, đói, đặc biệt là thèm đồ ngọt. Bệnh còn gây tổn thương nghiêm trọng cho mạch máu, tim, thận, mắt và thần kinh, do tăng glucose mạn tính trong thời gian dài dẫn đến những rối loạn chuyển hóa carbohydrate, protid và lipid. Nếu không điều trị khẩn cấp, bệnh tiểu đường có thể gây ra nhiều biến chứng nghiêm trọng cũng như các tác dụng phụ tiêu cực, như nhiễm toan ceton (diabetic ketoacidosis), tăng thẩm thấu do tăng glucose máu (nonketotic hyperosmolar), bệnh tim, đột quỵ, suy thận, loét bàn chân, giảm thị lực và mù. Ngoài ra, những người mắc bệnh tiểu đường có nhiều khả năng bị nhiễm COVID-19 hơn và có nguy cơ gặp biến chứng hoặc tử vong cao hơn [2].

Gần đây, bệnh tiểu đường đã trở thành nguyên nhân gây tử vong và bệnh tật hàng đầu trên thế giới. Theo Liên đoàn Tiểu đường Quốc tế, năm 2019 có khoảng 463 triệu người trên toàn thế giới mắc bệnh. Dự đoán đến năm 2045, số lượng người mắc tiểu đường sẽ tăng thêm khoảng 51%, đạt tới con số khoảng 700 triệu người [3]. Hình 1 cho thấy sự gia tăng của bệnh nhân mắc tiểu đường theo thời gian [4].

Bệnh tiểu đường được phân thành ba loại chính: loại 1, loại 2 và tiểu đường thai kỳ. Tiểu đường loại 1 phát sinh khi cơ thể không thể sản sinh insulin vì tế bào tụy đảm nhận chức năng này bị phá hủy. Tiểu đường loại 2 phát triển khi cơ thể trở nên đề kháng với insulin. Tiểu đường thai kỳ phát triển khi hormone ngăn chặn insulin được cơ thể sản sinh trong quá trình mang thai. Theo thống kê, năm 2019 có hơn 1,1 triệu người dưới 18 tuổi mắc bệnh tiểu đường loại 1, hơn 20 triệu trẻ em bị ảnh hưởng do tác hại của tiểu đường thai kỳ.



Hình 1. Biểu đồ gia tăng số lượng bệnh nhân mắc tiểu đường trên thế giới

Tuy nhiên, tiểu đường loại 2 là loại phổ biến nhất và thường thấy ở người trưởng thành, được phát hiện ở khắp các quốc gia. Năm 2019, thế giới có khoảng 374 triệu người có nguy cơ cao mắc tiểu đường loại 2, với khoảng 4,2 triệu ca tử vong. Nghiên cứu cho thấy cứ 2 người mắc bệnh tiểu đường thì có 1 người không được chẩn đoán, dẫn đến khả năng tử vong tăng lên. Nhìn chung, nhiều người mắc bệnh bệnh tiểu đường loại 2 hiếm khi biểu hiện bất kỳ triệu chứng nào, dẫn đến tăng các yếu tố nguy cơ gây ra biến chứng vì không được chữa trị kịp thời [5]. Do đó, phát hiện và điều trị sớm bệnh tiểu đường là một trong những bước quan trọng và cần thiết để làm giảm nguy cơ gây biến chứng nghiêm trọng và tử vong ở các bệnh nhân mắc bệnh này.

Nhiều nghiên cứu nhằm dự đoán bệnh nhân mắc bệnh tiểu đường bằng các phương pháp học máy đã được thực hiện [4]. Tuy nhiên, kết quả đạt được vẫn chưa khả quan lắm. Một trong những nguyên nhân là sự mất cân bằng dữ liệu trong tập dữ liệu mẫu, khi số lượng người bị tiểu đường bé hơn số người khỏe mạnh. Sự mất cân bằng dữ liệu làm giảm hiệu quả của các thuật toán phân lớp truyền thống vì các bộ phân lớp này có khuynh hướng dự đoán lớp đa số - lớp có số lượng phần tử nhiều hơn - và bỏ qua lớp thiểu số - lớp có số lượng phần tử ít hơn [6]. Nói cách khác, hầu hết các phần tử thuộc lớp đa số sẽ được phân lớp đúng và các phần tử thuộc lớp thiểu số cũng sẽ được dự đoán thuộc lớp đa số, kết quả là độ chính xác toàn thể (accuracy) của việc phân lớp rất cao trong khi độ nhạy (sensitivity) lại rất thấp.

Để khắc phục nhược điểm này, trong bài báo này chúng tôi đề xuất một phương pháp kết hợp giữa lựa chọn đặc trưng và kỹ thuật phân lớp dữ liệu mất cân bằng nhằm làm tăng chất lượng của việc dự đoán bệnh nhân mắc bệnh tiểu đường loại 2.

2. THUẬT TOÁN IG-ENANSMOTE DỰ ĐOÁN BỆNH TIỂU ĐƯỜNG

Bài báo này đề xuất thuật toán IG-ENaNSMOTE để dự đoán bệnh nhân mắc bệnh tiểu đường loại 2 dựa trên kỹ thuật lựa chọn đặc trưng và thuật toán sinh thêm phần tử.

Lựa chọn đặc trưng

Lựa chọn đặc trưng trong những nghiên cứu gần đây được cho thấy là có tác dụng trong việc cải thiện hiệu suất bài toán phân lớp dữ liệu mất cân bằng [7]. Ý tưởng cơ bản của kỹ thuật này là chỉ giữ lại một tập con các đặc trưng quan trọng của dữ liệu thay vì sử dụng tất cả các đặc trưng để thực hiện phân lớp. Có nhiều kỹ thuật lựa chọn đặc trưng khác nhau như Filter, Wrapper hay phương pháp kết hợp [8]. Một trong những phương pháp lựa chọn đặc trưng phổ biến thuộc nhóm Filter là sử dụng Độ lợi thông tin (Information Gain) để tính trọng số của các đặc trưng và giữ lại các đặc trưng có giá trị nhất dựa vào giá trị ngưỡng cho trước.

Với tập dữ liệu D, tập các đặc trưng F, Độ lợi thông tin IG(A) của đặc trưng $A \in F$ được xác định như sau:

$$IG(A) = H(D) - H(D|A) \quad (1)$$

Trong đó:

H(D) là entropy của tập dữ liệu D

H(D|A) là entropy có điều kiện của D với đặc trưng A

Thuật toán ENaNSMOTE để nâng cao hiệu suất phân lớp dữ liệu mất cân bằng

Nhằm nâng cao hiệu suất bài toán phân lớp dữ liệu mất cân bằng, nhiều phương pháp đã được đề xuất. Các phương pháp này được chia thành hai nhóm: nhóm phương pháp tiếp cận ở mức độ thuật toán như điều chỉnh xác suất ước lượng, sử dụng các hằng số phạt khác nhau cho các nhãn lớp khác nhau [9] và nhóm các phương pháp tiếp cận ở mức dữ liệu như sinh thêm các phần tử cho lớp thiểu số [10], giảm bớt các phần tử thuộc lớp đa số [9] hoặc các phương pháp kết hợp. Một số tác giả đã chỉ ra rằng các phương pháp tiếp cận ở mức dữ liệu hiệu quả hơn các phương pháp còn lại trong việc cải thiện độ chính xác sự phân lớp các tập dữ liệu mất cân bằng [9].

Một trong những phương pháp sinh thêm phần tử phổ biến, được áp dụng trong nhiều nghiên cứu khác nhau là phương pháp SMOTE. Phương pháp này cho phép các phần tử lớp thiểu số sinh thêm phần tử mới nằm giữa nó và một trong những phần tử láng giềng cùng nhãn lớp gần nhất của nó [10]. Tuy nhiên, SMOTE có nhược điểm là phụ thuộc vào tham số, nghĩa là số lượng láng giềng gần nhất *k* là tùy chọn. Ngoài ra, tất cả các phần tử lớp thiểu số đều sử dụng cùng một số lượng phần tử láng giềng gần nhất và cùng sinh một số lượng phần tử mới như nhau mà không quan tâm đến sự phân bố các đối tượng. Hơn nữa, SMOTE có khả năng sinh ra các phần tử mới nằm trong phạm vi lớp đa số, dẫn đến tình trạng chồng chéo giữa các lớp. Để khắc phục các nhược điểm trên, phương pháp ENaNSMOTE đã được đề xuất. Với mỗi phần tử thuộc lớp thiểu số, ENaNSMOTE sẽ sinh thêm phần tử mới nằm giữa nó và một trong những phần tử láng giềng tự nhiên mở rộng của nó [11] .

Gọi ENaN(y) là tập các phần tử láng giềng mở rộng của y∈D. Lúc đó, ∀x ∈D:

$$x \in ENaN(y) \Leftrightarrow x \in NN_\lambda(y) \vee y \in NN_\lambda(x) \Leftrightarrow x \in NN_\lambda(y) \vee x \in RNN_\lambda(y) \quad (1)$$

Trong đó: $NN_\lambda(y)$ là *λ láng giềng gần nhất của y trong D*

$$RNN_\lambda(y) = \{x \in | \, y \in NN_\lambda(x), y \in D\} \quad (2)$$

λ là giá trị riêng, được xác định bằng công thức (3) bên dưới:

$$\lambda = argmin_r\{(\forall r \in \{1,2,3\ldots\})(\forall y)(\exists x \neq y), (y \in NNr(x)) \wedge (x \in NNr(y)), x, y \in D\} \quad (3)$$

Để sinh thêm phần tử mới cho các phần tử thuộc lớp thiểu số, Hongjiao Guan và cộng sự [11] đã đề xuất thuật toán ENaNSMOTE như sau:

Thuật toán ENaNSMOTE
Input: Tập dữ liệu mất cân bằng *D*.
Output: Tập dữ liệu sau khi đã sinh thêm phần tử mới *Bal_D*.
1. Chia D thành hai tập con tương ứng với lớp thiểu số Pos và lớp đa số Neg
2. Tính giá trị ngen là số phần tử cần sinh thêm: ngen = nneg - npos;
3. Xác định tập láng giềng mở rộng ENaN = ENaN_Search(Pos);
4. Sinh thêm phần tử mới nằm giữa các phần tử lớp thiểu số và các láng giềng mở rộng của nó: Gen_Pos = SMOTE(Pos, ENaN, ngen);
5. Bal_D = Gen_Pos ∪ D.

Thuật toán ENaN_Search để tìm láng giềng mở rộng như sau:

Algorithm ENaN_Search
Input: *Tập dữ liệu D.*
Output: Tập láng giềng mở rộng *ENaN của D*
1. Khởi tạo $r = 1$, $\forall x_i \in D$, $NN_r(x_i) = \emptyset$, $RNN_r(x_i) = \emptyset$, $nb(x_i) = 0$;
2. while $r < |D|$ do
3. foreach $x_i \in D$ do
4. Xác định láng giềng gần nhất thứ r là x_j của x_i trong *D*;
5. *NNr(xi) = NNr(xi) ∪ {xj}, RNNr(xj) = RNNr(xj) ∪ {xi}, nb(xj) = nb(xj) + 1;*
6. $n(r) = \{|x_i|, nb(x_i) == 0\}$;
7. if $r > 1$ & $n(r) == n(r - 1)$ then
$\lambda = r - 1$;
8. foreach $x_i \in D$ do
9 . $ENaN(x_i) = NN_\lambda(x_i) \cup RNN_\lambda(x_i)$;
10. return *ENaN*;
11. else
12. $r = r + 1$;

3. THỰC NGHIỆM

Trong bài báo này, chúng tôi sử dụng bộ dữ liệu về bệnh nhân mắc bệnh tiểu đường: PIMA Indian diabetes dataset, là một bộ dữ liệu phổ biến thường được dùng trong các thực nghiệm dự đoán kết

quả bệnh tiểu đường loại 2. Bộ dữ liệu này chứa thông tin của các bệnh nhân có độ tuổi trên 20 từ Arizona, Hoa Kỳ. Các thông tin về bệnh nhân bao gồm số lần mang thai (Pregnancies), nồng độ glucose trong huyết tương (Glucose), huyết áp (BloodPressure), độ dày nếp gấp da cơ tam đầu (SkinThickness), chỉ số insulin (Insulin), chỉ số khối cơ thể (BMI), chức năng phả hệ bệnh tiểu đường (DiabetesPedigreeFunction), tuổi (Age) và có bị tiểu đường không (Outcome) được mô tả cụ thể trong Bảng 1.

Bảng 1. Thông tin về bộ dữ liệu PIMA Indian diabetes

STT	Tên thuộc tính	Loại dữ liệu	Miền dữ liệu
1	Pregnancies	Integer	0-17
2	Glucose	Integer	0-199
3	BloodPressure	Integer	0-122
4	SkinThickness	Integer	0-99
5	Insulin	Integer	1-846
6	BMI	Float	0-67.1
7	DiabetesPedigreeFunction	Float	0.078-2.42
8	Age	Integer	21-81
9	Outcome	Integer	0/1

Bộ dữ liệu này gồm 768 đối tượng, trong đó 268 đối tượng có kết quả bị tiểu đường loại 2, tỷ lệ mất cân bằng giữa lớp **tiểu_đường: không_tiểu_đường** là **1:1.87**.

Dữ liệu trên trước tiên được chuẩn hóa theo phương pháp chuẩn hóa Z-score để mỗi thuộc tính, trừ thuộc tính phân lớp, sau chuẩn hóa có mean bằng 0 và phương sai bằng 1. Quá trình này giúp giảm tác động của các phần tử ngoại lai lên tập dữ liệu. Trong bài báo này, chúng tôi đã sử dụng hàm normalize của gói lệnh SOM trong R.

Tiếp theo, chúng tôi tính giá trị Độ lợi thông tin (IG) cho các thuộc tính bằng hàm information_gain của gói FselectorRcpp trong R và giữ lại các thuộc tính có giá trị này lớn hơn 0, tức là loại bỏ các thuộc tính ít quan trọng. Độ lợi thông tin của các thuộc tính được thể hiện trong Bảng 2.

Bảng 2. Độ lợi thông tin (IG) của các thuộc tính

Thuộc tính	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	DiabetesPedigree Function	Age
IG	0.0272	0.1318	0	0	0.0412	0.0519	0.0144	0.0502

Dữ liệu sau xử lý được thực hiện kiểm chứng chéo 10-fold (10-fold crossvalidation). Trong mỗi lần lặp, lớp thiểu số của tập huấn luyện tương ứng được sinh thêm phần tử để đảm bảo có số lượng tương đương với số lượng phần tử của lớp đa số, sau đó kết hợp với lớp đa số để tạo ra tập huấn luyện mới. Trung bình cộng giá trị độ đo hiệu suất của kết quả phân lớp cho tập kiểm tra trong 10 lần lặp này chính là kết quả để đánh giá hiệu suất của mô hình.

Hình 2 mô tả toàn bộ các bước của mô hình chúng tôi đã đề xuất.

Trong bài báo này, chúng tôi sử dụng máy vector hỗ trợ (KSVM) thuộc gói lệnh kernlab làm bộ phân lớp chính để so sánh kết quả phân lớp bộ dữ liệu gốc không có can thiệp của thuật toán làm thay đổi số phần tử để xử lý sự mất cân bằng dữ liệu (ORIginal) và kết quả phân lớp có sử dụng thuật toán SMOTE thuộc gói lệnh smotefamily với kết quả khi sử dụng thuật toán ENaNSMOTE nhằm đánh giá tính hiệu quả của thuật toán này trong mô hình được đề xuất.

Để đánh giá hiệu quả của quá trình phân lớp, chúng tôi sử dụng các độ đo Recall (Độ hồi tưởng), Precision (Độ chính xác), F1, Sensitivity (Độ nhạy), Specificity (Độ đặc hiệu) và G-mean thay cho độ chính xác toàn thể (accuracy) như đối với việc phân lớp tập dữ liệu cân bằng. Khi phân lớp tập dữ liệu không cân bằng, Độ chính xác toàn thể này không thích hợp để xác định tính hiệu quả của mô hình phân lớp vì các bộ phân lớp thông thường sẽ cho ra độ chính xác toàn thể rất cao do phần lớn các phần tử đều được gán nhãn lớp là lớp đa số và rất hiếm phần tử được gán nhãn lớp của lớp thiểu số. Các giá trị Sensitivity, Specificity, G-mean được định nghĩa như sau:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (4), \qquad \text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (5)$$

$$\text{F1} = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall+Precision}} \quad (6)$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \quad (7), \quad \text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \quad (8)$$

$$\text{G} - \text{mean (Balanced accuracy)} = \sqrt{\text{Sensitivity} \times \text{Specificity}} \quad (9)$$

Ở đây, TP và FN lần lượt là số phần tử lớp thiểu số được dự đoán đúng và bị dự đoán sai so với nhãn lớp thực sự của chúng; TN và FP lần lượt là số phần tử lớp đa số được dự đoán đúng và sai so với nhãn lớp thực sự của chúng.

4. KẾT QUẢ THỰC NGHIỆM DỰ ĐOÁN BỆNH TIỂU ĐƯỜNG

Bảng 3 trình bày kết quả đánh giá hiệu suất dự đoán bệnh nhân mắc tiểu đường loại 2 của các phương pháp trước khi áp dụng chuẩn hóa dữ liệu: ORI-wo-norm, SMOTE-wo-norm, ENaNSMOTE-wo-norm và sau khi chuẩn hóa dữ liệu: ORI, SMOTE, ENaNSMOTE với phương pháp do chúng tôi đề xuất IG-ENaNSMOTE và phương pháp của Hongfang Zhou et al. [4], theo các độ đo Recall, Precision, F1, Sensitivity, Specificity và G-mean. Trong Bảng 3, những giá trị không tồn tại được thay bằng dấu -.

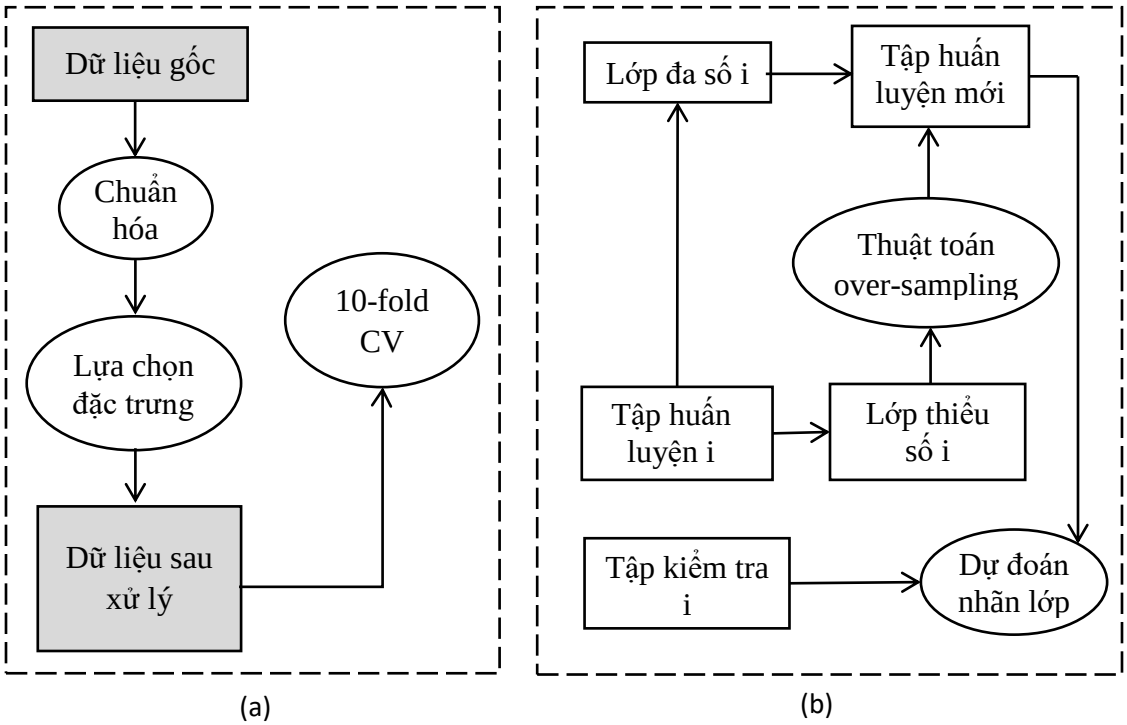
Bảng 3. Kết quả phân lớp (%).

Phương pháp	Precision	Recall	F1	Sensitivity	Specificity	G-mean
ORI-wo-norm	69.892	48.507	57.269	48.507	88.8	65.631
ORI	67.442	54.105	60.041	54.104	86	68.213
SMOTE-wo-norm	**70.745**	49.627	58.333	49.627	**89**	66.459
SMOTE	60.885	73.06	66.414	73.060	74.84	73.940
ENaNSMOTE-wo-norm	58.982	73.508	65.449	73.507	72.6	73.052
ENaNSMOTE	61.043	74.254	67.003	74.254	74.6	74.427

Hongfang Zhou et al. [4]	60.5	67.6	67.3	-	-	-
IG-ENaNSMOTE	62.155	**75.075**	**68.006**	**75.075**	75.5	**75.286**

So sánh kết quả khi áp dụng các phương pháp sinh thêm phần tử với kết quả của phương pháp ORI-wo-norm cho thấy hiệu quả của các phương pháp này trong xử lý sự mất cân bằng dữ liệu. Trong khi tỷ lệ số phần tử lớp thiểu số được dự đoán đúng bằng ORI thấp, tỷ lệ này tăng lên đáng kể, khi áp dụng các phương pháp sinh thêm phần tử. Cụ thể, Recall và Sensitivity của ORI tăng lần lượt là 1.12%, 24.55%, 25%, 25.75%, 26.57% so với các phương pháp SMOTE-wo-norm, SMOTE, ENaNSMOTE-wo-norm, ENaNSMOTE, IG-ENaNSMOTE. So với các phương pháp SMOTE-wo-norm, SMOTE, ENaNSMOTE-wo-norm, ENaNSMOTE, IG-ENaNSMOTE, F1 của phương pháp ORI tăng lần lượt là 1.06%, 9.15%, 8.18%, 9.73%, 10.74% và G-mean tăng lần lượt 0.823%, 8.31%, 7.42%, 8.8%, 9.66%.

Kết quả trên cũng thể hiện hiệu quả của phương pháp do chúng tôi đề xuất so với các phương pháp khác. Cụ thể, so với phương pháp ORI và ORI-wo-norm, giá trị Sensitivity và Recall tăng 26.57% và 20.97%, F1 tăng 10.74% và 7.97%, G-mean tăng 9.66% và 7.07%; so với phương pháp SMOTE và SMOTE-wo-norm, giá trị Sensitivity và Recall tăng lần lượt 25.45% và 2.02%, F1 tăng 9.67% và 1.59%, G-mean tăng 8.83% và 1.35%; so với phương pháp ENaNSMOTE và ENaNSMOTE-wo-norm, giá trị Sensitivity và Recall tăng lần lượt 1.57% và 0.82%, F1 tăng 2.56% và 1.00%, G-



Hình 2. Quá trình thực nghiệm dự đoán bệnh nhân mắc bệnh tiểu đường

Dữ liệu được chuẩn hóa và áp dụng phương pháp lựa chọn đặc trưng để loại bỏ các thuộc tính không cần thiết, sau đó áp dụng kiểm chứng chéo 10-fold (a). Với mỗi lần lặp i (i=1..10), lớp thiểu số của mỗi tập huấn luyện được sinh thêm phần tử để tập huấn luyện được cân bằng và áp dụng thuật toán phân lớp để dự đoán nhãn lớp cho tập kiểm tra tương ứng (b).

mean tăng 2.23% và 0.86%. Chúng tôi cũng thực hiện so sánh với phương pháp do Hongfang Zhou và các cộng sự đề xuất [4], kết quả cho thấy phương pháp của chúng tôi cho kết quả tốt hơn, cụ thể Sensitivity và Recall tăng 7.48% và F1 tăng 0.71%.

Bảng 3 cũng cho thấy vai trò của sự chuẩn hóa dữ liệu dựa trên kết quả so sánh giữa các phương pháp trước và sau khi chuẩn hóa. Recall và Sensitivy đã tăng 5.6%, F1 tăng 2.77%, G-mean tăng 2.58% khi phân lớp bằng KSVM. Khi áp dụng các phương pháp sinh thêm phần tử, chuẩn hóa dữ liệu cũng rất quan trọng. Ví dụ, giá trị Recall và Sensitivity trước và sau khi chuẩn hóa lần lượt là 47.627% và 73.06%, tăng 23.43%; giá trị F1 tăng 8.08% khi chuẩn hóa dữ liệu và G-mean tăng 7.48% khi áp dụng SMOTE làm phương pháp sinh thêm phần tử. Khi sử dụng phương pháp sinh thêm phần tử ENaNSMOTE, Recall và Sensitivity có áp dụng chuẩn hóa dữ liệu tăng 0.75%, F1 tăng 1.55%, G-mean tăng 1.38%, Precision tăng 2/06%, Specificity tăng 2%.

5. KẾT LUẬN

Trong bài báo này, chúng tôi đã trình bày một phương pháp làm tăng hiệu quả dự đoán bệnh nhân mắc bệnh tiểu đường loại 2 là IG-ENaNSMOTE. Chúng tôi cũng đã thực hiện đánh giá hiệu suất của thuật toán này và so sánh với các thuật toán khác. Kết quả thực nghiệm đã cho thấy thuật toán được đề xuất có hiệu quả tốt dựa trên các giá trị độ đo đánh giá hiệu suất Recall, Sensitivity, F1 và G-mean. Chúng tôi cũng thực hiện so sánh để thấy được hiệu quả của chuẩn hóa dữ liệu đối với việc dự đoán.

Tuy nhiên, do kích thước của tập dữ liệu dùng để thực nghiệm vẫn còn hạn chế, trong tương lai, chúng tôi sẽ tìm kiếm các tập dữ liệu thực khác để thực hiện việc kiểm thử đầy đủ hơn, từ đó có thể cho ra những kết quả chính xác và tin cậy hơn. Phương pháp này cũng có thể áp dụng để làm tăng hiệu quả dự đoán các bệnh khác như ung thư, bệnh tim, thận.

TÀI LIỆU THAM KHẢO

Tập dữ liệu PIMA Indian diabetes dùng để thực nghiệm trong bài báo này được download tại https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database vào ngày 24/11/2023.

[1] N. Barakat, A. P. Bradley, and M. N. H. Barakat, "Intelligible support vector machines for diagnosis of diabetes mellitus," *IEEE Trans. Inf.Technol. Biomed.,* vol. 14, no. 4, p. 1114–1120, Jul. 2010.

[2] R. Muniyappa and S. Gubbi, "COVID-19 pandemic, coronaviruses, and diabetes mellitus," *Amer. J. Physiol.-Endocrinol. Metabolism,* vol. vol. 318, no. no. 5, p. E736–E741, May 2020.

[3] "https://diabetesatlas.org/data/en/world/,".

[4] Hongfang Zhou, Yinbo Xin and Suli Li, "A diabetes prediction model based on Boruta feature selection and ensemble learning," *BMC Bioinformatics,* vol. 24, no. 224, 2023.

[5] G. Roglic, "WHO Global report on diabetes: A summary," *Int. J. Noncommunicable Diseases,* vol. 1, no. 1, p. 3–8, Jun. 2016.

[6] H. Sain and S. W. Purnami, "Combine Sampling Support Vector Machine for Imbalanced Data Classification," in *Procedia Comput. Sci.*, 2015.

[7] Alireza Moayedikia, Kok-Leong Ong, Yee Ling Boo, William GS Yeoh, Richard Jensen, "Feature selection for high dimensional imbalanced class data using harmony search," *Engineering Applications of Artificial Intelligence*, vol. 57, pp. 38-49, 2017.

[8] Saeys, Yvan & Inza, Iñaki & Larranaga, Pedro, "A review of feature selection techniques in bioinformatics," *Bioinformatics (Oxford, England)*, vol. 23, pp. 2507-17, 2007.

[9] H. He and E. A. Garcia, , "Learning from Imbalanced Data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, p. 1263–1284, Sep. 2009.

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE : Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, p. 321–357, Jan. 2002.

[11] Hongjiao Guan, Long Zhao, Xiangjun Dong, Chuan Chen, "Extended natural neighborhood for SMOTE and its variants in imbalanced classification," *Engineering Applications of Artificial Intelligence*, vol. 124 , 2023.

Title: A DIABETES PREDICTION METHOD BASED ON IG-ENANSMOTE ALGORITHM

Abstract: Diabetes is one of the common diseases and can cause serious damage to the human body. With time, this may result in death if there is no right treatment. Therefore, it is necessary to diagnose diabetes at an early stage and intervene early. In this paper, we propose a diabetes prediction method based on feature selection and oversampling technique to remove the redundant features and balance the dataset. The experimental results show that our approach is comparable to some other methods.

Keywords: Diabetes, Imbalanced data, Over-sampling, Feature selection

Lĩnh vực của bài báo: *Máy học*

NGUYỄN THỊ LAN ANH

Khoa Tin học – Trường ĐH Sư Phạm Huế

ĐT: 070-372-5257, Email: lananh257@gmail.com