THUẬT TOÁN HMU
TRONG BÀI TOÁN PHÂN LỚP DỮ LIỆU MẤT CÂN BẰNG

NGUYỄN THỊ LAN ANH
Trường Đại học Sư phạm, Đại học Huế
ĐT: 0120 372 5257, Email: lananh257@gmail.com

Tóm tắt: Phân lớp dữ liệu mất cân bằng là một bài toán quan trọng trong thực tế. Nhiều phương pháp đã được nghiên cứu nhằm nâng cao hiệu suất của bài toán phân lớp này. Trong bài báo này chúng tôi đề xuất một thuật toán làm giảm số lượng phần tử (Undersampling) dựa trên giá trị lề giả thuyết (hypothesis margin) của các đối tượng thuộc lớp đa số để cải thiện hiệu suất phân lớp tập dữ liệu mất cân bằng.

Từ khóa: Dữ liệu mất cân bằng, phương pháp làm giảm số lượng phần tử, lề giả thuyết, Hypothesis margin

1. GIỚI THIỆU

Trong những năm trở lại đây, vấn đề dữ liệu mất cân bằng là một trong những vấn đề quan trọng và đã nhận được nhiều sự quan tâm của các nhà nghiên cứu trên thế giới. Một tập dữ liệu được gọi là mất cân bằng khi số lượng phần tử thuộc về một nhãn lớp bé hơn nhiều so với các nhãn lớp khác. Trong phạm vi bài báo này chúng tôi chỉ đề cập đến bài toán phân loại hai lớp. Trong trường hợp đó, lớp có số lượng phần tử ít hơn được gọi là lớp thiểu số và lớp còn lại được gọi là lớp đa số.

Bài toán phân lớp dữ liệu mất cân bằng là một bài toán phổ biến trong thực tế, nhằm phát hiện các đối tượng hiếm nhưng quan trọng, chẳng hạn như bài toán phát hiện gian lận, phát hiện vị trí tràn dầu trên biển dựa vào ảnh chụp vệ tinh, các bài toán trong lĩnh vực tin sinh học như bài toán dự đoán cấu trúc protein, dự đoán tương tác giữa protein-protein, phân lớp microRNA…, cũng như các bài toán chẩn đoán bệnh trong y học. Trong một số trường hợp, tỷ lệ giữa các phần tử thuộc lớp thiểu số so với các phần tử thuộc lớp đa số có thể lên đến 1:100 hoặc 1:100,000 [1].

Khi áp dụng các thuật toán phân lớp truyền thống lên các tập dữ liệu mất cân bằng, đa số các phần tử thuộc lớp đa số sẽ được phân lớp đúng và các phần tử thuộc lớp thiểu số cũng sẽ được gán nhãn lớp là nhãn lớp của lớp đa số. Điều này dẫn đến kết quả là accuracy (độ chính xác) của việc phân lớp rất cao trong khi giá trị sensitivity (độ nhạy) lại rất thấp.

Nhiều phương pháp đã được đề xuất để giải quyết vấn đề này và được phân thành hai nhóm cơ bản: tiếp cận ở mức giải thuật và tiếp cận ở mức dữ liệu. Các phương pháp tiếp cận ở mức giải thuật hướng tới việc điều chỉnh các thuật toán phân lớp cơ bản để vẫn có hiệu quả cao trên các tập dữ liệu mất cân bằng như phương pháp điều chỉnh xác suất ước lượng [2], hay sử dụng các hằng số phạt khác nhau cho các nhãn lớp khác nhau [3],

Tạp chí Khoa học và Giáo dục, Trường Đại học Sư phạm Huế
ISSN 1859-1612, Số 02(42)/2017, tr. 101-108
Ngày nhận bài: 05/12/2016; Hoàn thành phản biện: 20/12/2017; Ngày nhận đăng: 13/3/2017

[4]... Các phương pháp tiếp cận ở mức dữ liệu nhắm tới thay đổi sự phân bố các đối tượng bằng cách sinh thêm các phần tử cho lớp thiểu số như SMOTE [5], OSD [6]... hay giảm bớt các phần tử thuộc lớp đa số để làm giảm sự mất cân bằng giữa các lớp đối tượng. Nhiều nghiên cứu 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 [1].

Sinh phần tử ngẫu nhiên (Random Oversampling) là phương pháp sinh thêm phần tử đơn giản nhất bằng cách tăng số lượng một số phần tử được chọn ngẫu nhiên thuộc lớp thiểu số để cân bằng tỷ lệ. Tuy nhiên, kỹ thuật này có nhược điểm là dễ dẫn đến tình trạng quá khớp với dữ liệu huấn luyện (overfitting). Ngoài ra, nếu tập dữ liệu có kích thước lớn thì chi phí thời gian và bộ nhớ cho giai đoạn phân lớp sẽ gia tăng đáng kể.

Trái lại, phương pháp Giảm số phần tử ngẫu nhiên (Random Undersampling) sẽ chọn ngẫu nhiên và loại bỏ một số phần tử thuộc lớp đa số để làm giảm tỷ lệ mất cân bằng của các tập dữ liệu. Phương pháp này tuy tốn ít chi phí về thời gian cũng như bộ nhớ cho quá trình phân lớp nhưng lại dễ làm mất các thông tin quan trọng của lớp đa số.

Trong bài báo này, chúng tôi đề xuất một phương pháp làm giảm số phần tử thuộc lớp đa số mới nhắm tới xử lý các đối tượng khó phân lớp và khắc phục nhược điểm đã đề cập.

2. ĐỘ ĐO ĐÁNH GIÁ HIỆU SUẤT PHÂN LỚP

Do các tập dữ liệu là không cân bằng, việc sử dụng độ đo accuracy làm cơ sở để đánh giá hiệu suất phân lớp sẽ không thể hiện được hết yêu cầu đặt ra là dự đoán cả hai nhãn lớp cần đạt được độ chính xác cao. Vì vậy, các độ đo khác thích hợp hơn thường được sử dụng làm độ đo hiệu suất của việc phân lớp, như:

$$\text{Sensitivity} = \text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{1}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \tag{2}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{3}$$

$$\text{F} - \text{measure} = \frac{(1+\beta^2).\text{Precision.Recall}}{\beta^2.\text{Precision+Recall}} \tag{4}$$

Trong đó, β là hệ số điều chỉnh mối quan hệ giữa Precision với Recall và thông thường β =1. F-measure thể hiện sự tương quan hài hòa giữa Precision và Recall. Giá trị của F-measure cao khi cả Precision và Recall đều cao.

G-mean là sự kết hợp của Sensitivity và Specificity, được tính bởi công thức:

$$\text{G} - \text{mean} = \sqrt{\text{Sensitivity}\times\text{Specificity}} \tag{5}$$

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

Trong phạm vi bài báo này, chúng tôi sử dụng F-measure và G-mean làm độ đo chính để đánh giá hiệu suất của sự phân lớp.

3. PHÂN LOẠI LỀ

Lề (margin), đóng vai trò quan trọng trong lĩnh vực học máy, thể hiện tính hiệu quả khi phân lớp của bộ phân lớp (classifier).

Có hai cách xác định giá trị lề cho một phần tử dựa trên quy tắc phân lớp [7]. Cách thứ nhất là đo khoảng cách từ phần tử đang xét tới biên quyết định được xác định bởi bộ phân lớp và lề trong trường hợp này gọi là lề phần tử (sample margin). Đối với cách thứ hai, lề là khoảng cách mà bộ phân lớp có thể di chuyển sao cho không làm thay đổi nhãn lớp của các phần tử đã được xác định, và được gọi là lề giả thuyết (hypothesis margin).

Trong trường hợp sử dụng bộ phân lớp láng giềng gần nhất, các kết quả sau đây đã được chứng minh là đúng [8]:

1. Lề giả thuyết là giới hạn dưới của lề phần tử.

2. Lề giả thuyết của phần tử x trong tập dữ liệu A được tính bởi công thức:

$$\theta_A = \frac{1}{2}(\|x - \text{nearestmiss}_A(x)\| - \|x - \text{nearesthit}_A(x)\|) \quad (6)$$

trong đó: $\text{nearesthit}_A(x)$ là phần tử gần nhất có cùng nhãn lớp với x trong A.

$\text{nearestmiss}_A(x)$ là phần tử gần nhất khác nhãn lớp với x trong A.

Từ đó có thể suy ra, nếu một tập các phần tử có giá trị lề giả thuyết lớn thì giá trị lề phần tử tương ứng của nó cũng lớn.

Do đó, chúng ta có thể áp dụng kết luận này vào bài toán xử lý dữ liệu mất cân bằng bằng phương pháp làm giảm bớt phần tử.

Giả sử phần tử x thuộc lớp đa số N được chọn để loại bỏ, lúc đó, lề giả thuyết của các phần tử y trong tập dữ liệu A sẽ là:

$$\theta_{A\setminus\{x\}}(y) = \frac{1}{2}\left(\|y - \text{nearestmiss}_{A\setminus\{x\}}(y)\| - \|y - \text{nearesthit}_{A\setminus\{x\}}(y)\|\right), \forall y \neq x$$

Ở đây, $\text{nearestmiss}_A(y)$, $\text{nearesthit}_A(y)$ lần lượt là phần tử gần nhất khác nhãn lớp và phần tử gần nhất cùng nhãn lớp của y trên tập A.

Nếu y_p thuộc vào lớp thiểu số P, thì:

$$\|y_p - \text{nearesthit}_{A\setminus\{x\}}(y_p)\| = \|y_p - \text{nearesthit}_A(y_p)\|$$

Và $\|y_p - \text{nearestmiss}_{A\setminus\{x\}}(y_p)\| \geq \|y_p - \text{nearestmiss}_A(y_p)\|$

Do đó: $\theta_{A\setminus\{x\}}(y_p) \geq \theta_A(y_p)$.

Tương tự, với y_n là phần tử thuộc lớp đa số N, $y_n \neq x$, ta có:

$$\|y_n - \text{nearesthit}_{A\setminus\{x\}}(y_n)\| \geq \|y_n - \text{nearesthit}_A(y_n)\|$$

và $\|y_n - \text{nearestmiss}_{A\setminus\{x\}}(y_n)\| = \|y_n - \text{nearestmiss}_A(y_n)\|$

Nên: $\theta_{A\setminus\{x\}}(y_n) \leq \theta_A(y_n)$.

Điều này có nghĩa rằng việc loại bỏ đi một phần tử thuộc lớp đa số làm tăng giá trị lề của các phần tử lớp thiểu số và giảm giá trị lề của phần tử thuộc lớp đa số. Do đó, nếu các phần tử được chọn để loại bỏ có lề lớn hơn các phần tử còn lại sẽ làm tăng khả năng phân lớp sai của bộ phân lớp. Hay nói cách khác, việc chọn các phần tử có giá trị lề giả thuyết bé nhất thay vì chọn một cách ngẫu nhiên để loại bỏ sẽ làm tăng hiệu suất của việc phân lớp.

4. PHƯƠNG PHÁP LÀM GIẢM PHẦN TỬ DỰA VÀO GIÁ TRỊ LỀ GIẢ THUYẾT

Dựa vào ý tưởng ở phần trên, chúng tôi đề xuất một phương pháp mới để xử lý bài toán phân lớp dữ liệu mất cân bằng là phương pháp làm giảm phần tử dựa vào giá trị lề giả thuyết, đặt tên là Hypothesis Margin based Undersampling (HMU). Phương pháp này ưu tiên chọn các phần tử có giá trị lề bé nhất để loại bỏ trước tiên nhằm tạo ra một tập dữ liệu dễ phân lớp hơn. Thuật toán được mô tả như sau:

HMU Algorithm
***Input**: lớp đa số **N**; số lượng phần tử cần loại bỏ **d**;*
Output**: lớp đa số sau khi đã làm giảm số phần tử **N**;

Begin

1. *nos = |N|- d*
2. *N* = N*
3. *while (|N*| > nos)*
4. *tính giá trị lề mar(x) của tất cả các phần tử x thuộc N* trên toàn bộ tập dữ liệu và lưu vào mảng @margin*
5. *sắp xếp mảng @margin*
6. *loại bỏ phần tử có giá trị lề tương ứng bé nhất trong mảng @margin*
7. *cập nhật lại N**
8. *end while*

End

Lề của các phần tử lớp đa số được tính dựa vào công thức (6).

Kích thước của lớp đa số sau khi làm giảm bớt số phần tử N* được xác định dựa vào số lượng phần tử cần loại bỏ d. Chỉ số d này phụ thuộc vào từng tập dữ liệu cụ thể.

Khoảng cách được sử dụng để xác định lề trong thuật toán này là khoảng cách Euclidean.

5. ĐÁNH GIÁ HIỆU SUẤT THUẬT TOÁN

Để đánh giá hiệu suất của quá trình phân lớp, chúng tôi tiến hành thực nghiệm trên 4 tập dữ liệu UCI [9] là Balance, Cmc, Haberman và Pima. Thông tin về số lượng thuộc tính, số phần tử, tỷ lệ mất cân bằng (số phần tử tập thiểu số:số phần tử tập đa số) của mỗi tập dữ liệu được mô tả ở Bảng 1. Tất cả các tập dữ liệu đều được chuẩn hóa bằng hàm normalize của gói lệnh SOM trong R trước khi tiến hành điều chỉnh tỷ lệ mất cân bằng cũng như phân lớp.

Bảng 1. *Các tập dữ liệu UCI*

Tập dữ liệu	**Số thuộc tính**	**Số phần tử**	**Tỷ lệ mất cân bằng**
Balance	4	625	1:11.75
Cmc	9	1473	1:3.42
Haberman	3	306	1:2.78
Pima	8	768	1:1.87

Sử dụng gói lệnh kernlab [10] trong R, chúng tôi tiến hành phân lớp 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 (KSVM), kết quả phân lớp có sử dụng thuật toán giảm số phần tử ngẫu nhiên (RUS) với kết quả có sử dụng thuật toán HMU nhằm đánh giá tính hiệu quả của thuật toán này.

Quá trình phân lớp được thực hiện như sau:

- Máy vector hỗ trợ (Support Vector Machine - SVM) với hàm nhân Gaussian RBF được sử dụng làm bộ phân lớp chính.

- Với mỗi tập dữ liệu, chúng tôi thực hiện mười lần 10-fold cross-validation (kiểm chứng chéo), nghĩa là:

Với mỗi lần thực hiện 10-fold cross-validation:

+ Tập dữ liệu được chia ngẫu nhiên thành 10 phần bằng nhau.

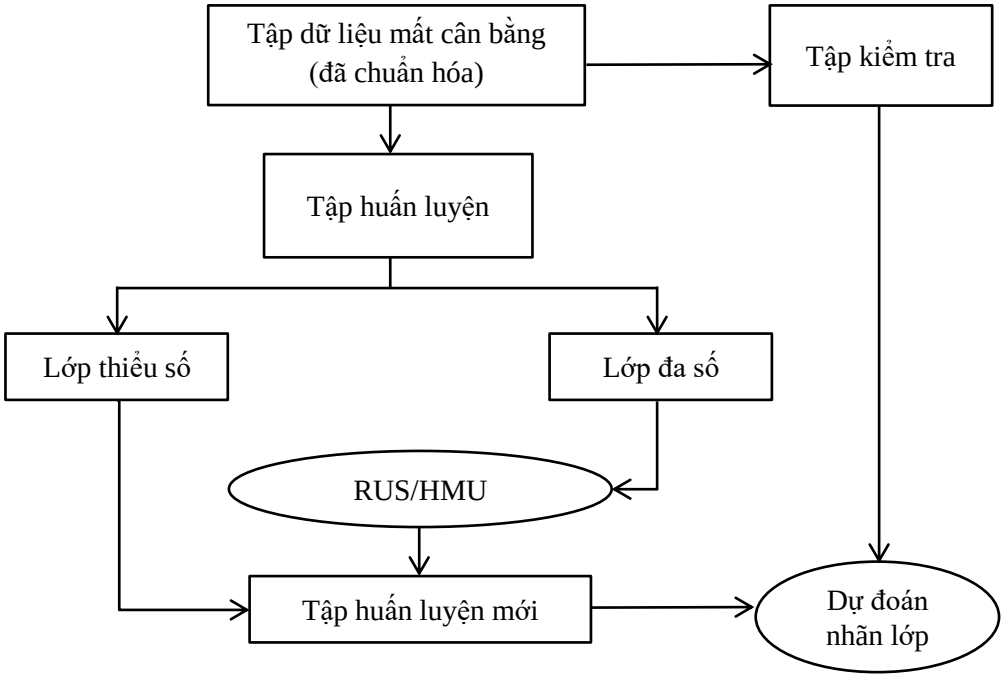
+ Lần lượt mỗi phần trong mười phần đó được chọn làm tập kiểm tra, chín phần còn lại tạo nên tập huấn luyện để xây dựng mô hình phân lớp. Với mỗi bộ tập kiểm tra và tập huấn luyện như thế, chúng tôi thu được các giá trị độ đo đánh giá hiệu suất tương ứng dựa trên số lượng các phần tử được phân lớp đúng và phân lớp sai của tập kiểm tra.

+ Kết quả của thu được từ mười bộ tập kiểm tra và huấn luyện chính là kết quả của một lần thực hiện 10-fold cross-validation.

Cuối cùng, các giá trị độ đo đánh giá hiệu suất Sensitivity, Specificity, F-measure và G-mean được tính bằng cách lấy giá trị trung bình cộng của mười lần thực hiện độc lập này.

- Sau khi áp dụng các thuật toán điều chỉnh tỷ lệ mất cân bằng, các tập dữ liệu mới có tỷ lệ số phần tử lớp thiểu số:lớp đa số là xấp xỉ 1:1.

Toàn bộ quá trình phân lớp đánh giá hiệu suất thuật toán được mô tả như ở Hình 1 bên dưới.



Hình 1. *Quá trình thực nghiệm phân lớp dữ liệu*

Kết quả đánh giá quá trình phân lớp các tập dữ liệu bằng các phương pháp được trình bày trong bảng 2 và bảng 3. Kết quả này cho thấy khi sử dụng thuật toán làm giảm phần tử để xử lý sự mất cân bằng trong dữ liệu, hiệu suất của quá trình phân lớp tăng lên đáng kể. Đồng thời kết quả cũng cho thấy thuật toán HMU đã hoạt động hiệu quả, cải thiện các giá trị F-measure và G-mean trong hầu hết các trường hợp.

Ví dụ, đối với Balance, tập dữ liệu có tỷ lệ mất cấn bằng khá lớn (1:11.75), KSVM hoàn toàn không phân lớp đúng một đối tượng thuộc lớp thiểu số nào (Sensitivity = 0% và Specificity = 100%). Khi áp dụng phương pháp giảm số phần tử ngẫu nhiên RUS, các giá trị Sensitivity, F-measure và G-mean đều tăng đáng kể (lần lượt tăng 77.76%, 18.60%, 58.47%). Tuy nhiên, các giá trị này vẫn thấp hơn so với khi áp dụng thuật toán HMU, cụ thể Sensitivity tăng 22.04%, F-measure tăng 2.29%, G-mean tăng 1.23% so với RUS.

Tương tự đối với hai tập dữ liệu Haberman và Cmc, phương pháp sử dụng thuật toán HMU cho các giá trị Sensitivity, F-measure và G-mean lớn hơn so với hai phương pháp còn lại. Còn đối với tập dữ liệu Pima, giá trị G-mean và F-measure của HMU là xấp xỉ so với RUS (F-measure nhỏ hơn 0.17%, G-mean lớn hơn 0.06%).

Bảng 2. *Kết quả phân lớp theo độ đo Sensitivity (%) và Specificity (%)của các tập dữ liệu UCI*

Tập dữ liệu	**Sensitivity (%)**			**Specificity (%)**		
	KSVM	**RUS**	**HMU**	**KSVM**	**RUS**	**HMU**
Balance	0.00	77.76	**99.80**	**100.00**	43.98	35.71
Haberman	23.46	52.22	**63.33**	**93.02**	76.04	71.73
Cmc	6.04	66.28	**75.35**	**98.62**	65.07	57.81
Pima	56.16	75.34	**78.06**	**87.74**	71.90	69.06

Bảng 3. *Kết quả phân lớp theo độ đo F-measure (%) và G-mean (%) của các tập dữ liệu UCI*

Tập dữ liệu	**F-measure (%)**			**G-mean (%)**		
	KSVM	**RUS**	**HMU**	**KSVM**	**RUS**	**HMU**
Balance	-	18.60	**20.89**	0.00	58.47	**59.70**
Haberman	32.85	47.75	**52.38**	46.71	63.01	**67.39**
Cmc	10.90	46.37	**47.12**	24.40	65.67	**66.00**
Pima	62.73	66.15	**66.21**	70.19	**73.59**	73.42

6. KẾT LUẬN

Phân lớp dữ liệu mất cân bằng là một bài toán quan trọng và được ứng dụng vào nhiều lĩnh vực khác nhau trong thực tế. Một trong những kỹ thuật nâng cao hiệu suất của bài toán này là sử dụng phương pháp làm giảm phần tử của lớp đa số. Trong bài báo này, chúng tôi đã trình bày một thuật toán làm giảm phần tử lớp đa số mới sử dụng giá trị lề giả thuyết và tiến hành các thực nghiệm để so sánh, đánh giá hiệu suất của thuật toán trên bốn tập dữ liệu chuẩn UCI. Kết quả thực nghiệm đã cho thấy rằng thuật toán do chúng tôi đề xuất có hiệu quả trên bốn tập dữ liệu này dựa trên các giá trị độ đo đánh giá hiệu suất Sensitivity, Specificity, F-measure và G-mean. Phương pháp này có thể kết hợp với một số các kỹ thuật khác như lựa chọn đặc trưng hoặc các phương pháp làm tăng phần tử để cho kết quả tốt hơn, đặc biệt là đối với các tập dữ liệu có kích thước lớn.

TÀI LIỆU THAM KHẢO

[1] H. He and E. A. Garcia (2009). “Learning from Imbalanced Data”. *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, 1263–1284.

[2] B. Zadrozny and C. Elkan (2001). *Learning and making decisions when costs and*

probabilities are both unknown, Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining - KDD'01, 204–213.

[3] R. Akbani, S. Kwek, and N. Japkowicz (2004). *Applying support vector machines to imbalanced datasets*, Proceedings of the 15th European Conference on Machine Learning, 39–50.

[4] X. Yang, Q. Song, and A. Cao (2005). *Weighted support vector machine for data classification*, Proceedings of the International Joint Conference on Neural Networks, vol. 2, 859–864.

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

[6] L. A. T. Nguyen *et al.* (2013). "Improving the Prediction of Protein-Protein Interaction Sites Using a Novel Over-Sampling Approach and Predicted Shape Strings". *Annu. Rev. Res. Biol.*, vol. 3, no. 2, 92–106.

[7] K. Crammer, R. Gilad-Bachrach, A. Navot, and N. Tishby (2002). "Margin Analysis of The Lvq Algorithm". *Neural Inf. Process. Syst.*, 462–469.

[8] R. Gilad-bachrach (2004). *Margin Based Feature Selection - Theory and Algorithms*, Proceedings of the 21st International Conference on Machine Learning, Banff, Canada.

[9] M. Lichman (2013). UCI Machine Learning Repository, *http://archive.ics.uci.edu/ml*, Irvine, CA: University of California, School of Information and Computer Science.

[10] A. Karatzoglou, T. U. Wien, A. Smola, K. Hornik, and W. Wien (2004). "kernlab – An S4 Package for Kernel Methods in R". *J. Stat. Softw.*, vol. 11, no. 9,1–20.

Title: HYPOTHESIS MARGIN BASED UNDERSAMPLING METHOD FOR DEALING WITH IMBALANCED DATA SETS

Abstract: Classifying the imbalanced data sets is one of the important issues. Many approaches were developed to handle this problem. In this paper, we present a novel Undersampling algorithm basing on the Hypothesis margin of samples to enhance the result of the imbalanced data sets classification.

Keywords: Imbalanced data, Undersampling, Hypothesis margin