CLASSIFYING IRIS FLOWER DATA USING ALGORITHMS NAÏVE BAYES, RANDOM FOREST AND KNN

Nguyen Van Nui
TNU - University of Information and Communication Technology

ARTICLE INFO

Received: 03/6/2021
Revised: 02/7/2021
Published: 14/7/2021

KEYWORDS

Data classifying
Naïve Bayes
Random Forest
KNN
Iris
Iris flower

ABSTRACT

Iris is a beautiful flower, representing luck and love courage, loyalty, and wisdom. Therefore, the classification and accurate prediction of Iris flower brings many important meanings in practice. Although there have been many scientific publications related to classification and prediction of Iris flowers, the classification and prediction performance of these publications still have certain limitations that need to be studied for further improvement. In this paper, the author proposes model to classify and predict Iris flowers on the basis of the application of the Weka toolkit and the Naïve Bayes, Random Forest and KNN algorithms. The results reveal that all three algorithms above give high accuracy (over 95%), so it is suitable for building model to classify Iris flowers. However, the two algorithms, Random Forest and KNN (k=3), show better stability and objectivity than the Naïve Bayes algorithm.

PHÂN LỚP DỮ LIỆU HOA IRIS SỬ DỤNG CÁC THUẬT TOÁN NAÏVE BAYES, RANDOM FOREST VÀ KNN

Nguyễn Văn Núi
Trường Đại học Công nghệ Thông tin và Truyền thông – ĐH Thái Nguyên

THÔNG TIN BÀI BÁO

Ngày nhận bài: 03/6/2021
Ngày hoàn thiện: 02/7/2021
Ngày đăng: 14/7/2021

TỪ KHÓA

Phân lớp dữ liệu
Naïve Bayes
Random Forest
KNN
Iris
Hoa Diên Vĩ

TÓM TẮT

Iris (hoa Diên Vĩ) là một loài hoa đẹp, đại diện cho sự may mắn, tình yêu, lòng dũng cảm, trung thành và sự khôn ngoan. Vì vậy việc phân lớp, dự đoán chính xác loài hoa Iris mang lại nhiều ý nghĩa quan trọng trong thực tiễn. Mặc dù đã và đang có rất nhiều công bố khoa học liên quan đến phân lớp, dự đoán loài hoa Iris, tuy nhiên hiệu năng phân lớp, dự đoán của những công bố này vẫn còn tồn tại những hạn chế nhất định cần được nghiên cứu để cải thiện hơn nữa. Trong bài báo này, tác giả đề xuất mô hình phân lớp dữ liệu, dự đoán hoa Iris trên cơ sở ứng dụng bộ công cụ Weka và các thuật toán Naïve Bayes, Random Forest và KNN. Kết quả cho thấy cả 3 thuật toán trên đều cho độ chính xác cao (trên 95%), vì vậy phù hợp để sử dụng cho việc xây dựng mô hình phân lớp dự đoán hoa Iris. Tuy nhiên, 2 thuật toán Random Forest và KNN (k=3) thể hiện sự ổn định và có tính khách quan tốt hơn so với thuật toán Naïve Bayes.

DOI: https://doi.org/10.34238/tnu-jst.4594

Email: nvnui@ictu.edu.vn

1. Giới thiệu chung

Iris (hoa Diên Vĩ) là một loài hoa được rất nhiều người yêu thích hiện nay (Hình 1). Trong văn hóa châu Âu, Diên Vĩ được xem là loài hoa đại diện của lòng dũng cảm, trung thành và sự khôn ngoan. Vì vậy, loài hoa này được chọn làm biểu tượng của nhiều gia đình hoàng tộc tại châu Âu. Không chỉ vậy, hoa Diên Vĩ còn được xem là loài hoa của sự may mắn và tình yêu. Do có giá trị cao về mặt truyền thống và kinh tế nên việc phân lớp, dự đoán chính xác loài hoa Iris mang lại nhiều ý nghĩa quang trọng trong thực tiễn.

Cùng với sự bùng nổ mạnh mẽ của công nghệ thông tin và trí tuệ nhân tạo như hiện nay, số lượng các nghiên cứu liên quan đến khai phá phát hiện tri thức nói chung; các phương pháp học máy, "tri thức con người" nói riêng, đang ngày càng tăng lên một cách mạnh mẽ. Trong số rất nhiều bài toán thực tế hiện nay; bài toán phân lớp, dự đoán loài hoa Iris cũng là một vấn đề cần được quan tâm nhất bởi ý nghĩa, giá trị rất thiêng liêng và to lớn của loài hoa này.

Trong những năm gần đây, có rất nhiều nhóm nghiên cứu về bài toán phân lớp, dự đoán. Đến nay, có rất nhiều công trình nghiên cứu sử dụng thuật toán học máy, trí tuệ nhân tạo đã được áp dụng thành công cho bài toán phân lớp, dự đoán [1]-[7]. JP Pinto và các cộng sự [1] đã đề xuất, áp dụng một số thuật toán phân lớp và hồi quy, ứng dụng cho bài toán phân lớp, dự đoán hoa Diên Vĩ. Năm 2011, Cao Thăng [5] đã công bố tài liệu một số ví dụ phân loại dùng SOM và MLP Neural Network. Trong nghiên cứu này, tác giả có đề cập đến bài toán phân lớp dự đoán hoa Diên Vĩ sử dụng SOM (Self-Organizing Map) và MLP (Multilayer Perceptron) Neural Network, …

Hình 1. *Iris Flower (hoa Diên Vĩ)*

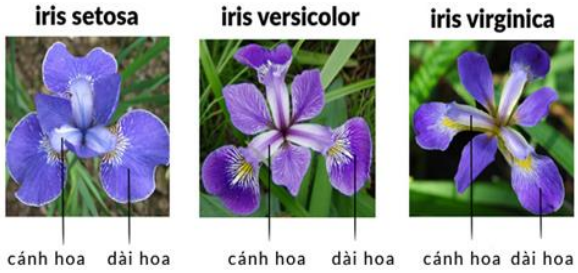
2. Xây dựng, huấn luyện mô hình

2.1. Thu thập, tiền xử lý dữ liệu

Tập dữ liệu hoa Iris hoặc tập dữ liệu của Fisher là tập dữ liệu đa biến được giới thiệu bởi nhà thống kê và nhà sinh vật học người Anh Ronald Fisher trong bài báo năm 1936 [8]. Việc sử dụng nhiều phép đo trong các bài toán phân loại như một ví dụ về phân tích phân biệt tuyến tính. Đôi khi nó được gọi là tập dữ liệu Iris của Anderson [900, vì Edgar Anderson đã thu thập dữ liệu để định lượng sự biến đổi hình thái của hoa Iris của ba loài liên quan [9].

Bộ dữ liệu bao gồm 150 mẫu (bản ghi) từ 3 loài Iris (Iris Setosa, Iris virginica và Iris versicolor), được thu thập từ kho dữ liệu học máy UCI [10]. Bốn đặc điểm được đo từ mỗi mẫu gồm: chiều dài và chiều rộng của đài hoa, chiều dài và chiều rộng của cánh hoa, tính bằng centimet. Dựa trên sự kết hợp của bốn đặc điểm này, Fisher dã phát triển một mô hình phân biệt tuyến tính để phân biệt các loài với nhau.

Bộ dữ liệu sau khi được rút gọn bao gồm 5 thuộc tính: Tên của loài hoa Iris (Iris Setosa, Iris Versicolour, Iris Virginica), chiều dài đài hoa, chiều rộng đài hoa, chiều dài cánh hoa, chiều rộng cánh hoa (Hình 2).



Hình 2. *Thông tin thuộc tính hoa Diên Vĩ*

Sau một số bước kỹ thuật tiền xử lý dữ liệu, bộ dữ liệu cuối cùng được sử dụng cho nghiên cứu này có thông tin thống kê chung về giá trị các thuộc tính (chiều dài, chiều rộng đài hoa; chiều dài, chiều rộng cánh hoa) được thể hiện ở Bảng 1.

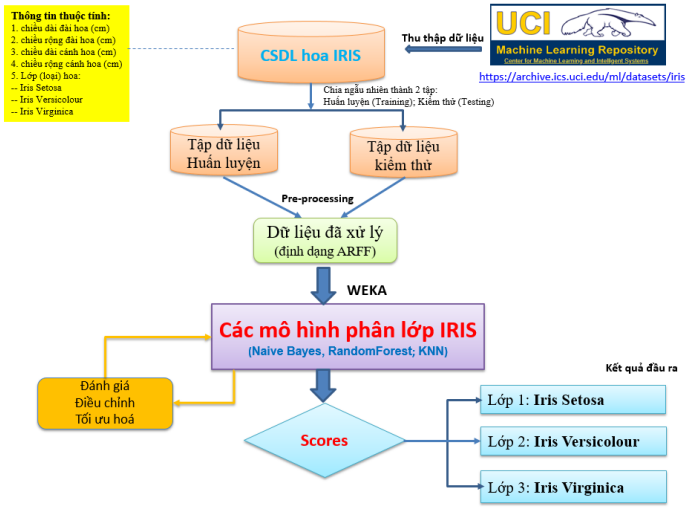
Bảng 1. *Giá trị trung bình đài hoa, cánh hoa*

Thuộc tính	Giá trị MIN	Giá trị MAX	Giá trị TB
Chiều dài đài hoa	4,3	7,9	5,84
Chiều rộng đài hoa	2,0	4,4	3,05
Chiều dài cánh hoa	1,0	6,9	3,76
Chiều rộng cánh hoa	0,1	2,5	1,20

2.2. Xây dựng và huấn luyện mô hình

Trong bài báo này, mô hình phân lớp dự đoán hoá Iris được xây dựng và huấn luyện trên cơ sở sử dụng bộ công cụ Weka; các thuật toán được sử dụng gồm có: Naïve Bayes, Random Forest và KNN.

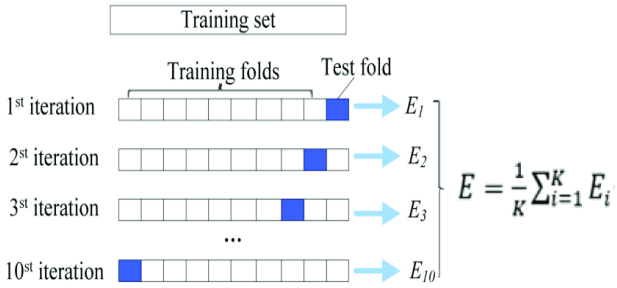
Mô hình tổng thể phân lớp dự đoán hoa Iris đề xuất trong bài báo này được thể hiện chi tiết ở Hình 3 bên dưới.



Hình 3. *Sơ đồ tổng thể phân lớp dự đoán hoa Iris*

Để đánh giá hiệu năng của mô hình, 2 phương pháp phổ biến được sử dụng đó là: đánh giá chéo 10 mặt (10-fold cross-validation) và kiểm thử độc lập (Independent testing) sử dụng bộ dữ liệu riêng biệt, độc lập với bộ dữ liệu huấn luyện (training dataset) [1]-[7], [11]-[14].

Theo phương pháp đánh giá chéo 10 mặt (10-fold cross-validation), tập dữ liệu huấn luyện sẽ được chia ngẫu nhiên thành 10 tập con bằng nhau, lần lượt mỗi tập con sẽ được dùng cho vai trò kiểm thử, trong khi 9 tập còn lại được dùng làm dữ liệu huấn luyện (Hình 4).



Hình 4. *Mô hình đánh giá kiểm tra chéo 10 mặt*

Các đại lượng thông dụng được sử dụng để đo lường và đánh giá hiệu năng của mô hình bao gồm: Accuray (độ chính xác), MCC (hệ số tương quan Matthews và Error Rate [6]-[12].

$$ACC = \frac{TP+TN}{P+N}; \qquad Error\ Rate = \frac{FP+FN}{P+N}$$

$$MCC = \frac{(TP\times TN)-(FN\times FP)}{\sqrt{(TP+FN)\times(TN+FP)(TP+FP)(TN+FN)}}$$

Trong đó:

P: Số bản ghi Positive trong tập dữ liệu.

N: Số bản ghi Negative trong tập dữ liệu.

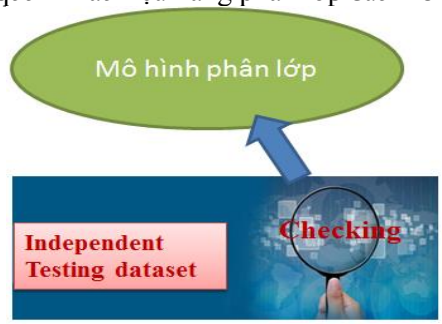
TP: Số bản ghi Positive được dự đoán là Positive.

TN: Số bản ghi Negative được dự đoán là Negative.

FP: Số bản ghi Negative được dự đoán là Positive.

FN: Số bản ghi Positive được dự đoán là Negative.

Ngoài ra, phương pháp kiểm thử, đánh giá độc lập cũng được sử dụng để đánh giá hiệu năng của mô hình phân lớp, dự đoán. Như hiển thị ở Hình 5, theo phương pháp đánh giá kiểm thử độc lập, hiệu năng của mô hình sẽ được xác định bằng việc sử dụng một bộ dữ liệu kiểm thử hoàn toàn khác biệt và không trùng lặp với bộ dữ liệu huấn luyện đã dùng cho việc huấn luyện mô hình (Independent testing dataset). Việc sử dụng bộ dữ liệu kiểm thử độc lập này sẽ giúp ta kiểm tra, đánh giá một cách khách quan nhất hiệu năng phân lớp của mô hình.



Hình 5. *Mô hình kiểm thử độc lập*

3. Kết quả và một số thảo luận

3.1. Kết quả huấn luyện và đánh giá mô hình phân lớp theo phương pháp đánh giá chéo 10 mặt

Như đã trình bày trước đó, trong nghiên cứu này, tác giả tiến hành sử dụng kết hợp thuật toán của máy vector hỗ trợ và bộ công cụ Weka để xây dựng mô hình phân lớp dự đoán hoa Iris.

Trong bài báo này, tác giả lựa chọn phương pháp đánh giá chéo 10 mặt (10-fold cross-validation) để đánh giá hiệu năng của mô hình phân lớp, dự đoán. Theo thông tin tổng hợp ở Bảng 2, cả 3 thuật toán Naïve Bayes, Random Forest và KNN (k=3) đều có độ chính xác cao, đạt trên 95%. Trong đó, thuật toán Naïve Bayes thể hiện là tốt nhất cho bài toán phân lớp dự đoán hoa Diên Vĩ, với độ chính xác đạt 96,0% và tỉ lệ lỗi chỉ ở mức 4,0%.

Bảng 2. *Kết quả đánh giá mô hình bằng phương pháp đánh giá chéo 10 mặt*

Thuật toán	Accuracy	Recall	MCC	Error Rate
Naïve Bayes	96,0%	96%	0,94	4,0%
Random Forest	95,3%	95,3%	0,93	4,6%
KNN (k=3)	95,3%	95,3%	0,93	4,6%

3.2. Kết quả đánh giá mô hình sử dụng phương pháp kiểm thử độc lập

Như đã đề cập trước đó, phương pháp đánh giá độc lập giúp kiểm chứng khả năng thực nghiệm của mô hình trong trường hợp thực tế, khách quan nhất. Để thực hiện được việc này, một bộ dữ liệu kiểm thử độc lập đã được xây dựng bao gồm 50 bản ghi.

Hiệu năng của mô hình đánh giá bởi phương pháp kiểm thử độc được thể hiện chi tiết ở Bảng 3. Rất may mắn, kết quả cho thấy cả 3 thuật toán cũng đều đạt kết quả tốt với độ chính xác trên 94%. Tuy nhiên, thông qua Bảng 2 và Bảng 3, ta có thể thấy rằng, 2 thuật toán Random Forest và KNN (k=3) có độ chính xác khi đánh giá bởi phương pháp đánh giá chéo 10 mặt thấp hơn so với kết quả đánh giá bởi phương pháp kiểm thử độc lập. Điều này cho thấy, với bài toán phân lớp dự đoán hoa Diên Vĩ này, 2 thuật toán Random Forest và KNN (k=3) có sự ổn định tốt hơn thuật toán Naïve Bayes.

Bảng 3. *Kết quả đánh giá mô hình bằng phương pháp kiểm thử độc lập*

Thuật toán	Accuracy	Recall	MCC	Error Rate
Naïve Bayes	94,1%	94,1%	0,91	5,9%
Random Forest	96,1%	96,1%	0,94	3,9%
KNN (k=3)	96,1%	96,1%	0,94	3,9%

4. Kết luận

Hoa Diên vĩ là một loài hoa có ý nghĩa và giá trị rất lớn cả về vật chất và tinh thần. Do đó, bài toán phân lớp, dự đoán chính xác loài hoa Iris có ý nghĩa khoa học và mang thực tiễn cao trong cuộc sống. Trong bài báo này, tác giả đề xuất cách tiếp cận sử dụng kết hợp các thuật toán Naïve Bayes, Random Forest, KNN và bộ công cụ Weka để xây dựng, huấn luyện mô hình hỗ trợ cho bài toán phân lớp dự đoán loài hoa Diên Vĩ. Kết quả cho thấy, việc kết hợp bộ công cụ Weka và các thuật toán trên cho thấy sự phù hợp trong việc phân lớp dự đoán hoa Iris. Các thuật toán đều cho kết quả phân lớp dự đoán khá tốt, với độ chính xác đạt trên 95%. Tuy nhiên, hai thuật toán Random Forest và KNN (k=3) thể hiện sự ổn định và có tính khách quan tốt hơn so với thuật toán Naïve Bayes.

Lời cảm ơn

Tác giả xin được bày tỏ lòng biết ơn đến Trường Đại học Công nghệ thông tin và Truyền thông đã hỗ trợ một phần tài chính cho nghiên cứu này theo đề tài cấp cơ sở mã số: T2021-07-02.

TÀI LIỆU THAM KHẢO/ REFERENCES

[1] J. P. Pinto, S. Kelur, and J. Shetty, "Iris Flower Species Identification Using Machine Learning Approach," 2018 4th International Conference for Convergence in Technology (I2CT), SDMIT Ujire, Mangalore, India. Oct 27-28, 2018.

[2] M. Swain, S. K. Dash, S. Dash, and A. Mohapatra, "An approach for Iris Plant Classification Using Neural Network," *International Journal on Soft Computing (ÍC)*, vol. 3, no. 1, pp. 79-89, February 2012.

[3] C. Geetha, R. Ram, and N. Vali, "Iris-flower Classification," *Eurasian Journal of Analytical Chemistry*, vol. 12, no. 3, pp. 51-63, 2017.

[4] A. Eldem, H. Eldem, and D. Üstün, *A model of Deep Neural Network for Iris Classification with Different Activation Functions*, 978-1-5386-6878-8/18/$31.00 ©2018 IEEE, 2018.

[5] T. Cao, Some examples of classification using SOM and *MLP Neural Network*, July 11, 2013.

[6] T. X. Tran and V. N. Nguyen, "Classifying protein s-farnesylation sites with support vector machine and decision tree," *TNU Journal of Science and Technology*, vol. 204, no. 11, pp. 149-154, 2019.

[7] H. J. Kao, V. N. Nguyen, K. Y. Huang, W. C. Chang, and T. Y. Lee, "SuccSite: Incorporating Amino Acid Composition and Informative k-spaced Amino Acid Pairs to Identify Protein Succinylation Sites," Genomics, Proteomics and Bioinformatics (Q1, SCI, IF: 6.615), June 2020.

[8] R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, vol. 7, pp. 179-188, 1936.

[9] E. Anderson, "The Species Problem in Iris," *Annals of the Missouri Botanical Garden*, vol. 23, no. 3, pp. 457-509, 1936.

[10] D. Dua and C. Graff, *UCI Machine Learning Repository*. Irvine, CA: University of California, School of Information and Computer Science, 2019.

[11] K. Lee and V. N. Nguyen, "SNARE-CNN: a 2D convolutional neural network architecture to identify SNARE proteins from high-throughput sequencing data," *Peer J Computer Science*, vol. 5, 2019, Art. no. e177, doi: https://doi.org/10.7717/peerj-cs.177.

[12] V. N. Nguyen and H. M. Nguyen, "Identification of protein S-Farnesyl cysteine prenylation sites based on substrate specificities," *International Journal of Science and Research (IJSR)*, vol. 7, no. 6, pp. 758-763, June 2018.

[13] V. N. Nguyen, T. X. Tran, H. M. Nguyen, H. T. Nguyen, and T. Y. Lee, "A new schema to identify S-farnesyl cysteine prenylation sites with substrate motifs," in *Advances in Intelligent Systems and Computing ICTA 2016*, in Advances in Information and Communication Technology, vol. 538, Springer, Cham., 2017, doi: 10.1007/978-3-319-49073-1.

[14] V. M. Bui and V. N. Nguyen, "The prediction of Succinylation site in protein by analyzing amino acid composition" in *Advances in Information and Communication Technology. ICTA 2016*, in Advances in Intelligent Systems and Computing, vol. 538, Springer, Cham., doi: 10.1007/978-3-319-49073-1.