JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 24, No. 2: 225~230, Feb. 2020

머신러닝 기반 유클리드 거리를 이용한 붓꽃 품종 분류 재구성

남수태[1] · 신성윤[2] · 진찬용[3*]

A Reconstruction of Classification for Iris Species Using Euclidean Distance Based on a Machine Learning

Soo-Tai Nam[1] · Seong-Yoon Shin[2] · Chan-Yong Jin[3*]

[1]Lecturer, Institute of General Education, Pusan National University, Busan, 46241, Korea
[2]Professor, School of Computer Information & Communication Engineering, Kunsan National University, Kunsan, 54150 Korea
[3*]Professor, Division of Information & Electronic Commerce, Wonkwang University, Iksan, 54538, Korea

요 약

기계학습은 데이터를 기반으로 한 컴퓨터를 학습시켜 컴퓨터 스스로 데이터의 경향성을 파악하게 하여 새로운 입력 데이터의 출력을 예측하도록 하는 알고리즘이다. 기계학습은 크게 지도학습, 비지도학습, 강화학습으로 나눌 수 있다. 지도학습은 데이터에 대한 레이블이 주어진 상태로 기계를 학습시키는 방법이다. 즉, 데이터 및 레이블의 쌍을 통해 해당 시스템의 함수를 추론하는 방법으로 새로운 입력 데이터에 대해서 추론한 함수를 이용하여 결과를 예측한다. 그리고 예측하는 결과 값이 연속 값이면 회귀분석, 예측하는 결과 값이 이산 값이면 분류로 사용된다. 새로운 붓꽃 데이터 Sepal length(5.01)과 Sepal width(3.43)을 이용하여 기초 데이터와 유클리드 거리를 분석하였다. 분석결과, 테이블 3의 8번(5, 3.4, setosa), 27번(5, 3.4, setosa), 41번(5, 3.5, setosa), 44번(5, 3.5, setosa) 그리고 40번(5.1, 3.4, setosa)의 데이터 순으로 유사도가 높은 붓꽃으로 분류되었다. 따라서 이론적 실무적 시사점을 제시하였다.

ABSTRACT

Machine learning is an algorithm which learns a computer based on the data so that the computer can identify the trend of the data and predict the output of new input data. Machine learning can be classified into supervised learning, unsupervised learning, and reinforcement learning. Supervised learning is a way of learning a machine with given label of data. In other words, a method of inferring a function of the system through a pair of data and a label is used to predict a result using a function inferred about new input data. If the predicted value is continuous, regression analysis is used. If the predicted value is discrete, it is used as a classification. A result of analysis, no. 8 (5, 3.4, setosa), 27 (5, 3.4, setosa), 41 (5, 3.5, setosa), 44 (5, 3.5, setosa) and 40 (5.1, 3.4, setosa) in Table 3 were classified as the most similar Iris flower. Therefore, theoretical practical are suggested.

키워드 : 데이터마이닝, 머신러닝, 분류, 지도학습, 회귀분석

Keywords : Data mining, Machine learning, Classification, Supervised learning, Regression analysis

Received 30 September 2019, **Revised** 2 October 2019, **Accepted** 2 November 2019

*** Corresponding Author** Chan-Yong Jin (E-mail:jcy85366@wku.ac.kr, Tel:+82-63-850-6567)
Professor, Division of Information & Electronic Commerce, Wonkwang University, Iksan, 54538, Korea

Open Access **http://doi.org/10.6109/jkiice.2020.24.2.225** print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. INTRODUCTION

Machine learning is a learning of computer algorithms based on data. In addition, it is an algorithm which predicts the output of new input data by summarizing the tendency of the learned data. Machine learning can generally be divided into supervised learning, unsupervised learning, and reinforcement learning [1]. Supervised learning is a way of learning a machine with a given label of data. Therefore, it is a method of inferring a function of the system through a pair of data and a label, and predicting a result using a function inferred on new input data. As a result, if the predicted result value is a continuous value, it is a regression analysis. If the predicted result value is a discrete value, it is used as a classification. The typical regression algorithm is a linear regression algorithm. And, classification algorithms are include decision tree, support sector machine, and k-nearest neighbors, etc. [2, 3]. Unsupervised learning is a way of learning a machine without a label for the data. When data are randomly distributed, clustering algorithms such as k-means, etc. [4], which group similar data together, are typical. Thus, when an agent performs an action in a given state, the environment changes state in response. The resulting reward is sent to the agent. The agent then checks the value and decides the next action. In other words, the agent is a learning method that takes an optimal action after calculation to maximize the rewards it can receive. The goal of reinforcement learning is to use a reward for action to maximize the average of future rewards.

Machine learning is a field of artificial intelligence in computer science, which has evolved from research into pattern recognition and computer learning theory. And, machine learning is a system that learns and makes predictions based on empirical data and improves its own performance, and can be a technology for researching and building algorithms for this purpose. Also, machine learning algorithm is a method of constructing a specific model in order to extract predictions and decisions based on input data, rather than executing strictly defined static program instructions. In addition, machine learning is used in almost all fields including computer science, including computer vision (character recognition, object recognition, face recognition), natural language processing (automatic translation, conversation analysis), speech recognition and handwriting recognition, information search and search engine (text mining, spam filter, extraction, summarizing, recommended system), bioinformatics (gene analysis, protein classification, disease diagnosis), computer, graphics and games (animation, virtual reality), robot (route search, autonomous car, object recognition and classification) and other fields.

It can be applied to our real life in a variety of ways, utilizing classification, which is one of the methods of machine learning. Monitoring the quality of products processed in a manufacturing plant in real time is an important issue from the viewpoint of improving product reliability. The problem of distinguishing the defective product from the produced product is time-consuming and expensive to solve by manpower inspection. Existing image based defect detection methods have a large amount of computation and are difficult and difficult to process in real time.

Iris data is a dataset introduced by biologist Ronald Fisher for linear discriminant analysis in his 1936 article “The use of multiple measurements in taxonomic problems” [5]. This data set is a public data set that has long been used in machine learning and statistics. There are three species of Iris data used here, Iris setosa, Iris versicolor, and Iris virginica. Since the Irises are classified by the ratio of the size of the petals to the sepals, the data are measured by measuring the length and width of each part of the petals and consisted of 150 records. It is a data set that records the length and width of the sepal and the length and width of the petals according to the type of Irises. Analyzing this data provides a useful knowledge of what differences are in terms of the sepal and petal lengths of Irises. This data set is suitable for classification, one of the popular areas of machine learning. It is used as a prediction problem to measure the species of Iris using Iris size.

Ⅱ. RESEARCH METHODOLOGY

K-nearest neighbor algorithm is used for classification learning and is a very simple and efficient non-parametric method proposed by Hart in 1968. It is a very intuitive method of finding the k-nearest individuals in the training dataset for a single entity based on the similarity between the samples and assigning the highest frequency group within the k-sets. There are many ways to measure similarity within k-nearest neighbors. Methods of measuring distance between entities include Euclidean distance, Manhattan distance, Minkowski distance, Mahalanobis distance, Chebyshev distance, and Hamming distance, etc.. The most commonly used and widely known of these is the Euclidean distance. Fig. 1 shows three distances to measure the distance between entities [6, 7, 8].

$$Euclidean\,Distance\ D(X, Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

$$Manhattan\,Distance\ D(X, Y) = \sum_{i=1}^{n}|x_i - y_i|$$

$$Minkowski\,Distance\ D(X, Y) = (\sum_{i=1}^{n}|x_i - y_i|^p)^{\frac{1}{p}}$$

Fig. 1 Distances measuring similarity of algorithm

As the data distribution changes, the results of the similarity measurement method appear differently. More important than similarity measurement in the k-nearest neighbors are determining k-values. When determining k-values, you should find the k-value with the best classification performance based on the data. If the k-value is too large, the probability of belonging to the highest frequency in the existing training data increases, and it cannot be classified in detail.

Conversely, when the k-value is too small and the k-value is 1, the outlier data are highly affected. Table 1 below shows a part of the data sets will be used in the research algorithm. The data set Iris to be used to verify the algorithms used in this model is a data set frequently used in libraries such as Python as well as data analysis and machine learning. In addition, Iris is a French chrysanthemum, which is introduced by the statistician Fisher and consists of 150 records as measurement data, including the width and length of each part of the petal. The data consist of six fields in total, the field consists of the order (No), sepal length (Sepal length), sepal width (Sepal width), petal length (Petal length), petal width (Petal width), and flower species (Species).

Table. 1 Raw data set for three species of Iris

No	Sepal length	Sepal width	Petal length	Petal width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.3	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3.0	1.4	0.1	setosa
14	4.3	3.0	1.1	0.1	setosa
15	5.8	4.0	1.2	0.2	setosa
~	~	~	~	~	~
~	~	~	~	~	~
150	5.9	3	5.1	1.8	virginica

Ⅲ. DATA ANALYSIS AND RESULT

Using the length and width of Sepal and Petal, the Iris data provided by Fisher, we classify the species and verify their accuracy. Based on the result, the new data is used to validate the algorithm in the same way. In this study, it is important to check through Euclidean distance, which has been verified in previous studies. First of all, the chart of three kinds Iris (setosa,

versicolor, and virginica) using the Iris raw data is shown in Fig. 2 below.

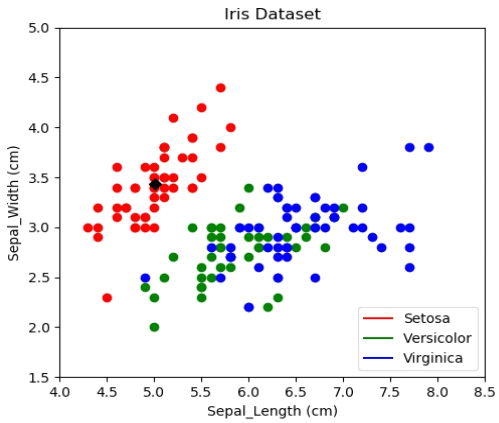

Fig. 2 The chart of Iris using raw data

In the chart for the basic data, setosa is located in the upper left. Next, versicolor is located diagonally from the upper right to the lower left, and virginica is divided diagonally from the upper right to the lower center. Fig. 2 demonstrates which the data are displayed without any processing based on the basic data.

Table. 2 New Iris data values

New item	
Sepal length	5.01
Sepal width	3.43
Pepal length	1.46
Pepal width	0.25
Species	setosa

Based on the basic data described earlier, the important issue is how to classify the new Iris sepal and petal length and width. Table 2 below shows the basic data values for the length, width of the petal and the sepal of the new Iris. Then, we went into the main body and finished preparation for analysis. In principle, we begin the classification process using the new Iris data.

Ⅳ. CONCLUSIONS

First, the Euclidean distance is calculated using the new iris data sepal. We check the Euclidean distance by using the length and width of the new Iris data sepal. Next, the Euclidean distance is calculated using the length and width of the new Iris data sepal and the Euclidean distance is calculated using the length and width of the petal. The Euclidean distance between the raw data and Sepal length (5.01), Sepal width (3.43) in Table 2 was measured. The analysis result is shown as Table 3 below. The closest Euclidean distance data for the new data were in the order of 8 (5, 3.4, setosa), 27 (5, 3.4, setosa), 41 (5, 3.5, setosa), 44 (5, 3.5, setosa) and 40 (5.1, 3.4, setosa) in Table 3.

Table. 3 The Euclidean distance between Iris new data and Sepal values

No	Distance	Sepal length	Sepal width	Species	Rank
1	0.1140	5.1	3.5	setosa	6
2	0.4438	4.9	3	setosa	31
3	0.3860	4.7	3.2	setosa	23
4	0.5263	4.6	3.1	setosa	38
5	0.1703	5	3.6	setosa	10
6	0.6107	5.4	3.9	setosa	40
7	0.4111	4.6	3.4	setosa	29
8	0.0316	5	3.4	setosa	1
9	0.8081	4.4	2.9	setosa	50
10	0.3479	4.9	3.1	setosa	18
11	0.4743	5.4	3.7	setosa	34
12	0.2121	4.8	3.4	setosa	14
13	0.4785	4.8	3	setosa	35
14	0.8301	4.3	3	setosa	52
15	0.9742	5.8	4	setosa	63
16	1.1904	5.7	4.4	setosa	85
17	0.6107	5.4	3.9	setosa	40
18	0.1140	5.1	3.5	setosa	6
19	0.7829	5.7	3.8	setosa	48
20	0.3808	5.1	3.8	setosa	20
21	0.3912	5.4	3.4	setosa	26
22	0.2846	5.1	3.7	setosa	17
23	0.4438	4.6	3.6	setosa	32
24	0.1581	5.1	3.3	setosa	9

No	Distance	Sepal length	Sepal width	Species	Rank
25	0.2121	4.8	3.4	setosa	14
26	0.4301	5	3	setosa	30
27	0.0316	5	3.4	setosa	1
28	0.2025	5.2	3.5	setosa	13
29	0.1924	5.2	3.4	setosa	11
30	0.3860	4.7	3.2	setosa	23
31	0.3912	4.8	3.1	setosa	25
32	0.3912	5.4	3.4	setosa	26
33	0.6964	5.2	4.1	setosa	43
34	0.9127	5.5	4.2	setosa	55
35	0.3479	4.9	3.1	setosa	18
36	0.2302	5	3.2	setosa	16
37	0.4950	5.5	3.5	setosa	37
38	0.2025	4.9	3.6	setosa	12
39	0.7463	4.4	3	setosa	46
40	0.0949	5.1	3.4	setosa	5
41	0.0707	5	3.5	setosa	3
42	1.2398	4.5	2.3	setosa	90
43	0.6519	4.4	3.2	setosa	42
44	0.0707	5	3.5	setosa	3
45	0.3808	5.1	3.8	setosa	20
46	0.4785	4.8	3	setosa	35
47	0.3808	5.1	3.8	setosa	20
48	0.4701	4.6	3.2	setosa	33
49	0.3962	5.3	3.7	setosa	28
50	0.1304	5	3.3	setosa	8

Second, let's calculate the Euclidean distance between the new iris data values (Sepal length, Sepal width, Petal length, Petal width) and the raw data (Table 1). In addition to the previous analysis, all five data with the smallest Euclidean distance were classified as Species (setosa). Thus, Euclidean distance is calculated using the 4 variables as Sepal length (5.01), Sepal width (3.43), Petal length (1.46), and Petal width (0.25) in the new data (Table 2). The Euclidean distance between Iris new data (Table 1) and Iris Sepal length, Sepal width, Petal length, and Petal width in the new data (Table 2) are shown below. The closest Euclidean distance data for the new data were in the order of 8 (5, 3.4, 1.5, 0.2, setosa), 40 (5.1, 3.4, 1.5, 0.2, setosa), 1 (5.1, 3.5, 1.4, 0.2, setosa), 18 (5.1, 3.5, 1.4, 0.3, setosa), and 50 (5, 3.3, 1.4, 0.2, setosa) in Table 4. Therefore, it was confirmed through empirical analysis which the results of this study were the same as the previous studies.

Table. 4 The Euclidean distance between Iris new data and Sepal, Petal values

No	Dist-ance	Sep-L.	Sep-W.	Pet-L.	Pet-W.	Spec.	Rank
1	0.1382	5.1	3.5	1.4	0.2	setosa	3
2	0.4507	4.9	3	1.4	0.2	setosa	27
3	0.4208	4.7	3.2	1.3	0.2	setosa	25
4	0.5302	4.6	3.1	1.5	0.2	setosa	36
5	0.1873	5	3.6	1.4	0.3	setosa	7
6	0.6731	5.4	3.9	1.7	0.4	setosa	41
7	0.4184	4.6	3.4	1.4	0.3	setosa	24
8	0.0714	5	3.4	1.5	0.2	setosa	1
9	0.8118	4.4	2.9	1.4	0.2	setosa	44
10	0.3809	4.9	3.1	1.5	0.1	setosa	16
11	0.4786	5.4	3.7	1.5	0.2	setosa	31
12	0.2590	4.8	3.4	1.6	0.2	setosa	12
13	0.5051	4.8	3	1.4	0.1	setosa	34
14	0.9171	4.3	3	1.1	0.1	setosa	47
15	1.0095	5.8	4	1.2	0.2	setosa	48
16	1.2005	5.7	4.4	1.5	0.4	setosa	49
17	0.6489	5.4	3.9	1.3	0.4	setosa	39
18	0.1382	5.1	3.5	1.4	0.3	setosa	3
19	0.8204	5.7	3.8	1.7	0.3	setosa	45
20	0.3861	5.1	3.8	1.5	0.3	setosa	19
21	0.4616	5.4	3.4	1.7	0.2	setosa	29
22	0.3242	5.1	3.7	1.5	0.4	setosa	13
23	0.6412	4.6	3.6	1	0.2	setosa	38
24	0.3809	5.1	3.3	1.7	0.5	setosa	17
25	0.4910	4.8	3.4	1.9	0.2	setosa	33
26	0.4551	5	3	1.6	0.2	setosa	28
27	0.2076	5	3.4	1.6	0.4	setosa	8
28	0.2124	5.2	3.5	1.5	0.2	setosa	10
29	0.2076	5.2	3.4	1.4	0.2	setosa	9
30	0.4136	4.7	3.2	1.6	0.2	setosa	22
31	0.4184	4.8	3.1	1.6	0.2	setosa	23
32	0.4208	5.4	3.4	1.5	0.4	setosa	26
33	0.7135	5.2	4.1	1.5	0.1	setosa	42
34	0.9160	5.5	4.2	1.4	0.2	setosa	46
35	0.3537	4.9	3.1	1.5	0.2	setosa	15
36	0.3509	5	3.2	1.2	0.2	setosa	14
37	0.5226	5.5	3.5	1.3	0.2	setosa	35

No	Dist-ance	Sep-L.	Sep-W.	Pet-L.	Pet-W.	Spec.	Rank
38	0.2590	4.9	3.6	1.4	0.1	setosa	11
39	0.7649	4.4	3	1.3	0.2	setosa	43
40	0.1145	5.1	3.4	1.5	0.2	setosa	2
41	0.1819	5	3.5	1.3	0.3	setosa	6
42	1.2510	4.5	2.3	1.3	0.3	setosa	50
43	0.6731	4.4	3.2	1.3	0.2	setosa	40
44	0.3835	5	3.5	1.6	0.6	setosa	18
45	0.6009	5.1	3.8	1.9	0.4	setosa	37
46	0.4849	4.8	3	1.4	0.3	setosa	32
47	0.4088	5.1	3.8	1.6	0.2	setosa	21
48	0.4766	4.6	3.2	1.4	0.2	setosa	30
49	0.4014	5.3	3.7	1.5	0.2	setosa	20
50	0.1520	5	3.3	1.4	0.2	setosa	5

The implications and limitations of this study are as follows. Implications: using the open dataset, the Euclidean distance using algorithms in machine learning was measured and reproduced to provide researchers with basic knowledge. This study verified the algorithm using public data and prepared a comparative study on the existing algorithm using new data. Unfortunately, there is a limitation which no new dataset were found and no comparative analysis was possible.

REFERENCES

[1] S. Cho, D, Jung, S, Lee, M, Shin, and H. Park "Survey on Machine Learning Algorithms for SDN/NFV Automation," *The Journal of Korean Institute of Communications and Information Sciences*, vol. 44, no. 1. Jan. 2019.

[2] J. R. Quinlan, "Induction of Decision Trees," *Machine Learning*, vol. 1, no, 1, pp. 81-106, Mar. 1986.

[3] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18-28, Jul. 1998.

[4] J. A. Hartigan, and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100-108, Jan. 1979.

[5] Wiley Online Library, The Use of Multiple Measurements in Taxonomic Problems [Internet]. Available: https://doi.org/10.1111/j.1469-18091936.t-b02137.x.

[6] S. Y. Shin, and H. C. Lee, "Realistic Enhancement of 3D Expressions for Building Expressions with Hologram," *Journal of the Korea Institute of Information & Communication Engineering*, vol. 23, no. 09, pp. 1104-1109, Sep. 2019.

[7] H. M. Lee, and S. Y. Shin, "Design of The Wearable Device considering ICT-based Silver-care," *Journal of the Korea Institute of Information & Communication Engineering*, vol. 22, no. 10, pp. 1347-1354, Oct. 2018.

[8] S. P. Kim, and J. M. Kim, "A Study on Open Source Software Business Model based on Value," *Asia-pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology*, vol. 7, no. 2, pp. 237-244, Feb. 2017.

남수태(Soo-Tai Nam)

2011년 부산대학교 경영정보시스템 박사수료
2014년 원광대학교 정보관리학 박사
2017년 ~ 2019년 전 군산대학교 컴퓨터통신공학 시간강사
2015년 ~ 현 원광대학교 정보전자상거래학부 초빙교수
2019년 ~ 현 부산대학교 교양교육원 강사
※관심분야 : MIS, E-Business, Technology Management, Big-Data, Internet of Things

신성윤(Seong-Yoon Shin)

1994년 군산대학교 컴퓨터과학 학사
1997년 군산대학교 컴퓨터통신공학 석사
2003년 군산대학교 컴퓨터과학과 박사
2013년 ~ 현 한국정보통신학회 상임이사, 편집위원장, 수석부회장 등
2006년 ~ 현 군산대학교 교수
※관심분야 : Computer Engineering, Multimedia System

진찬용(Chan-Yong Jin)

1984년 고려대학교 경영학 학사
1987년 한국과학기술원 경영과학 석사
1990년 한국과학기술원 경영과학 박사수료
2009년 서남대학교 경영정보학 박사
2014년 ~ 2015년 UCLA, visiting scholar
1990년 ~ 현 원광대학교 교수
※관심분야 : MIS, E-Business, Venture Start-Up, Big-Data