ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20211102113116
Original scientific paper

A Study on the Features Selection Algorithm Based on the Measurement Method of the Distance Between Normal Distributions for Classification in Machine Learning

Byungju SHIN, Minwoo KIM, Bohyun WANG, Joon S. LIM*

Abstract: Feature selection is an important technique that simplifies machine learning models to easily understand, shorten learning time, and reduce curve over-fitting or under-fitting. This paper presents a shape selection algorithm based on a method of investigating similarities between sampled shape values for classification variables (classes). This is based on the premise that the lower the similarity, the higher the usefulness of class classification. The confidence interval of a normal distribution is used to measure similarity. It is judged that the more overlapping the confidence intervals, the higher the similarity. The smaller the duplication of the confidence interval, the lower the similarity, and if the similarity is low, it can be used as a criterion for classification. Therefore, I propose an equation to apply this method. To confirm the usefulness of the equation, a colorectal cancer dataset with about 2000 genes was used and comparative experiments were performed with other feature selection algorithms. The comparison algorithms were Gini Index (10 features), mRMR (10 features), and relational matrix algorithms (7 features). Artificial neural networks were generally used as machine learning algorithms, and comparative verification was performed based on the rib one-out cross-validation method. As a result of the experiment, the results of the Gini index (85.487%), mRMR (87.09%), and relational matrix algorithms (87.09%) were better than those of 88.71% by selecting 10 features. In addition, experiments on iris, wine, glass, music emotions, seeds, and Japanese collection datasets were conducted on multiple classification problems. In the case of wine, the accuracy was 98.8% when all functions were used, but six functions were removed, resulting in 99.4% accuracy. In the case of music sensitivity, the accuracy was 51.7% when all 54 features were used, but when 20 features were removed, it improved to 61.3%. In the case of seeds, it was found that when the number of seeds decreased from 7 to 5, it slightly improved from 93.3% to 93.8%. In the case of iris, glass, and Japanese vowels, the accuracy did not increase even though the function was removed. Therefore, it can be concluded that features can be easily and effectively selected from the multi-class classification problem using the method proposed in this paper.

Keywords: classification; distance; feature selection; Gaussian distribution; similarity

1 INTRODUCTION

The human genes have about 30 000 exons that store information. Using a machine-learning algorithm on these genes to predict for a certain disease, such as colon cancer, can become quite chaotic. Some genes may be the cause of colon cancer, but others may have nothing to do with it. Furthermore, even if there is a correlation, it can sometimes cause a misjudgement when the relationship is shown insignificant.

Therefore, if the goal is to predict whether there is a particular disease by inputting features in the machine learning algorithm, features that are highly related with the disease should be selected appropriately. This is called feature selection. If features are selected appropriately, then the problem can be treated in the stochastic problem domain, rather than the chaos problem domain. Most learning models that we use deal with stochastic problems. Regardless of the learning model type, an operation is required to select appropriate features for accurate prediction [1].

In this paper, i propose a feature selection method and show examples of experiments for verifying its effectiveness. The experiments deal with various problems, from a classification problem of determining whether there is a specific disease using gene information to binomial classification and multinomial classification problems using multiple datasets. In the case of a disease determination problem, a correlation can be expected between the relevant genes and the disease, provided the difference in the gene values between the normal state and the diseased state can be distinguished significantly. These genes deserve to be feature candidates. If there is ambiguity in distinguishing between the values of a normal case and an abnormal case, then it is not used as a feature because it is considered irrelevant to the pertinent disease. This paper is about a method of finding genes that have significant difference in the values in terms of probability among many genes.

As the proportion of unrelated features (or noises) in the dataset increases, the machine learning converges less, ultimately leading to divergence. Appropriate feature selection eliminates noise features, preventing divergence of learning, and simplifies models to facilitate easy understanding for researchers and users, as well as reducing learning time [2, 3].

Existing feature selection methods include a global search method, which evaluates all subsets in the entire features space. However, it is not a realistic alternative because, as the size of the entire feature space increases (i.e., as the number of features to be considered increases), the number of subsets increases exponentially. Instead, a sequential forward (SFS) method and a sequential backward search (SBS) method can be used. Starting with an empty set, the SFS method adds a feature by finding the best feature in the feature space and then repeats the evaluation process of finding and adding the next best feature until no more good results can be obtained. Conversely, the SBS method is used to find and remove the worst features individually from the entire feature space. These methods facilitate appropriate features faster than the global search method. However, whenever a feature is added or removed, the learning has to be performed for its evaluation. If the size of the feature space is large, a significant amount of time and computing power are required to evaluate a good model based on machine learning.

The method proposed in this paper evaluates the interrelationships. The feature selection methods using the interrelationships include mutual information [4], *t*-test [5], Gini index [6], chi-square [7], maximum relevance-minimum redundancy (mRMR) [8], and Bhattacharyya distance [9, 10], and the relational matrix algorithm produces better results than these methods [11]. The relational matrix algorithm produces good results, but it relies on NEWFM with the limitation in scalability. NEWFM is an algorithm that classifies classes using the Fuzzy function. RM selects features by analyzing the relationship between the Fuzzy functions inside the NEWFM. Therefore, RM is bound to be very dependent on NEWFM. This study is the result of efforts made to improve this [11-15].

These experiments aimed to select effective features for classification among the features of various datasets and to reduce the uncertainty of predictions and bring them closer to stochastic problems. In other words, I aimed to show experimentally that the method of selecting features could be distinguished stochastically and training the machine learning algorithm would be effective in increasing the classification accuracy.

2 RELATED WORKS

To examine the performance of the method proposed in this study, i compared it with that of three algorithms: the relational matrix algorithm, Gini coefficient, and mRMR. Among them, the relational matrix algorithm starts on two premises.

1. "When event A occurs, if events B and C are frequently observed, A and B and A and C may have correlations, if not causality relationships".

2. In the first premise, "although B and C are usually observed simultaneously when event A occurs, they are not always observed simultaneously".

Based on the first premise, it can be thought that event A has occurred if event B is observed. Similarly, if event C is observed, it can be thought that event A has occurred. Based on the second premise, it can be thought that event A has occurred when event B is observed, even if event C is not observed; the reverse case is also the same. In this case, all you have to know is whether event B or C has occurred to determine whether event A has occurred. Therefore, event B or C is said to be complementary. The characteristic of this algorithm is that this phenomenon is actually observed by the learning algorithm. For this, the learning algorithm uses NEWFM because NEWFM has a structure that can evaluate individual features, and it is suitable to use in the relational matrix algorithm.

The second feature selection algorithm used in the comparison is a feature selection method using the Gini coefficient. Gini index or Gini coefficient [6, 16], which is used as a statistical index that shows the degree of inequality, can be obtained through the Lorenz curve. In economics, the Lorenz curve shows the probability distribution when *y*% income is distributed among the bottom *x*% households in a graph of the cumulative distribution function.

If the principle of the Gini coefficient is applied, the coefficient value can be used as an indicator to determine the purity, which indicates the degree of how well the feature classifies the classes. When the purity is high, the feature is determined to be suitable for use. When 20 genes were selected from the colon dataset based on this method, the accuracy obtained through the leave-one-out cross-validation (LOOCV) was 88.71% in the case of training with a Bayesian network.

The third algorithm is maximum relevance minimum redundancy (mRMR) [17]. This algorithm uses mutual information between variables to find a subset of features that have high maximum relevancy and low minimum redundancy. When 20 genes were selected from the colon dataset based on this method, the accuracy obtained through LOOCV was 83.87% in the case of training with the Bayesian network.

3 METHOD OF SELECTING FEATURES THROUGH MEASUREMENT OF DISTANCE BETWEEN NORMAL DISTRIBUTIONS

Numerous statistics used in our everyday lives and many statistical data follow normal distribution. Normal distributions are important distributions because they enable various statistical analyses by just assuming that certain data follow a normal distribution.

Assuming that the data follow a normal distribution, the confidence interval can be obtained by using the mean and the standard deviation. If the confidence interval is assumed to be 2 sigma (σ), it is, at least, equivalent to setting a valid range for the parameter of interest.

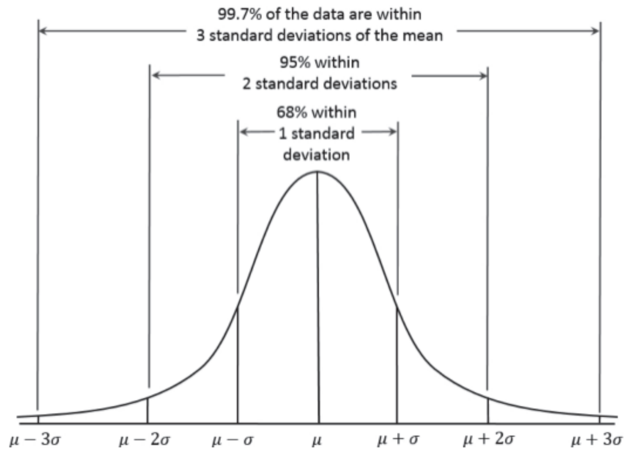


Figure 1 Confidence interval of the normal distribution

In statistics, the minimum level of significance (allowed limit) that becomes the baseline of a hypothesis test can be interpreted as the limit value of a reliable range. The confidence interval estimates the probability of the parameter included in the confidence interval, and a 95% confidence interval implies that the level of significance is 0.05. The idea of this study is to use the confidence interval of a group that is estimated to be a different group, instead of the *p*-value. To be able to say that groups A and B are different groups, the confidence interval of group A and that of group B must not overlap (Fig. 2a). If there is a section where the confidence intervals overlap, this section is ambiguous, which may be group A or group B (Fig. 2b).

Therefore, if the overlapping section is small, there is a high possibility that the two groups are different groups. A simple equation needs to be created to measure this. If it is required to have no overlap within the range of a 95% confidence interval of different distribution, then the confidence interval of the different distribution B must not overlap. This requires satisfying Eq. (1).

$$\left|\mu_i - \mu_j\right| > 2\sigma_i + 2\sigma_j \tag{1}$$

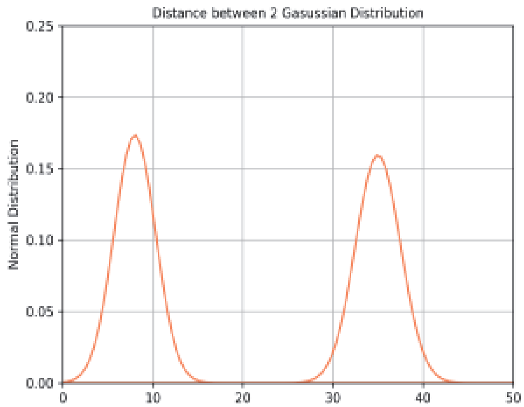
If both sides are divided by a positive number ($2\sigma_i + 2\sigma_j$), the inequality sign does not change. Thus, the equation can be converted into Eq. (2).

$$\frac{|\mu_i - \mu_j|}{2\sigma_i + 2\sigma_j} > 1 \tag{2}$$

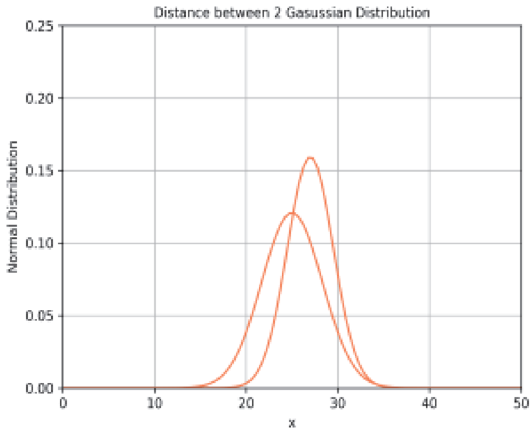
From Eq. (2), Eq. (3) for measuring the distance between normal distributions can be defined as follows:

$$GD(\mu_i, \sigma_i, \mu_j, \sigma_j) = \frac{|\mu_i - \mu_j|}{2\sigma_i + 2\sigma_j} \tag{3}$$

To satisfy a 99% confidence interval condition in Eq. (3), σ_i and σ_j have to be multiplied by 3, instead of 2.

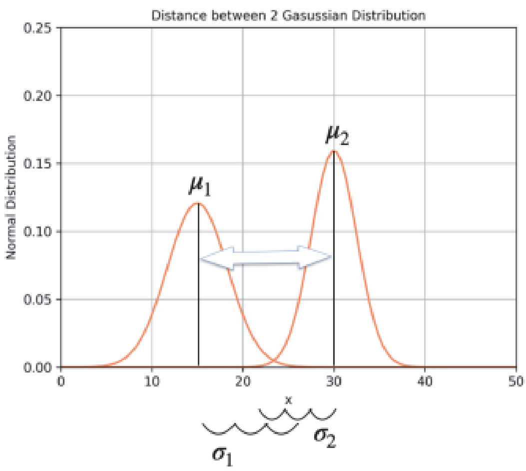


(a) μ_i= 8, σ_i= 2.3 and μ_j= 35, σ_j= 2.5

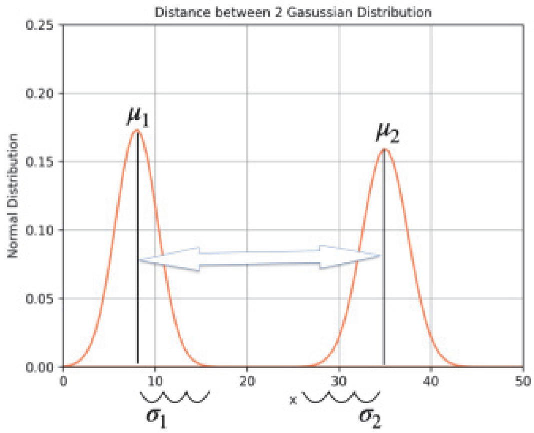


(b) μ_i= 25, σ_i= 3.3 and μ_j= 27, σ_j= 2.5

Figure 2 Evaluation criteria for the distance between normal distributions



(a) When $\frac{|\mu_1 - \mu_2|}{2\sigma_1 + 2\sigma_2} < 1$



(b) When $\frac{|\mu_1 - \mu_2|}{2\sigma_1 + 2\sigma_2} > 1$

Figure 3 Examples of distance measurement between normal distributions

3.1 Measurement Method for the Distance Between Normal Distribution in a Binomial Classification Problem

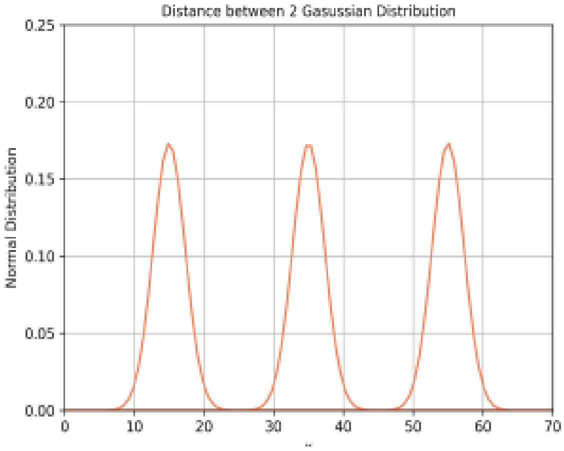
The mean and the standard deviation of each class are obtained for the features to be selected by applying the equation of distance measurement between normal distributions, and they are applied to (3) to obtain the result. This is the measurement method of the distance between normal distributions.

If the values obtained are sorted in descending order, a list sorted from the farthest distance between the distributions to the nearest distance can be obtained, and based on this list, a binary search method is proposed. This feature selection method applies the following algorithm to find the value from the sorted list. When there are features of n dimensions, the experimental results of using all features of n dimensions are compared to those of using $n/2$ features from the farthest feature in the distribution to the $n/2$th feature. If the result of training with $n/2$ features is better, then it is divided by $n/4$, and the result is compared to the result of using the $n/2$ features. If the result of training with n features is better than that of using $n/2$ feature, the experiment is performed with $3n/4$ features.

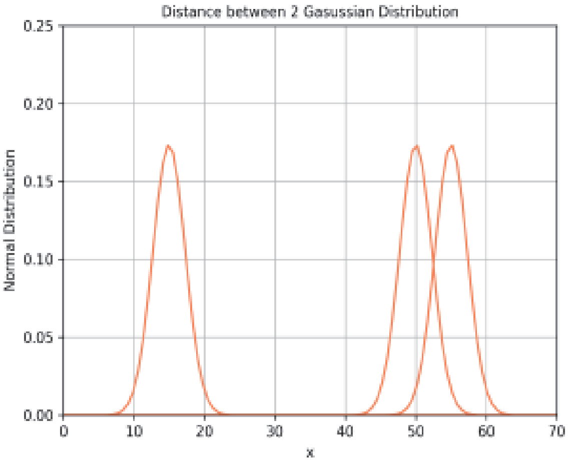
In this study, the features selection methods were comparatively experimented by selecting the number of features according to the results of some previously published papers to compare the effectiveness between the measurement method of the distance between normal distributions and existing feature selection algorithms.

3.1 Measurement Method for the Distance Between Normal Distributions in a Multinomial Classification Problem

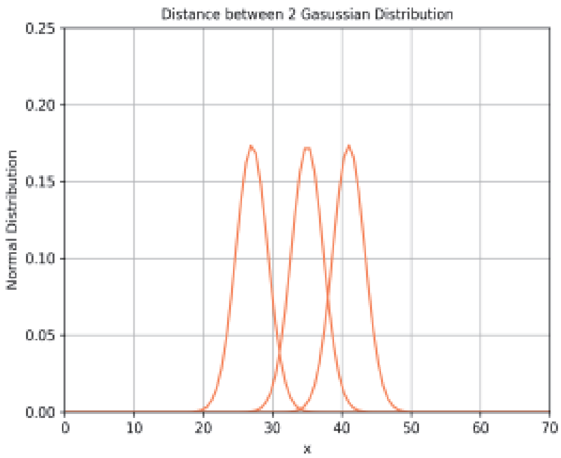
In the case of classifying three class types, if the distribution of the samples (feature values) extracted as the population for each class is the same as in Fig. 4, then Fig. 4a will be the best case for classifying the classes using these feature values because there is almost no overlapping section between the three distributions. In Fig. 4b, although the first distribution can be distinguished from the rest of the distributions, there is a large overlapping section between the two remaining distributions. Therefore, they cannot be classified, or the accuracy may decrease significantly due to the section that is ambiguous. Fig. 4c is the worst-case among the three. In this case, the feature is not suitable to use.



(a) Feature example



(b) Feature example



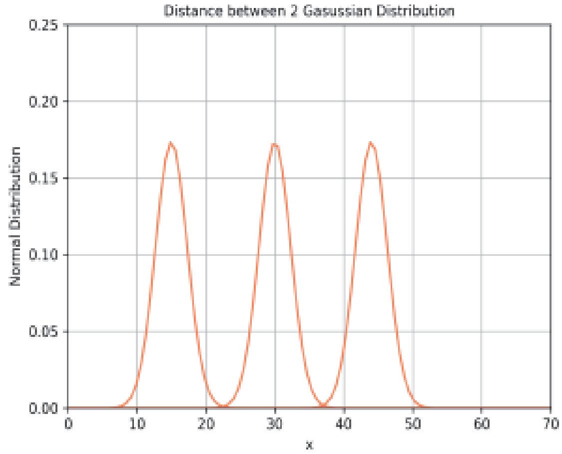
(c) Feature example

Figure 4 Examples of distance measurement between normal distributions

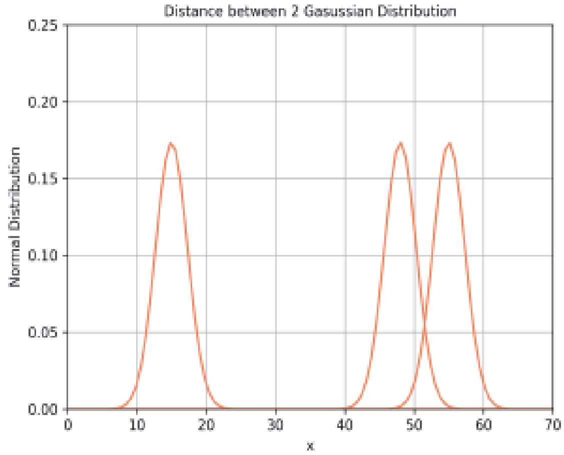
In the case of calculating using the equation of distance measurement between normal distributions using binary classification, we need as many calculations as the number of cases that can be made by selecting two elements from the set of n elements (nC$_2$) to measure the distance for n classes Eq. (4).

$$\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} GD\left(\mu_i,\sigma_i,\mu_j,\sigma_j\right) \tag{4}$$

However, a second equation may be needed to evaluate them because of the following reason.



(a) (μ_1 = 15, σ_1 = 2.3), (μ_2 = 30, σ_2 = 2.3), (μ_3 = 44, σ_3 = 2.3)



(b) (μ_1 = 15, σ_1 = 2.3), (μ_2 = 48, σ_2 = 2.3), (μ_3 = 55, σ_3 = 2.3)

Figure 5 Examples of multiple distributions (2)

The sum of *GD* in Fig. 5a is 6.30, and the sum of *GD* in Fig. 5b is 8.67. The calculation result seems to be good in Fig. 5b, but as it has an overlapping region, Fig. 5a will rather be a better choice. The feature selection for multiple classes cannot be performed simply by the distance measurement only. Therefore, an equation that calculates additional points (*AP*) is required. The *AP* can be obtained through Eq. (5).

$$\begin{cases} AP\left(\mu_i,\ \sigma_i,\ \mu_j,\ \sigma_j\right)=1, \text{ when } GD\left(\mu_i,\ \sigma_i,\ \mu_j,\ \sigma_j\right)\geq 1 \\ AP\left(\mu_i,\ \sigma_i,\ \mu_j,\ \sigma_j\right)=0, \text{ when } GD\left(\mu_i,\ \sigma_i,\ \mu_j,\ \sigma_j\right)<1 \end{cases} \tag{5}$$

If the result value of $GD\left(\mu_i,\ \sigma_i,\ \mu_j,\ \sigma_j\right)$ exceeds 1, the distance measurement equation is divided by 100, as shown in Eq. (6), to perform the scaling to distinguish it from the value obtained using $AP\left(\mu_i,\ \sigma_i,\ \mu_j,\ \sigma_j\right)$. Specifically, 100 is an arbitrarily set value, and it was selected as a safe value based on the experience because there were cases that the result value of $GD\left(\mu_i,\ \sigma_i,\ \mu_j,\ \sigma_j\right)$ was close to 10 in the experiments. It can be adjusted by the experimenter.

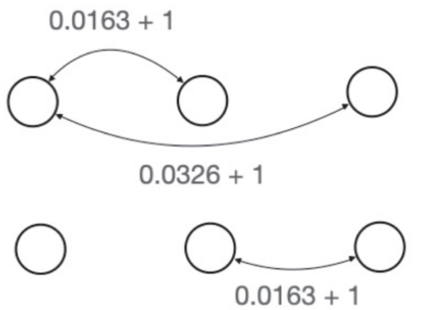
$$\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\frac{GD\left(\mu_i,\sigma_i,\mu_j,\sigma_j\right)}{100} \tag{6}$$

As the overlapped region decreases, a higher *AP* can be obtained. For example, if three classes have no overlap (i.e., if the result values of $GD\left(\mu_i,\sigma_i,\mu_j,\sigma_j\right)$ obtain 1 point each), 3 points are received as *AP*. If two distributions are overlapped, 2 points are received. If all three are overlapped, the *AP* is 0.

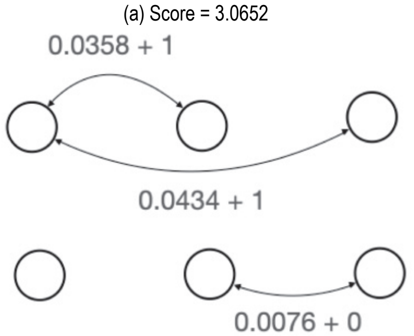
If Eq. (5) and Eq. (6) are combined, they can be expressed as the modified Eq. (7).

$$\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\frac{GD\left(\mu_i,\ \sigma_i,\ \mu_j,\ \sigma_j\right)}{100} + AP\left(\mu_i,\ \sigma_i,\ \mu_j,\ \sigma_j\right) \tag{7}$$

Calculate (7), the score of Fig. 6a is 3.07, and that of Fig. 6b is 2.09. Therefore, if (7) is used, it is determined that (a) is a better feature.



(a) Score = 3.0652



(b) Score = 2.0869

Figure 6 Examples of calculation for multinomial class distributions

Eq. (7) is still valid in the binary classification. Therefore, it can be used without distinguishing the binary and multinomial classification equations.

This method can be represented by Algorithm 1. Finally, the features are sorted based on the score to select and used as many features as desired. Thereafter, the classifier changes depending on the experiment, and in this study, an artificial neural network (ANN) was used.

Algorithm 1. Measurement method for the distance between normal distributions

for i = 0 to number of features do
for j = 0 to number of classes –1 **do**
calculate μ_{ij}, σ_{ij} of features for each class
End for
End for
for f = 0 to number of features **do**
totalscore[f] = 0
fori = 0 to number of classes – 1 **do**
for j = i + 1 to number of classes **do**
calculate GD
if GD ≥ 1 **do**
AP = 1
else if GD < 1 **do**
AP = 0
total score[f] + = ($\frac{GD}{100}$ + AP)
End for
End for
End for
Sort the features list by score
Select as many features as you like from the features list

4 EXPERIMENTS AND ANALYSIS

4.1 Experiments and Analysis for Binomial Classification

4.1.1 Feature Selection Experiments for Colon Cancer Dataset (Binomial Classification)

To check the validity of the proposed method, i selected features of the colon cancer dataset used by Jeyachidra [18]. This dataset is a dataset used in the Gene Expression Project at Princeton University. It has information on 2000 genes (features) selected through the first selection process and consists of data for 22 normal persons and 40 colon cancer patients. For this dataset, machine learning was performed by selecting features based on the measurement method of the distance between normal distributions.

The mean and the standard deviation of the normal class and the colon cancer class are calculated for 2000 features, respectively, to measure the distance between the normal distributions. The mean and standard deviation of each class calculated for each feature was used to calculate the score using the equation of distance measurement between normal distributions. The calculated scores were sorted and ranked. Fig. 7 shows the normal distributions of the normal tissue class and colon cancer tissue class for the top 10 features obtained through the calculations.

When the distribution of the top 10 features was examined, I found that there was no clear distinguishable value between the two distributions. However, the bottom 10 features showed almost completely overlapped shapes, as shown in Fig. 8. Tab. 1 summarizes the features selected based on different algorithms: the list of top 10 features obtained using the Gini index and mRMR, respectively, which were used by Jeyachidra [18], the list of 7 features selected using the relational matrix algorithm, and the list of ten features obtained using the Gaussian distribution distance.

Table 1 Top primary features according to different algorithms

Feature Selection Method	Index of Selected Features (Genes)
Gini Index	1671, **249**, **493**, **765**, **1423**, **513**, 1771, **245**, 267, **1772**
mRMR	1671, **249**, **493**, **765**, **1772**, 625, 1042, **1423**, **513**, 1771
Relational Matrix	72, **245**, **249**, 286, **493**, **765**, **1772**
Gaussian Distribution Distance	**249**, **765**, **1772**, **493**, **1423**, 245, 1582, 267, **513**, 780

I compared the accuracy among different selection methods by applying the same machine learning algorithm. The machine learning algorithm used in the experiments was an ANN, which has two hidden layers. Each hidden layer consists of nodes corresponding to 1.5 times the number of selected features. Sigmoid was used as the activation function, and the output layer was set up using Softmax. The loss function was configured to use the categorical cross-entropy. For the validation, the method used by Jeyachidra [16] was adopted.

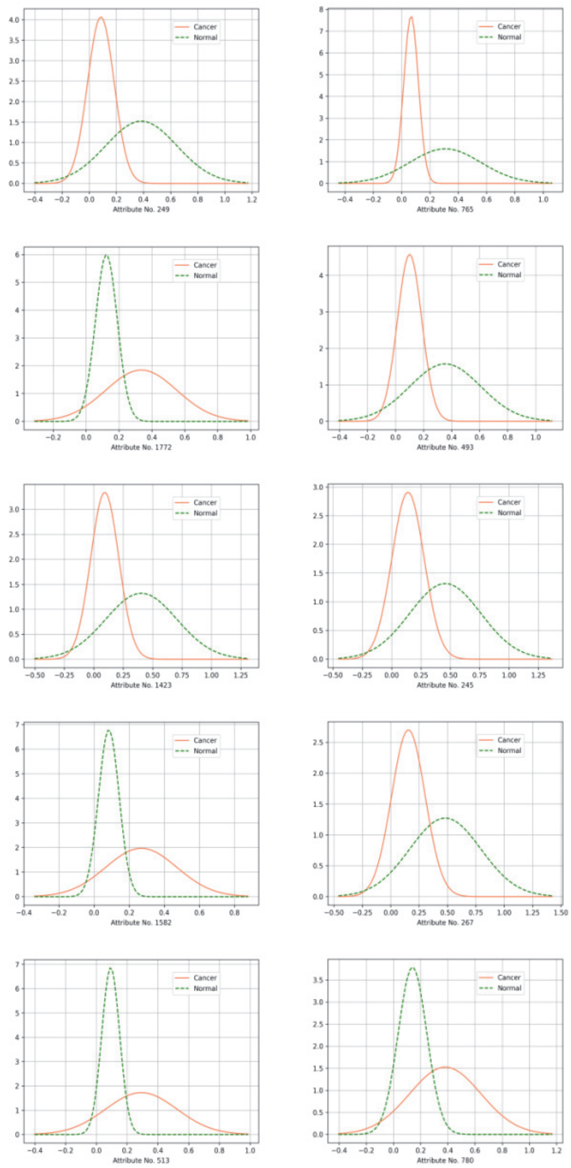


Figure 7 Normal distribution graphs for the top 10 features

Tab. 2 shows the result of LOOCV for the features selected based on the Gini index, mRMR, relational matrix, and the measurement method of the distance between normal distributions. LOOCV stands for Leave-One-Out Cross Validation. The Leave-One-Out CV (LOOCV) method is a method of making a total of N (number of samples) models, excluding only one sample when making each model, and calculating test set performance with other samples to average N performance.

Table 2 LOOCV results for the 10 features

Classifier	Feature Selection Method	Number of Selected Features	Accuracy / %	Time / sec
Artificial Neural Network	Gini Index	10	85.48	0.144
	mRMR	10	87.09	365.895
	Relational Matrix	7	87.09	-
	Gaussian Distribution Distance	10	**88.71**	**0.013**
		7	87.09	

In the experimental results, the accuracy showed slight differences, but the results of measuring the time consumed by each method of feature selection showed that the measurement method of the distance between normal distributions required the least amount of time. In this case, it seems that the experiment should be conducted by adding feature candidates of the next highest ranks individually, based on the SFS method, rather than trying to find the minimal features.

The features of lower ranks obtained in this experiment were not helpful in obtaining a higher classification accuracy. When they were used together in the training, they had to be removed because they hindered the training. However, when the SBS method was applied, it took a lot of time and cost to find optimal features. At this point, the results of the feature selection method proposed in this paper, were similar to those of the three methods compared.

In conclusion, the result was good, only when it had a statistically significant difference. In fact, the results of LOOCV for these were slightly different in each experiment. This is a basic characteristic of machine learning algorithms. Considering that the value to be observed in the experiment is the expected value and not the maximum value, it can be said that the four algorithms show similar results. However, the method proposed in this paper shows a better result in terms of efficiency because it requires less time compared to the other methods.

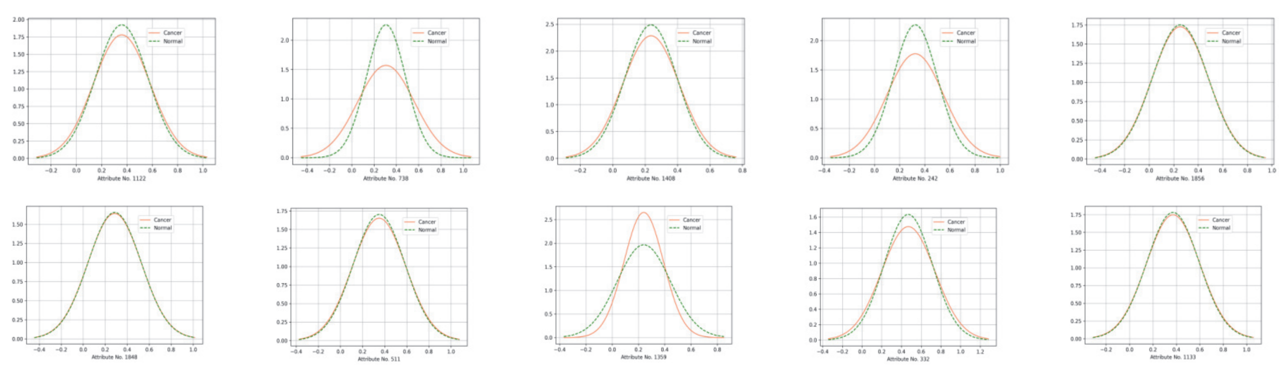


Figure 8 Normal distribution graphs for the bottom 10 features

4.1.2 Feature Selection Experiment for the Breast Cancer Dataset (Binomial Classification)

The second experiment for binomial classification used the breast cancer dataset (diagnostic) of the University of Wisconsin in the machine learning repository of UCI [19-23]. This dataset consists of 30 features and 212 malignant and 357 benign data. The features of the dataset consist of values such as radius, texture, circumference, and area obtained by actual measurements for breasts.

Table 3 10-fold CV result of the breast cancer dataset

Number of Features	Accuracy
27	**0.9841**
26	0.9824
24	0.9824
30	0.9807
25	0.9807
21	0.9807
20	0.9807
29	0.9806
22	0.9806
28	0.9788
19	0.9771

For this dataset, the distance between normal distributions was measured, and the features were sorted in descending order based on the scores. Then, the ANN was used to perform the training and validation by removing features one by one, starting from the worst until only one feature remained. K-fold cross-validation was used as the validation method, and the value of k was set to 10. Tab. 3 shows the number of selected features and the corresponding accuracy. The best result was 98.41%, which was obtained when three features were removed. It was better than 0.9807, when all 30 features were used.

4.2 Experiments and Analysis for Multinomial Classification

4.2.1 Feature Selection Experiment for Iris Dataset (Multinomial Classification)

Experiments were conducted using the iris dataset to examine the final equation of distance measurement between normal distributions considering the multi-class classification. The iris dataset [25] has three classes, namely, Setosa, Versicolour, and Virginica, and each has 50 data samples. This dataset has four features (i.e., Sepal Length, Sepal Width, Petal Length, and Petal Width). Tab. 4 shows the Gaussian distance score for each feature of the iris dataset.

Table 4 Gaussian Distance Scores of each feature of Iris dataset

Dataset	Features	Gaussian Distance Score	Order
Iris	Sepal Length	1.635	3
	Sepal Width	0.968	4
	Petal Length	7.682	1
	Petal Width	6.914	2

The scores recorded in Tab. 4 are the scores of the distance measurement between normal distributions shown in Fig. 9. Tab. 5 shows the accuracy when features were selected based on the scores shown in Tab. 5.

In brief, the best result was shown when all features were used. However, while the accuracy was 97% when the top two features were used, it was only 79% when the bottom two features were used, showing a difference of 0.18. This demonstrates the validity of the method of determining the ranking order.

Table 5 Experiment result of Iris dataset

Dataset	**Features**	**Accuracy**
Iris	Sepal Length, Sepal Width, Petal Length, Petal Width	0.98
	Sepal Length, Petal Length, Petal Width	0.97
	Petal Length, Petal Width	0.97
	Sepal Length, Sepal Width, Petal Width	0.96
	Sepal Length, Sepal Width	0.79

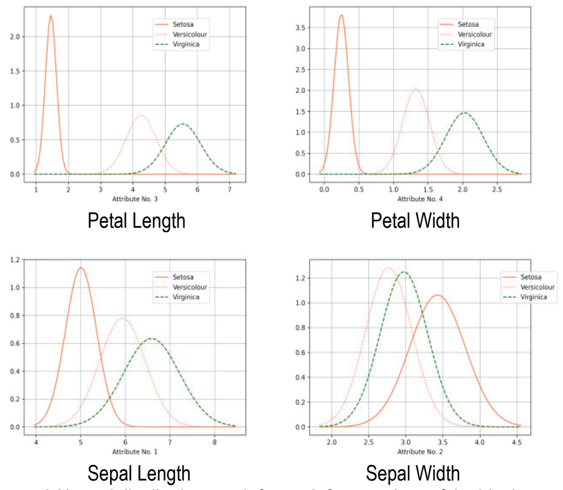


Figure 9 Normal distribution graph for each feature class of the iris dataset

4.2.2 Feature Selection Experiments for Various Datasets (Multinomial Classification)

These experiments were conducted using the datasets used by Ping Zhong [24], and 10-fold cross-validation was used to perform the feature selection, training, and validation. Tab. 6 shows the results.

Table 6 Experiment result of the wine, glass, music emotion, seeds, and Japanese vowels dataset

Dataset	Classes	Using All Features		Using Selected Features		Explanation
		Number of Features	Accuracy	Number of Selected	Accuracy	
Wine	3	13	0.988	7	0.994	Classification of Alcohol Types
Glass	6	9	0.659	9	0.659	Classification of Glass Types
Music Emotion	4	54	0.517	34	0.613	Classification of Emotions
Seeds	3	7	0.933	5	0.938	Classification of Wheat Varieties
Japanese Vowels	9	12	0.802	12	0.802	Classification of Speakers

The glass and Japanese vowels datasets showed the highest accuracy when all the features were used. The wine dataset showed the highest accuracy when six features were removed. In the case of the music emotion dataset, a 0.1 increase was shown in the accuracy when 20 features were removed. In the case of the seed's dataset, the number of features was reduced by two although the accuracy did not show a large difference. These results imply that, when the computational processing time was reduced in the training process of machine learning, there was a time and cost wise benefit in the end.

5 CONCLUSION

Machine learning performed on a well-organized and properly processed (clean) dataset often guarantees good results. The process of obtaining clean data involves finding potential candidates from large data. One such process for obtaining clean data is feature selection. The goal of feature selection is to increase the accuracy of prediction or classification by removing the uncertainty and securing the certainty for the prediction or classification to the maximum degree possible. Sometimes the number of features is a matter of efficiency, depending on the machine learning environment. Therefore, feature selection is similar to a see-saw game in the sense that there is a trade off between improvement and efficiency of accuracy. For example, the colon cancer dataset (one of the datasets used as input for the experiments in this study) is used to predict colon cancer and to check for issues related to human health. Moreover, it is regarding a life threatening condition. In this case, the efficiency can be traded to an agreeable degree to prioritize improvement in the prediction accuracy

The colon cancer dataset used in the experiments had small samples, and as I could not secure separate data for testing and validation, I used the LOOCV, but LOOCV is vulnerable to outliers. Suppose that, if any one of the 62 data points used is an outlier, then using them in the training data would distort the learning and using them in the validation data would compromise the accuracy of learning. Furthermore, in the case of life-related experiments, one cannot argue that a certain algorithm is superior based on a small number of data. Therefore, this paper is a report confirming that through LOOCV the method developed in the research process is as effective as other methods.

In recent years, many theories have come up in the fields of data science and artificial intelligence, including machine learning, and the advancement of various sensors has triggered an advancement of the Internet of Things. This in turn has led to the creation of a large volume of data called big data. This provides us with a wonderful opportunity to apply machine learning to an increasing number of worthy targets. In such a scenario, the methods of selecting features or the roles of feature selection that processes data becomes increasingly important.

The proposed method is expected to reduce the computational turn-around time, which is one of the key issues in deep learning. In the future, studies will be conducted on machine learning algorithms that can be applied in real life, along with studies on deep learning by improving it through combinations of the proposed method and machine learning algorithms.

Acknowledgements

This research was supported by the Ministry of Science and ICT (MSIT), Korea, under the Information Technology Research Center (ITRC) support program (IITP-2022-2017-0-01630), supervised by the Institute for Information & communications Technology Promotion (IITP). "This work was supported by the Gachon University research fund of 2019 (GCU-2019-2019-0766)".

6 REFERENCES

[1] Guyon, I. & A. Elisseeff (2003). An introduction to variable and feature selection. *JMLR*, *3*.
[2] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer. https://doi.org/10.1007/978-1-4614-7138-7
[3] Bermingham, M. L., Pong-Wong, R., Spiliopoulou, A., Hayward, C., Caroline, I., Rudan, H., Campbell, A. F., Wright, J. F., Wilson, F., Agakov, P., Navarro, P., & Haley, C. S. (2015). Application of high-dimensional feature selection: evaluation for genomic prediction in man. *Scientific Reports*, *5*. https://doi.org/10.1038/srep10312
[4] Liu, X., Krishnan, A., & Mondry, A. (2005). An entropy based gene selection method for cancer classification using microarray data. *BMC Bioinformatics*, *6*, 1-14. https://doi.org/10.1186/1471-2105-6-76
[5] Li, J., Su, H., Chen, H., & Futscher, B. W. (2007). Optimal search-based gene subset selection for gene array cancer classification. *IEEE Transactions on Information Technology in Biomedicine*, *11*, 398-405. https://doi.org/10.1109/TITB.2007.892693
[6] Gini, C. (1912). *Variabilità e Mutuabilità. Contributoallo Studio delle Distribuzioni e delle Relazioni Statistiche*. C. Cuppini, Bologna.
[7] Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, *3*, 1289-1305.
[8] Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*, 1226-1238. https://doi.org/10.1109/TPAMI.2005.159
[9] Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, *35*, 99-109.
[10] Reyes-Aldasoro, C. C. & Bhalerao, A. (2006). The Bhattacharyya space for feature selection and its application to texture segmentation. *Pattern Recognition*, 39, 812-826. https://doi.org/10.1016/j.patcog.2005.12.003
[11] Shin, B., Wang, B., & Lim, J. S. (2019). Relational matrix algorithm for feature selection in a fuzzy neural network. *Basic & Clinical Pharmacology & Toxicology*, 124, 114-115.
[12] Lim, J. S. (2005). Extracting minimized feature input and fuzzy rules using a fuzzy neural network and non-overlap area distribution measurement method. *Journal of the Korean Institute of Intelligent Systems*, *15*(5), 599-604. https://doi.org/10.5391/JKIIS.2005.15.5.599
[13] Lim, J. S. (2009). Finding features for real-time premature ventricular contraction detection using a fuzzy neural network system. *IEEE Transactions on Neural Networks and Learning Systems*, *20*, 522-527. https://doi.org/10.1109/TNN.2008.2012031
[14] Lim J. S. & Gupta, S. (2004). Feature selection using weighted neuro-fuzzy membership functions. *The 2004 International Conference on Artificial Intelligence (IC-AI'04)*, *1*, 1301-1315.
[15] Lim, J. W., Shin, B. J., & Lim, J. S. (2017). A match count method (mcm) for feature selection with cancer datasets in a neuro-fuzzy system. *Engineering and Bio Science, International Journal of Pharma and Bio Sciences*, 236-242.
[16] Zhu, W. & Lin, Y. (2013). Using gini-index for feature weighting in text categorization. *Journal of Computer and System Sciences*, *9*, 14. https://doi.org/10.2991/icibet-14.2014.22

[17] Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27, 1226-1238. https://doi.org/10.1109/TPAMI.2005.159

[18] Jeyachidra, J. & Punithavalli, M. (2013). A comparative analysis of feature selection algorithms on classification of gene microarray dataset. *Information Communication and Embedded Systems*, 1088-1093. https://doi.org/10.1109/ICICES.2013.6508165

[19] Shik, L. J. (2004). Extracting Wisconsin Breast Cancer Prediction Fuzzy Rules Using Neural Network with Weighted Fuzzy Membership Functions. *Korea Information Processing Society*, *11B*, 717-722. https://doi.org/10.3745/KIPSTB.2004.11B.6.717

[20] Wolberg, W. H., Street, W. N., Heisey, D. M., & Mangasarian, O. L. (1995). Computerized Breast Cancer Diagnosis and Prognosis from Fine Needle Aspirates. *Archives of Surgery*, *130*, 511-516. https://doi.org/10.1001/archsurg.1995.01430050061010

[21] Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (1994). Machine learning techniques to diagnose breast cancer from fine-needle aspirates. *Cancer Letters*, *77*, 163-171. https://doi.org/10.1016/0304-3835(94)90099-X

[22] Street, W. N., Wolberg, W. H., & Mangasarian, O. L. (1993). Nuclear Feature Extraction for Breast Tumor Diagnosis. *IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology*, *1905*, 861-870. https://doi.org/10.1117/12.148698

[23] Mangasarian, O. L., Street, W. N., & Wolberg, W. H. (1995). Breast cancer diagnosis and prognosis via linear programming. *Operations Research*, *43*(4), 570-577. https://doi.org/10.1287/opre.43.4.570

[24] Zhong, P. & Fukushima, M. (2005). A regularized nonsmooth Newton method for multi-class support vector machines.

[25] Lim, J. S. (2004). Finding fuzzy rules for iris by neural network with weighted fuzzy membership function. *The International Journal of Fuzzy Logic and Intelligent Systems*, *4*(2), 211-216. https://doi.org/10.5391/IJFIS.2004.4.2.211

[26] Shin, B., Kim, M., Wang, B., & Lim, J. S. (2021). Features Selection Method Based on the Measurement of the Distance Between Normal Distributions for Classification. *International Journal of Advanced Science and Technology*, *149*.

Contact information:

Byungju SHIN
k2soft
(06140) 4th Floor, 39, Teheran-ro, 33-gil, Gangnam-gu, Seoul, Republic of Korea
E-mail: bjshin@k2soft.kr

Minwoo KIM
k2soft
(06140) 4th Floor, 39, Teheran-ro, 33-gil, Gangnam-gu, Seoul, Republic of Korea
E-mail: minwoo2305@gmail.com

Bohyun WANG
Gachon University,
(13120) 531 AI Building, 1342,
Seongnam-daero, Sujeong-gu, Seongnam-si, Gyeonggi-do, Republic of Korea
E-mail: bhwang99@daum.net

Joon S. LIM
(Corresponding author)
Gachon University,
(13120) 310 AI Building, 1342,
Seongnam-daero, Sujeong-gu, Seongnam-si, Gyeonggi-do, Republic of Korea
E-mail: jslim@gachon.ac.kr