Finding the Optimal Data Classification Method Using LDA and QDA Discriminant Analysis

SeungJae Kim\textsuperscript{1} and SungHwan Kim\textsuperscript{2}\textsuperscript{*}

Abstract

With the recent introduction of artificial intelligence (AI) technology, the use of data is rapidly increasing, and newly generated data is also rapidly increasing. In order to obtain the results to be analyzed based on these data, the first thing to do is to classify the data well. However, when classifying data, if only one classification technique belonging to the machine learning technique is applied to classify and analyze it, an error of overfitting can be accompanied. In order to reduce or minimize the problems caused by misclassification of the classification system such as overfitting, it is necessary to derive an optimal classification by comparing the results of each classification by applying several classification techniques. If you try to interpret the data with only one classification technique, you will have poor reasoning and poor predictions of results. This study seeks to find a method for optimally classifying data by looking at data from various perspectives and applying various classification techniques such as LDA and QDA, such as linear or nonlinear classification, as a process before data analysis in data analysis. In order to obtain the reliability and sophistication of statistics as a result of big data analysis, it is necessary to analyze the meaning of each variable and the correlation between the variables. If the data is classified differently from the hypothesis test from the beginning, even if the analysis is performed well, unreliable results will be obtained. In other words, prior to big data analysis, it is necessary to ensure that data is well classified to suit the purpose of analysis. This is a process that must be performed before reaching the result by analyzing the data, and it may be a method of optimal data classification.

Keywords: Machine Learning, LDA, QDA, SVM, Classification Analysis

1. Introduction

With the emergence of the 4th industrial revolution, AI (artificial intelligence)\textsuperscript{[1-3]}, big-data\textsuperscript{[4-6]}, Internet of Things (IOT)\textsuperscript{[7-9]}, etc. 4 Core technologies that will lead the tea industrial revolution are also at the center of the general public's topic. This meaning is not only studied and used by large corporations, government offices, universities and research institutes, but also by the businessmen and those who are preparing for employment. This means that they try to maximize income, profit and self-competence by using core technologies of the fourth industry. In other words, individuals, companies, and the state use industrial robots based on the technology of artificial intelligence (AI) to reduce employment costs and increase production to increase profits. In addition, in order to understand the characteristics of consumers, big data analysis (BDA) techniques are introduced in places where goods are sold to classify personal information of consumers and analyze each consumption type, thereby expanding profits through accurate sales for consumers. These AI technologies and BDA technologies must show sophisticated results so that errors do not occur. The importance of data cannot be overlooked in order to show sophisticated results. With the recent introduction of AI technology, the use of data is rapidly increasing, and newly generated data is also rapidly increasing. In order to obtain the results to be analyzed based on these data, the first thing to do is to classify the data well. However, when identifying and classifying data, if the data is classified and analyzed by applying only one discrimination or classification technique belonging to the machine learning (ML) technique\textsuperscript{[10-12]}, an error of overfitting can be
accompanied. Because in order to reduce or minimize the problems caused by misclassification of the classification system such as overfitting[13-15], the optimum classification must be derived by comparing each classification result by applying several classification techniques. If you try to interpret the data with only one classification technique, you will have poor reasoning and poor predictions of results.

Among BDA techniques, the field of classifying data is the field of ML and is called data mining. Classification analysis (CA), which belongs to the new learning technique, includes Decision Tree (DT)[16-18], Random Forest (RF)[19,20] and Support Vector Machine (SVM)[21]. There are[22], Logistic Regression (LR) Analysis[23], Linear Discriminant Analysis (LDA)[24,25] and Quadratic Discriminant Analysis (QDA)[26,27]. In BDA, CA uses this analysis method to train and test data to classify data. The numerical values obtained through CA are analyzed statistically to determine the degree of classification of the data.

In this study, we will look at data from various perspectives before analyzing data in data analysis, and seek ways to classify data optimally by applying LDA and QDA of various classification techniques such as linear classification or nonlinear classification. In order to obtain the reliability and sophistication of statistics as a result of BDA, it is necessary to analyze the meaning of each variable and the correlation between the variables. If the data is classified differently from the hypothesis test from the beginning, even if the analysis is performed well, unreliable results will be obtained. In other words, prior to BDA, it is necessary to ensure that data is well classified to suit the purpose of analysis. This is a process that must be performed before reaching the result by analyzing the data, and it may be a method of optimal data classification.

2. Machine Learning Definition

ML is a process in which computers can predict the future by collecting and analyzing data on their own so that machines can make decisions and infer like humans. This is a technology that gives a computer the ability to infer and predict by learning data having a specific meaning (ML) based on an algorithm[28]. His ML is largely divided into three categories: the first is supervised learning, the second is unsupervised learning, and the third is reinforcement learning.

First, supervised learning finds a prediction function from specific data and predicts the result of a new input value. Second, unsupervised learning learns random data by itself. Third, reinforcement learning reinforces learning through errors. In addition, ML has sophisticated classification capabilities through deep learning (DL)[29]. Whether classifying data using such ML or classifying data using DL, handling and processing the data itself means data mining.

The ML techniques to be applied in this study are LDA and QDA of DA methods that discriminate and classify data as a field of data mining.

3. Data Mining Definition

Data mining is a method that is mainly used when analyzing multivariate data. Multivariate refers to all the situations that are necessary for the result to come out when there is one result. In other words, it refers to all necessary variables (multivariate). This data mining is to convert the data stored for simple business processing into a data format suitable for the purpose of analysis and analyze it. In addition, it makes it easy to refine and analyze data so that data built for general business purposes can be analyzed for research or knowledge management purposes. It also belongs to the money learning technique as an analysis method that allows interpretation from various aspects.

3.1. Features of Data Mining

Data mining uses large amounts of data and can process vast amounts of data in statistical inference based on existing samples by using computer learning algorithms. It can be used in all fields such as statistics, database, pattern recognition (PR), ML, and AI. Because
it is based on an empirical method, it is based on an empirical method. Model) can be created and generalized.

Data mining techniques are largely divided into two categories: goal-oriented (defined) and undefined goal (undefined). Each has three techniques internally. First, there are classification, estimation, and prediction methods for goal orientation. Second, there are grouping or association rules and clustering methods for unknown targets. And the data structure of data mining has a causal relationship.

3.2. Discriminant Analysis

The DA is called ‘Fisher's Discriminant Analysis’, and it is a process of measuring variables when each group of individuals is already known and making discriminant formulas that can distinguish each group using these variables. In other words, DA refers to an analysis method that finds a criterion that can determine from which population these samples were extracted by using information possessed by samples extracted from two or more parent groups.

Types of DA include LDA and QDA. DA using LDA is used as a method of reducing the dimension in terms of maximizing the separation between classes. DA using QDA is a quadratic discrimination function that maximizes group K.

3.3. Analysis Method of Discriminant Analysis

When analyzing data using DA, the reliability of the analysis result may vary depending on what data type the data used is. The independent variable(IV) must be composed of an isometric scale or a ratio scale, and the dependent variable(DV) must be composed of a nominal or sequence scale to analyze the relationship between the independent variable and the dependent variable. That is, the dependent variable should be described as an independent variable.

DA aims to measure the discriminant ability of how well each individual is divided into a group by using a discriminant function, and to predict which group a new object will be classified into. The value obtained by substituting the values of the discriminant variables of a new individual into the discriminant function is called the discriminant score. In addition, in order to obtain the reliability of the result of DA, there must be a certain number of samples. In the whole sample, the number of IV should be 3 directors, and the number of samples in the DV group should be less than the number of IV.

3.4. Analysis stage of Discriminant Analysis

The analysis stage of DA consists of four stages. First, the IV is selected. The second is to derive the plate change function equation. Third, verify the accuracy. Fourth, predict. Step-by-step explanations are given in Table 3.

<table>
<thead>
<tr>
<th>Table 1. Models and Applications of CA and RA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Result</td>
</tr>
<tr>
<td>Apply</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2. Types of Discriminant Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Division</td>
</tr>
<tr>
<td>Covariance matrix identity</td>
</tr>
<tr>
<td>Number of samples</td>
</tr>
<tr>
<td>Classification flexibility</td>
</tr>
</tbody>
</table>
CA classifies objects with multiple features into groups or categories, and generates efficient classification rules from a training sample. If classification is made according to these classification rules, the verification process by classification model must be performed. In the verification process, the degree of classification can be verified by calculating and comparing the misclassified rates among the total classification rates. The misclassification rate can be obtained by dividing the number of misclassified objects by the total number of objects. Table 4 is a table of classification rates by DA. Figure 3 shows the step of verifying the classification model by DA.

### Classification Model Verification of DA

CA classifies objects with multiple features into groups or categories, and generates efficient classification rules from a training sample. If classification is made according to these classification rules, the verification process by classification model must be performed. In the verification process, the degree of classification can be verified by calculating and comparing the misclassified rates among the total classification rates. The misclassification rate can be obtained by dividing the number of misclassified objects by the total number of objects. Table 4 is a table of classification rates by DA. Figure 3 shows the step of verifying the classification model by DA.

### Discriminant Analysis Experiment

This study seeks to find a way to optimally classify data using the DA method of data mining techniques. Types of DA include Linear DA and QDA. DA using LDA is used as a method of reducing the dimension in terms of maximizing the separation between classes. DA using QDA is a quadratic discrimination function that maximizes group K. In this experiment, both the LDA method and the QDA method will be applied. Classify the total number of samples into train data and test data, calculate the classification rate through DA, and calculate the data classification rate according to DA through the verification process of the classification model, and find out the reliability according to the classification rate.

#### 4.1. Subject and Method of Experiment

This study attempts to find out the degree of smartphone addiction among adolescents using the LDA method and QDA method of DA among data mining techniques. Data was collected by conducting a questionnaire survey, and four meaningful independent variables that could express adolescents’ smartphone addiction were selected from the survey data and used for analysis. The total number of data used in the analysis is 232.

As an experimental method, the tool used for analysis was coded with the version of R program (3.6.3) in Windows 10 environment. As DA techniques for data classification, the LDA method, which is a linear DA, and the QDA method, which is a quadratic discrimination function, are compared.
and the QDA method, which is a second-order linear DA, were used. Each method divides into a train (training) set and a test (test) set through an internal operation among all data, and calculates the classification rate after learning each. The smaller the difference, the better it can be said. Figure 4 is a part of 232 data used for DA, and the file format is (*.csv) file.

4.2. Discriminant Analysis Using LDA

LDA is used as a method of reducing the dimension in terms of maximizing the separation between classes. In other words, the dimension is reduced by classifying objects into several categories.

In this case, LDA is an analysis method used when variance and covariance matrices are the same regardless of the category of data when classifying data.

a. Data Segmentation

The data division divides a total of 232 data into train data and test data. At this time, train data is set to 70% of the total data, and test data is set to 30% of the total data. The code below is a code that divides train data and test data by a ratio of 7:3. Each of the two divided data sets was subjected to plate analysis, and the number of each data was divided into 169 train data and 63 test data.

```r
set.seed(123)
ind <- sample(2, nrow(lda1), replace = TRUE, prob = c(0.7, 0.3))
train <- lda1[ind==1, ]
test <- lda1[ind==2, ]
```

b. Test for Identity of Covariance Matrix

LDA is an analysis method used when the variance and covariance matrices are the same regardless of the category. As a result of testing the identity part of the covariance matrix, the P value was lower than 0.05, so a nonlinear analysis method rather than a linear analysis method should be used.

```
Chi-Sq (approx.) = 82.317, df = 12, p-value = 1.489e-12
```

However, since it is a study to find out the degree of data classification of LDA and QDA, I will use the linear analysis method.

c. Train

As a result of DAs using 169 data classified as trains in LDA analysis, it was found that the addiction of adolescents' cell phone use is due to “impulsiveness” by the coefficients of linear discriminants. Looking at the linear discrimination coefficient, it can be interpreted that the impulse of LD1 is the highest value as 1.333, and the discriminant power is the best, and the estimated classification rate is 0.9998, which is 99.98%.

<table>
<thead>
<tr>
<th>Coefficients of linear discriminants:</th>
</tr>
</thead>
<tbody>
<tr>
<td>LD1</td>
</tr>
<tr>
<td>Phone_time</td>
</tr>
<tr>
<td>Addictive</td>
</tr>
<tr>
<td>SNS_time</td>
</tr>
</tbody>
</table>

Proportion of trace:

<table>
<thead>
<tr>
<th></th>
<th>LD1</th>
<th>LD2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9998</td>
<td>0.0002</td>
<td></td>
</tr>
</tbody>
</table>

In the LDA analysis, the correct classification rate by cross-analysis was 55%. The calculation of the correct classification rate can be obtained by summing all the values of the diagonal components that intersect and dividing by 169, the total number of train data. Figure 5 shows the crosstabulation table of train data.

d. Test

As a result of discriminant analysis using 63 data classified as tests in LDA analysis, it was found that the addiction of adolescents’ cell phone use is due to...
‘impulsiveness’ by the coefficients of linear discriminants. Looking at the linear discrimination coefficient, it can be interpreted that the impulsiveness in LD1 is the highest as 2.73, and the discriminant power is the best, and the estimated classification rate is 0.8161, which is 81.61%.

In the LDA analysis, the correct classification rate by cross analysis was 58.7%. The calculation of the correct classification rate can be obtained by summing all the values of the diagonal components that intersect and dividing by 63, the total number of test data. Figure 6 shows the cross-analysis table of the test data.

e. LDA Analysis Result
As a result of analysis of all data, the estimated classification rate was 98.77%, which was lower than the estimated classification rate of 0.9998 of the train model and higher than 0.8161 of the test model, and the correct classification rate was 53.01%. Therefore, the prediction equation based on the linear discrimination coefficient of the train model of the DA model is as follows.

4.3. Discriminant Analysis Using QDA
DA using QDA, a nonlinear DA method, is a quadratic discrimination function that maximizes group K. QDA is an analysis method used when the variance and covariance matrices are different for each category of data when classifying data.

a. Train
As a result of the nonlinear DA using 169 data classified as trains in the QDA analysis, the correct classification rate by the cross-analysis table was 59.17%. The calculation of the correct classification rate can be obtained by summing all the values of the diagonal components that intersect and dividing by 169, the total number of train data. Figure 8 shows the crosstabulation of train data.

When comparing the correct classification rate of the QDA analysis method and the correct classification rate of the LDA analysis method, the train model correct classification rate of LDA is 55%, and the train model correct classification rate of QDA is 59.17%.

<table>
<thead>
<tr>
<th>Coefficients of linear discriminants:</th>
</tr>
</thead>
<tbody>
<tr>
<td>LD1</td>
</tr>
<tr>
<td>Phone_time</td>
</tr>
<tr>
<td>Addictive</td>
</tr>
<tr>
<td>SNS_time</td>
</tr>
</tbody>
</table>

Proportion of trace:
LD1 0.8161 0.1839
correct classification rate of QDA is 59.17%. QDA showed a correct classification rate of 4.1% higher than that of LDA. Therefore, in the train model, it can be said that the QDA analysis method has a 4.1% higher correct classification rate that can optimally classify the data than the LDA analysis method.

b. Test
As a result of nonlinear DA using 63 data classified as test in QDA analysis, the correct classification rate by cross-analysis table was 52.38%. The calculation of the correct classification rate can be obtained by summing all the values of the diagonal components that intersect and dividing by 63, the total number of test data. Figure 9 shows the cross-analysis table of the test data.

When comparing the correct classification rate of the QDA analysis method and the correct classification rate of the LDA analysis method, the correct classification rate of the LDA test model is 58.7%, and the correct classification rate of the QDA test model is 52.38%. QDA showed a 6.32% lower accuracy rate than LDA. Therefore, in the test model, it can be said that the QDA analysis method has a 6.32% lower correct classification rate that can optimally classify data than the LDA analysis method.

c. QDA Analysis Result
As a result of the analysis of the Quadratic DA model, the correct classification rate by the cross-analysis table was 55.6%, which was 3% higher than the train model 53.01% of the LDA analysis.

5. Conclusion

With the recent emergence of the 4th Industrial Revolution, core technologies that will lead the 4th Industrial Revolution such as AI, Big-data, and IOT are also at the center of the general public's topic. Individuals, companies, and the state use industrial robots based on the technology of AI to reduce employment costs and increase production to increase profits, and to introduce BDA techniques to large or small stores to determine consumer consumption patterns. By analyzing, you are expanding your profits. These AI technologies and BDA technologies must show sophisticated results so
that errors do not occur. The importance of data cannot be overlooked in order to show sophisticated results.

With the recent introduction of AI technology, the use of data is rapidly increasing, and newly generated data is also rapidly increasing. In order to obtain the results to be analyzed based on these data, the first thing to do is to classify the data well. However, when classifying data, if only one classification technique belonging to the ML technique is applied to classify and analyze it, an error of overfitting can be accompanied. In order to reduce or minimize the problems caused by misclassification of the classification system such as overfitting, it is necessary to derive an optimal classification by comparing each classification result by applying several classification techniques. If you try to interpret the data with only one classification technique, you will have poor reasoning and poor predictions of results.

This study seeks to find a method for optimally classifying data by looking at data from various perspectives and applying various classification techniques such as LDA and QDA, as a process before data analysis. As a result of the analysis of LDA, the estimated classification rate of the entire data was 98.77%, which was lower than the estimated classification rate of 0.9998 of the train model and higher than 0.8161 of the test model, and the correct classification rate was 53.01%. Therefore, the prediction equation based on the linear discrimination coefficient of the train model of the DA model is as follows. Looking at the analysis results of the QDA model, the correct classification rate by the cross-analysis table was 55.6%, which was 3% higher than the train model 53.01% of the LDA analysis. In order to improve the reliability of BDA results, it is necessary to ensure that data is classified well for the purpose of analysis. This study will be a methodology for optimal data classification as a process that must be carried out before reaching the results by analyzing data.

In the future, using methods such as DA or CA of several branches belonging to data mining, we will find a plan for optimal data classification and research on improvements for performance improvement. As a result, this is to improve and improve recognition ability through discrimination and classification of objects that are important in the realization of technology using DA and CA.

References


