ISSN 2616-5775 Vol. 6, Issue 9: 17-24, DOI: 10.25236/AJCIS.2023.060903

Research on Distinguishing Biological Species by Data Model and Linear Discriminant Analysis

Yanming Zhang*, Qihong Wu, Sijie Yang

Beijing National Day School, Beijing, China
zymzakryzzz@163.com
**Corresponding author*

Abstract: *This article explores the classification of lizards based on their distinct pholidosis and morphological characteristics using various data attributes. The authors aim to construct a classification model that takes advantage of data attributes for both simplicity and accuracy. Additionally, the article aims to propose an adaptive model that provides recommendations according to the precision requirements of biologists and the computational environment, enhancing the model's applicability. The authors employ Fisher's and Bayesian methods from linear discriminant analysis for classification, leveraging the linear structure to ensure the model's simplicity. A novel aspect of this work is the development of a discriminative power index for variables. This index prioritizes variables with strong discriminative abilities, thus simplifying computations and improving efficiency. The results align with those obtained through exhaustive searches for optimal solutions. Furthermore, the constructed model offers classification criteria and prediction accuracy under different variable combinations, enabling biologists to adjust variables based on accuracy needs and computational constraints. This functionality enhances the model's suitability for various real-world research scenarios.*

Keywords: *Classification problem, Linear discriminant analysis, Fisher's method, Bayesian method*

1. Introduction

There exists a certain correlation between lizard species and gender with their pholidosis and morphological characteristics. The objective is to establish criteria for classifying gender and species of lizards based on their various pholidosis and morphological characteristics. These criteria should be simple, effective, and practically meaningful. Upon an initial observation of a data set of 564 lizards of 8 species belonging to genus *Darevskia*, the authors noted significant differences in the average values of various variables within most categories. However, due to substantial within-category differences and the interference of multiple variables, data overlap prevents straightforward classification. Therefore, the authors focus on constructing standards that comprehensively evaluate the within-category data dispersion and between-category distances. Such standards evaluate data separability, guiding data transformation. Additionally, they render the classification criteria practically meaningful, allowing for parameter interpretation. The second focus is ensuring high classification accuracy while simplifying the classification method. It is crucial to identify methods for constructing straightforward classification criteria, a pivotal prerequisite for model applicability.

Combining previous research, this article employs linear discriminant analysis for modeling, particularly Fisher's method for binary classification. Fisher's method, initially proposed by Fisher, was used to address a two-flower classification problem [1]. The core idea is to find a line within the multidimensional space composed of variables, projecting all data points onto it to achieve dimension reduction. The method involves optimizing minimal within-class dispersion and maximal between-class distance to find the line that maximally separates the data. The resulting discriminant function is expressed as a linear combination of variables. Evidently, this method suits the dataset's features and the requirement for simplified classification criteria. Given the abundance of variables in the problem, even a linear discriminant function is not easily computed directly, necessitating variable optimization. For multi-classification problems, the authors incorporate Bayesian methods' ideas and procedures. This method calculates a discriminant function for each category using posterior probabilities. Predicting classification entails determining which discriminant function yields the highest value for a data point. Due to the need for simplification, variable selection remains essential.

2. Model 1: Submodel for Binary Classification Problems

2.1. Overall Description of Submodel 1

Model 1 combines the principles of Fisher's method in linear discriminant analysis to model binary classification problems involving only two categories. The discriminant function computed by this method involves a linear summation of all variables. Building upon this function, this article assesses the discriminative abilities of individual variables and simplifies the linear discriminant function for increased ease of use while maintaining predictive accuracy. Furthermore, the model can output classification criteria and their accuracies for all possible variable combinations. Regardless of the variables available to researchers, corresponding classification criteria can be derived for prediction. The model can also filter classification criteria based on desired classification prediction accuracy, providing simpler criteria.

2.2. Establishment Process of Submodel 1

2.2.1. Construction of Discriminant Function and Objective Function

The following construction process draws inspiration from Fisher's proposed Fisher's method [1].

The objective is to find a linear discriminant function that, when mapping the data from the two groups to the one-dimensional space it resides in, effectively separates the two classes by maximizing the inter-class data distance and minimizing the intra-class data dispersion (as shown in Figure 1).

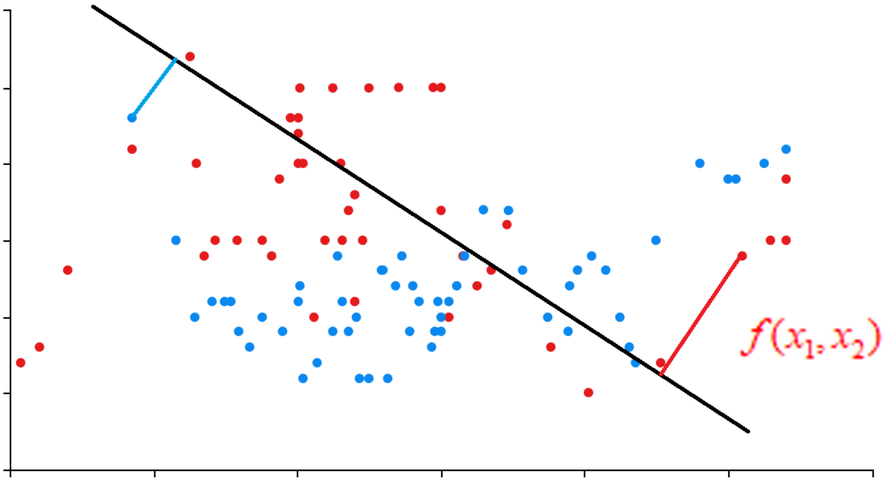

Figure 1: Illustration of Mapping by Discriminant Function

Firstly, the linear discriminant function is considered as follows :

$$X = \sum_{i=1}^{k} \lambda_i v_i \tag{1}$$

Where *k* represents the total number of variables, i.e., the dimensionality of the vector space in which the dataset resides. *vi* denotes different variables of the data, and *λi* represents the sought constants.

The difference in mean coordinates of the two classes' data on *X* can effectively reflect the inter-class distance of the processed data. Thus, the following function *D* is defined, where *di* is the difference in means of the two classes' data for the variable *vi*:

$$D = \sum_{i=1}^{k} \lambda_i d_i \tag{2}$$

To emphasize the minimization of intra-class dispersion in the processed data, that is, to encourage tighter clustering of data within each group, the following definition of S_{pq} is introduced:

$$S_{pq}=\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\left(x_{p1i}-\overline{x_{p1}}\right)\left(x_{q1j}-\overline{x_{p1}}\right)+\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\left(x_{p2i}-\overline{x_{p1}}\right)\left(x_{q2j}-\overline{x_{p1}}\right) \tag{3}$$

Where $x_{p1i}$and x_{p2i} represent the values of the first category for the p variable, while x_{q1i} and x_{q2i} represent the values of the second category for the q variable. n_1 and n_2 respectively denote the total number of data points in category 1 and category 2. Hence, S_{pq} assesses the dispersion of the two classes of data across the p and q variables. S is defined, which reflects the intra-class dispersion of the processed data for the two classes :

$$S=\sum_{p=1}^{k}\sum_{q=1}^{k}\lambda_p\lambda_q S_{pq} \tag{4}$$

With the above two functions, the classification problem can be transformed into an optimization problem. Since we aim for a larger inter-class data distance, implying a larger value for *D* is preferable. Simultaneously, since we desire a smaller intra-class dispersion, implying a smaller value for *S* is preferable, we define the following objective function:

$$\max\left(\frac{D^2}{S}\right) \tag{5}$$

2.2.2. Variable Selection

Based on the fundamental idea of the linear discriminant function, variables with high discriminative power should ideally exhibit a large inter-class distance and a small intra-class dispersion. According to this criterion, the following expression is formulated to quantify the discriminative power *C* of any variable x_i:

$$C_{xi}=\frac{\left|\mu_{xi1}-\mu_{xi2}\right|}{S_{xi1}{}^{2}+S_{xi2}{}^{2}} \tag{6}$$

The numerator represents the absolute difference between the means of the two categories under a certain variable, indicating inter-class distance. The denominator represents the sum of variances of the two categories under the same variable, indicating intra-class variance. Due to the substantial numerical differences between means and variances, normalization is performed separately on the numerator and denominator during calculations.

Based on practical requirements, a certain number of variables with strong discriminative power are selected for computation.

2.2.3. Calculation of Discriminant Function

The following calculation process is adapted from Fisher's proposed Fisher's method [1].

Taking the derivative of the objective function (5) yields the following equation:

$$\frac{1}{2}\frac{\partial S}{\partial\lambda}=\frac{S}{D}\frac{\partial S}{\partial\lambda} \tag{7}$$

Substituting (2) and (4), the following system of equations is obtained:

$$\sum_{j=1}^{k}S_{ij}\lambda_j=d_i,\left(i=1,2,\ldots,k\right) \tag{8}$$

This is a system of *k* linear equations with *k* variables, and k equations can determine a unique solution for the set of λ values, thereby deriving a unique solution for the discriminant function.

2.2.4. Construction of Classification Method

Once the discriminant function is obtained, it is necessary to establish the basis for classification. The discriminant function establishes a mapping relationship that projects a multidimensional dataset onto a one-dimensional line. Constructing a classification method on this one-dimensional line requires calculating the boundary point that separates the two categories. The approach adopted in this article comprehensively considers the impact of both inter-class dispersion and means. Dispersion is represented by the within-class standard deviation. Given that the data distribution characteristics of each category can be approximated by a normal distribution, the within-class standard deviation can roughly measure data spread. The proportion of standard deviations can be used to calculate the relative position of the boundary point between the two categories.

Consequently, the coordinates of the boundary value *Z* have the following expression:

$$Z = \mu_1 + (\mu_2 - \mu_1)\frac{s_1}{s_1 + s_2}, (\mu_2 > \mu_1) \tag{9}$$

Where μ_1 and μ_2 represent the means of category 1 and category 2, respectively, while s_1 and s_2 represent the standard deviations of category 1 and category 2. When predicting classification, if the discriminant function value of the data is greater than the boundary value, it is assigned to category 2; if it's smaller, it is assigned to category 1; and if they are equal, it could be assigned to either category.

2.3. Solving Submodel 1

2.3.1. Solving Problem 1

Problem 1 is an an attempt to use the submodel 1 to distinguish species five from other species using the FBNr(variable femoral pore number on the right side of lizards). With only one variable, data are on a one-dimensional axis, so the above method for constructing a classification method can be directly applied. By calculating the average values and standard deviations of each category's data and substituting them into (9), we obtain:

$$Z = 12.093 \tag{10}$$

Therefore, when the value of FBNr is less than 12.093, the lizard belongs to species five, and when the value is greater than 12.093, it belongs to other species. Since FBNr takes integer values, the case of equality doesn't need to be considered. Under this classification method, the classification accuracy can reach 99.3%.

2.3.2. Solving Problem 2

In this problem, the training set and test set have an 8:2 ratio, and the performance of the classification method is evaluated using the test set.

Problem 2 is an attempt to use the submodel 1 to find the most accurate variables for classifying species five and other species. As there are two variables, the discriminant function is given as follows:

$$X = \lambda_i x_i + \lambda_j x_j \tag{11}$$

Where *i* and *j* are two distinct integers ranging from 1 to 23, and *xi* and *xj* are two distinct morphological or pholidosis variables.

This yields a system of binary linear equations:

$$\begin{aligned} S_{ii}\lambda_i + S_{ij}\lambda_j &= d_i \\ S_{jj}\lambda_j + S_{ji}\lambda_i &= d_j \end{aligned} \tag{12}$$

By calculating the discriminative power *C* for each variable, two variables with the strongest discriminative abilities, MBS(number of dorsal scales) and VSN(ventral scale number on the middle line), are identified. Once the corresponding *λi* and *λj* are obtained, the original dataset can be mapped onto this discriminant function. Subsequently, by calculating the means and standard deviations of the processed data for the two categories and substituting them into formula (9), the classification coordinates can be determined. This classification method achieves a prediction accuracy of 100.0%.

The discriminant function is as follows:

$$X = -0.736v_1 - 0.667v_2 + 50.407 \tag{13}$$

Classification Criteria: Input the data into the discriminant function to obtain the corresponding X value. When X is greater than 4.001, the data belongs to species 5. When X is less than 4.001, the data belongs to species 1-4 or species 6-8. When X equals 4.001, the data could belong to any category.

2.4. Conclusion of Submodel 1

The above analysis presents the classification criteria and accuracy based on existing data for binary classification problems. Balancing accuracy and simplicity, recommended solutions and the theoretically maximum accuracy are provided. Biologists can adjust variables based on their accuracy requirements and computational environment to receive corresponding accuracy feedback. However, since the actual data comprises eight species, Submodel 1 cannot classify them individually. In Submodel 2, the Bayes algorithm from LDA is applied to achieve more detailed classification.

3. Model 2: Submodel for Multi-classification Problems

3.1. Overall Description of Submodel 2

Building upon Model 1, Model 2 extends the method for constructing classification criteria to address multi-classification problems. This article combines the research approach for multi-classification problems in Bayesian methods [2]. It calculates corresponding discriminant functions for each category and assigns a category based on the equation with the highest function value. To simplify classification criteria, Model 2 extends the method of measuring variable discriminative power proposed in Model 1, making it applicable to multi-classification problems for simplifying discriminant functions. Moreover, this model can also output classification criteria for all variable combinations, enhancing practical applicability.

3.2. Establishment Process of Submodel 2

3.2.1. Variable Selection

Building upon the fundamental idea of the linear discriminant function, variables with high discriminative power should ideally exhibit a large inter-class distance and small intra-class dispersion. Similar to Model 1, Model 2 extends the approach for assessing discriminative power, constructing the following expression to quantify the discriminative power C for any variable *vi*:

$$C_i = \frac{\sum_{j=1}^{g} \left| \mu_{ij} - \overline{x} \right|}{\sum_{j=1}^{g} s_{ij}^{\ 2}} \tag{14}$$

The numerator represents the sum of the absolute differences between the means of each category under a certain variable and the overall mean, indicating inter-class distance. The denominator represents the sum of variances of each category under the same variable, indicating intra-class variance. Due to the substantial numerical differences between means and variances, normalization is performed separately on the numerator and denominator during calculations.

3.2.2. Main Body of the Model

For the main body of Model 2, this article adopts the Bayesian method for construction [2][3].

According to Bayes' theorem, it can be derived that the posterior probability of a data point belonging to a certain category is proportional to the product of the prior probability and the probability density here:

$$p_i(a) = cf_i(a)\pi_i \tag{15}$$

Pi(*a*) represents the posterior probability that data point *a* belongs to category *i*, *fi*(*a*) signifies the probability density for data point *a*, *πi* stands for the prior probability that data points belong to category *i*, *c* represents a constant term.

Since the distribution of data within each category approximates a normal function, the probability density function can be expressed using the normal distribution function. Taking the logarithm of both sides of the equation and rearranging leads to the discriminant function concerning the classification of data point *a* into category *i*.

3.3. Solving Submodel 2

In this problem, the training set and test set have an 8:2 ratio, and the performance of the classification method is evaluated using the test set.

Problem 5 requires classifying lizards based on their species. With 8 different species, this constitutes a multi-classification problem. Initially, discriminative power C is used to select variables, and the 7 variables with the strongest discriminative power are chosen: CSN(collar scale number), SCGr(number of superciliary granules on the right), aNDSr(average number of dorsal scales along one abdomen scale near limb on the right), SVL(snout-vent length), ESD(length of the posterior half of the pileus), HW(width of the head before the tympanic hole), MO(mouth opening). At this stage, the prediction accuracy is 92.7%. Results and classification criteria are provided in the appendix.

When the number of variables is unrestricted, the combination of variables that yields the highest accuracy includes: CSN, FPNr, SCGr, SMr(number of scales between the masseteric shield and the supratemporal scale on the right), MTr(number of scales between masseteric and tympanum shields on the right), aNDSr, SVL, TRL(trunk length), HL(head length), ESD, HW, HH(head height near the occipital plate), MO, FFL(total forelimb length), HFL(total hindlimb length). This set of 16 variables achieves a prediction accuracy of 99.1%. The accuracy of the combination of variables selected based on discriminative power is 6.4% lower than the highest-performing combination across all possible combinations.

3.4. Conclusion of Submodel 2

Model 2 uses the Bayesian method to model multi-classification problems. The model produces a corresponding number of linear discriminant functions based on the number of categories. By calculating the category corresponding to the discriminant function with the highest value, data can be classified. The provided classification criteria exhibit high accuracy, all surpassing 90%. Moreover, linear discriminant functions are simple and interpretable. Biologists can use these classification criteria to support biological theories. Additionally, the model can output discriminant function parameters for all variable combinations, making it more versatile in practical applications.

4. Model Verification

4.1. Validity Analysis

(1) Through research, it was found that in the field of biological classification, accuracy above 90% is often considered excellent. All recommended classification solutions in this paper achieve an accuracy of over 90% [4][5].

(2) The model in this paper uses linear discriminant functions, which significantly simplifies the structure while meeting accuracy requirements. In comparison to methods like support vector machines and neural networks, the model is more concise and reduces computational time (as shown in Figure 2 and Figure 3).

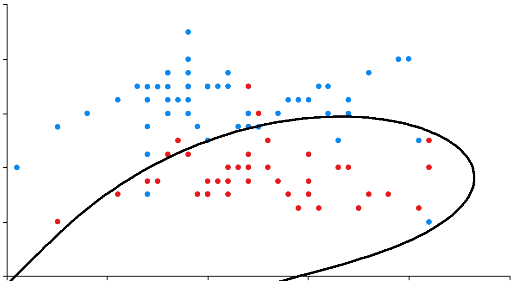

Figure 2: Illustration of Support Vector Machine's Discriminant Function

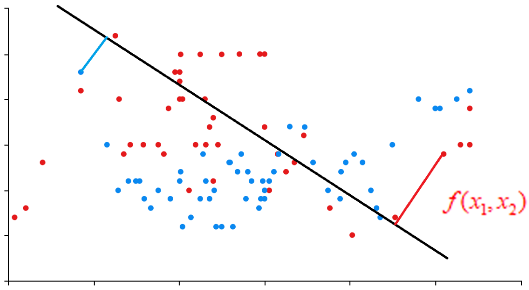

Figure 3: Illustration of Linear Discriminant Analysis's Discriminant Function

(3) After iterating through all possible variable combinations, the recommended solution prioritizing variables with strong discriminative power and the theoretically optimal solution differ in accuracy by only 1.8%. However, this selection method significantly reduces computational time by 99.2%.

4.2. Advantages of the Model

(1) The model's measurement of variable discriminative power C is scientifically accurate, and the accuracy of the predicted best variable combination closely aligns with the actual best variable combination.

(2) While satisfying the fundamental accuracy requirements, this classification model simplifies the structure. In contrast to support vector machines and neural networks, even with a high number of variables and categories, this model can still provide interpretable linear discriminant functions as classification criteria. It meets the needs of biologists for usability and aids in uncovering internal patterns behind observed phenomena.

(3) While greatly simplifying the model, this simplified model still fulfills the classification accuracy requirements common in biological research. Additionally, the flexibility of this classification model allows for customization based on individual biologist needs, striking a balance between simplicity and accuracy.

4.3. Disadvantages of the Model and Proposed Improvements

In multi-classification problems, the number of discriminant functions is equal to the number of categories. As the number of categories increases, the computation complexity also increases due to the larger number of discriminant functions to be calculated. Based on the Fisher method for multi-class situations, the number of calculated discriminant functions is the smaller value between (g-1) and k (where g is the number of categories and k is the number of variables). One improvement idea is to develop a method that generates a comprehensive discriminant function. By comparing this function with the original discriminant functions and aiming to minimize the loss of discriminative power, the optimal comprehensive discriminant function can be identified. It would only be necessary to calculate this comprehensive discriminant function, and based on different threshold values, multi-class data can be classified. This function would still maintain its linear characteristics and require only one calculation for classification, ensuring the simplicity of the classification criterion. Moreover, optimizing for the loss of discriminative power would also reduce the loss of model accuracy.

5. Conclusion

In this study, the Fisher algorithm from Linear Discriminant Analysis and the Bayesian algorithm were applied to solve the binary and multi-classification problems of lizard gender and species. While ensuring accuracy, the model was simplified to distinguish species solely through the linear combination of variables. This study introduced an index for variable discriminative power, streamlining variable quantity and computation processes, resulting in a 99.2% reduction in computation time. The succinct model benefits biologists by providing a classification method applicable under various computational conditions, aiding the development of interpretable theories. The study presented a linear discriminant equation with fewer variables while maintaining a prediction accuracy of at least 90%. Moreover, the model can provide classification criteria and prediction accuracy for all variable combinations, enhancing flexibility for diverse applications.

References

[1] Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7(2), 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.

[2] George H. Dunteman. (1984) Introduction to multivariate analysis. Thousand Oaks, CA: Sage Publications.

[3] Hardle, W., Simar, L. (2007) Applied Multivariate Statistical Analysis. Springer-Verlag Berlin and Heidelberg. pp. 289-303.

[4] Abbas, H., Altameemi, A., Farhan, H. (2019) Biological Landmark Vs Quasi-landmarks for 3D Face Recognition and Gender Classification. International Journal of Electrical and Computer Engineering (IJECE), 9(5), 4069-4076. https://doi.org/10.11591/ijece.v9i5.pp4069-4076.

[5] Wührl, L., et al. Diversity Scanner: Robotic Discovery of Small Invertebrates With Machine Learning Methods. Biorxiv, Cold Spring Harbor Laboratory, May 2021. Available from: https://doi.org/10.1101/2021.05.17.444523.