IMMC

Control Number: IMMC23253807
Problem: E

Summary Sheet.

Nowadays, biologists need to find a key set of features in order to distinguish a certain type of lizards, Darevskia, because they are sometimes difficult to discriminate. This paper gives a convenient way to use single feature classification. In addition, it also provides a method on selecting features and classifying in a multi-group and multi-feature condition.

In response to this problem, we develop our original **single feature classification** method. It measures the frequency of points appearing near the data that is going to be distinguished using a judging radius. All of the near points appearing in the judging circle are calculated. The most appropriate value of this judging radius can be computed by using **cross-analysis**. **Cross-analysis** is when we put already know data points to our model and check the accuracy of the algorithm.

We also develop an approach for **multiple feature classification**. First, since the quantity of the different features are different, we use min-maximum method to do the **normalization** of our data. Second, we discover that 26 features given by the problem require a huge amount of calculation, so we need to do **multi-feature selection** using a specific **determinant**. We consider the internal and the external length between groups so that the determinant measure a feature's ability in distinguishing different groups. The determinant has a positive correlation of external group length and a negative correlation with internal group length. Furthermore, we successfully verify our methods through our results in the **Linear Discriminant Analysis**(**LDA**). Third, in our **LDA** classification, by calculating all the eigenvector of the dimension conversion matrix, our data matrix is projected to a space in a lower dimension in this process. This process enables us to distinguish between the groups.

To present our results in a vivid way, we have shown the steps in each process in choosing features and using the LDA with clear graphs. What's more, we also show all of the accuracy for different features and evaluate each process of classification. For instance, our accuracy in distinguishing species5 is 98.3% and two of the most essential factors that determine the species5 are feature FPNr and HFL respectively. Therefore, we increase the flexibility for biologists to decide the number of features they will choose and which method they will use.

Keywords: single feature classification,cross-analysis, multiple feature classification, normalization, determinant, Linear Discriminant Analysis(LDA)

Contents

1 Introduction

1.1 Background

The ability to distinguish closely related species in the wild is crucial to biologists. With this ability, they are able to focus on the species relating to their research topic, thus saving a large amount of time. However, many closed related species have similar exterior appearance. Biologists have found some effective features to distinguish them, such as color. However, these features can not be quantified. Using those features will create difficulty and ambiguity for the scientific research, so during the process of researching in the wild, biologists are in desperate need to find out the species of a certain animal through an effective criterion. Besides, this criterion should base on some measurable features.

To solve this problem, the measurement data of 564 lizards of 8 species belonging to genus Darevskia are used and various math models are developed to create criteria that can identify the species of lizards with a great accuracy.

1.2 Problem Restatement

In this essay, the following problems will be coped with.

Question 1. Build a criterion based on Femoral Pore Number on the right side (FPNr), so the lizard5 can be distinguished from other lizard species as accurately as possible

Question 2. Build a criterion based on two of the variables in the measured morphometric and pholidosis characteristics, so that lizard5 can be distinguished from other lizard species as accurately as possible

Question 3. Build a criterion so as to distinguish lizards' sex, regardless of their species.

Question 4. Build a set of criteria which can distinguish lizard species living in the same area ,including:

(a) species6 and species7

(b) species1 and species2

(c) species3, species4 and species5

Question 5. Build a set of criteria to predict the sex and species of all lizards, as accurately as possible.

Besides, there are two general requirements. First, all the criteria should be accompanied by their performance metrics. Second, criteria should be relatively simple, so that a biologist can use them in the field condition only with a graphing calculator

2 Assumptions And Justifications

Assumption 1: We ignore the effects of the eighth species on different groups, which means that we can independently do a classification of the eighth species.

Justification 1: The sample size of the eighth group is small, so this assumption will largely decrease the calculating amount and will not affect the accuracy significantly.

Assumption 2: We assume that there is no other recessive variables in the process of distinguishing sex.

Justification 2: It requires a huge amount of calculation to calculate a expression that has large linear coefficients with sex, so to simplify our model, we will only use and analyze the data given.

Assumption 3: The set of data directly relates to our classification. Additionally, that means that the sample size will influence the categorizing results.

Justification 3: Since no any additional results are given, we will assume that the results we get from the 564 lizards have direct relationships between our classification to the lizards.

3 Variables

Variables	Descriptions
m_{ij}	the value of FPNr of the jth lizard in species$i$
r	judging radius
F_i	the frequency of the species$i$ within the circle
I_s	Index value
n_{ij}	a set of data in the ith attribute
x_{ij}	the standardized value of a_{ij}
y_{ij}	the sample average in the ith attribute
q	the quantity of q data points in the ith attributes
r_i	the difference between maximum and minimum of x_{ij}
$G_i(orG_j)$	the ith(jth) group
A_i	the lizard 's data for different features
n_j	the number of lizards in G_j
μ_i	the mean of each group G_j
μ	the total mean of all data
D	the between class matrix
I	in-class matrix
W	the matrix
λ_i	the designated eigenvalues of W
V_i	the eigenvectors of W
U	the whole set
t_{right}	all the correctedly classified individuals
t_{all}	all the classified individuals

4 Single Feature Classification Model

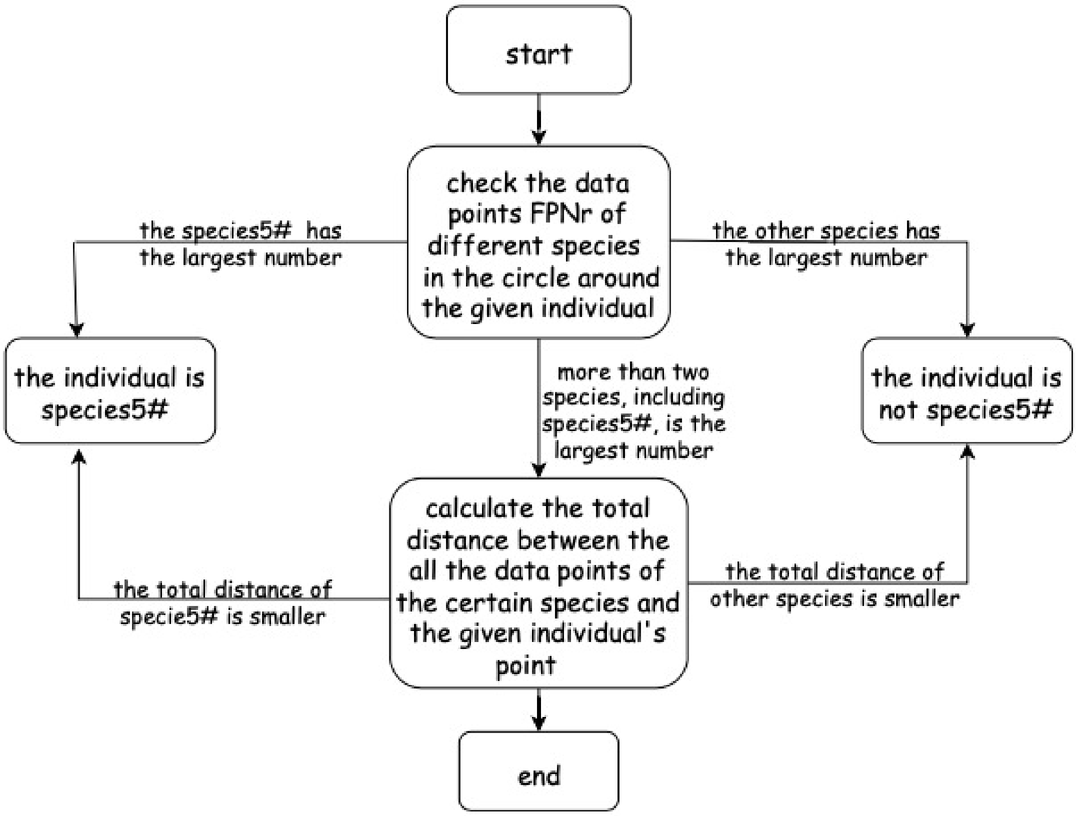

start
check the data points FPNr of different species in the circle around the given individual
the species5# has the largest number
the other species has the largest number
the individual is species5#
the individual is not species5#
more than two species, including species5#, is the largest number
calculate the total distance between the all the data points of the certain species and the given individual's point
the total distance of specie5# is smaller
the total distance of other species is smaller
end

Figure 1: The process of the solution of Task 1

The framework of distinguishing species5 by FPNr data is represented in Figure 1. Given a data point t, our goal is to decide whether t belongs to species5 or not. We use judging radius r to estimate this. Define set $K = \{x \mid t - r \leq x \leq t + r\}$. Then the number of all the eight species whose FPNr data $x \in K$ is calculated. The frequency of the species$i$ is F_i.

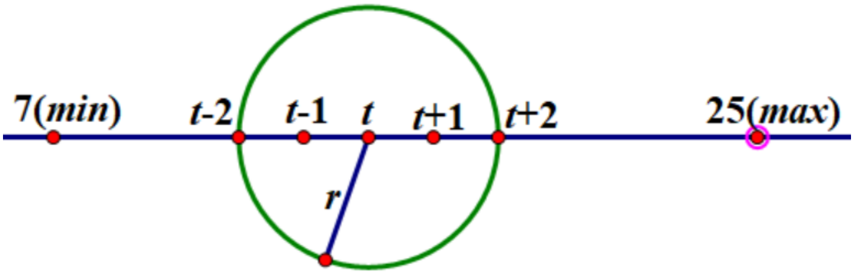


Figure 2: The sketch map of single feature classification

Rank these data and take the maximum F_j. For the different value of j, we have the following results:

Case 1. If there does not exist another F_l such that $F_l = F_j$.

(a) If $j = 5$, then t belongs to the 5th species.

(b) If $j \neq 5$, then t belongs to other species.

Case 2. If there exists at least one l such that $F_l = F_j$ for all $i \leq m$. Then we calculate index value I_s,where s is the species$s$.

$$I_s = \sum_{m_{sm} \in K, F_s = F_j} (|m_{sm} - t|) \tag{1}$$

We then rank I_s and find the smallest I_c.

(a) If $c = 5$, then t belongs to 5 species.

(b) If $c \neq 5$, then t belongs to other species.

We need to use cross-over analysis to determine the value of r. In addition, we will assume that a known data $m_{ij} = t$ and do the algorithm above. 10 percent of species 5 and 10 percent of other species are chosen for this analysis.

We will change r such that the accuracy of classification reaches the maximum value. And we conclude that the best r value is 2. The results and the evaluation is given in the Results part.

5 Multiple Feature Classification Model

The whole process of multiple feature classification is shown in Figure 3 below

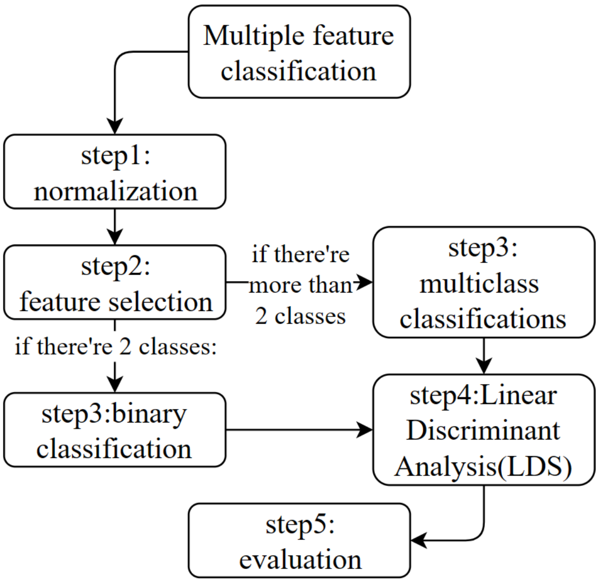

Multiple feature classification
step1: normalization
step2: feature selection
if there're more than 2 classes
step3: multiclass classifications
if there're 2 classes:
step3:binary classification
step4:Linear Discriminant Analysis(LDS)
step5: evaluation

Figure 3: The whole process for multiple feature classification

5.1 Normalization

To start with, a process called variable normalization will be done because the quantity of the numbers are not the same. For a set of data in the ith attribute a_{ij}, where $1 \leq j \leq Z$. We define the standardized value of a_{ij} as x_{ij}, where:

$$x_{ij} = \frac{a_{ij} - \min \left\{a_{ij}\right\}_{j=1}^{Z}}{r_i} \tag{2}$$

where r_i is the range in the data of the ith attribute.

$$r_i = \max \left\{a_{ij}\right\}_{j=1}^{Z} - \min \left\{a_{ij}\right\}_{j=1}^{Z} \tag{3}$$

5.2 Feature Selection

Definition 1. ***Distance between two Vectors***

A_1 and A_2 are two n dimension vectors.

$$\begin{aligned} A_1 &= (x_{11}, x_{12}, ..., x_{1n}) \\ A_2 &= (x_{21}, x_{22}, ..., x_{2n}) \end{aligned} \tag{4}$$

so the distance between x_1, x_2 is:

$$d(A_1, A_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2} \tag{5}$$

Definition 2. ***Group Average Distance*** $D(G_1, G_2)$

G_1 and G_2 are two groups with n dimension sample vector. $|G_1| = n_1$,$|G_2| = n_2$, then the group average distance of G_1 and G_2 is:

$$D(G_1, G_2) = \frac{1}{n_1 n_2} \sum_{A_i \in G_1, A_j \in G_2} d(A_i, A_j) \tag{6}$$

Definition 3. ***Internal Average Distance*** $I(G)$

We regard the Internal Average Distance in a group G as:

$$I(G) = \frac{\sum_{A_i, A_j \in G} d(A_i, A_j)}{|G|} \tag{7}$$

Assume a determine value for the ith attribute is D_i:

$$D_i = \frac{D(G_1, G_2)}{I(G_1) + I(G_2)} \tag{8}$$

Then rank D_i and we will use the m biggest features as the set of features.

Note that in the process of selection, Python is applied, which is a graphing calculator. This means that this part of criterion can be calculated on a graphing calculator and this part satisfies the requirement of "simple function" in the problem.

5.3 Linear Discriminant Analysis Categorization

5.3.1 The Whole Process

The Linear Discriminant Analysis (LDA) categorization is a mature method of categorization whose goal is to project the original data matrix onto a lower dimensional space. In this lower dimensional space, we can successfully discriminate different groups. We will briefly summarize the LDA process below. For more details of the process, check the reference.

Step 1. Assume that we have already normalized and chosen n features from the previous feature selection. Suppose the sample size is Z(in this case, $Z = 564$).Therefore, our data matrix can be represented as:

$$\begin{bmatrix} x_{11} & x_{12} & x_{13}\ldots\ldots & x_{1n} \\ x_{2,1} & x_{2,2} & x_{2,3}\ldots\ldots & x_{2n} \\ . & . & .... & . \\ . & . & .... & . \\ . & . & .... & . \\ . & . & .... & . \\ x_{Z1} & x_{Z2} & x_{Z3}\ldots\ldots.. & x_{Zn} \end{bmatrix} \tag{9}$$

Step 2. The groups we are going to distinguish are G_1,G_2,... and G_s. Each sample is represented by a point in the n dimension space. We also regard the data matrix as $(\{A_1, A_2, ...A_n\})^T$, where A_i is a lizard 's data for different features. First compute the mean of each group G_j.

$$\mu_j = \frac{1}{n_j} \sum_{A_i \in G_j} A_i \tag{10}$$

where n_j is the number of lizards in G_j.

Step 3. Then we'll compute the total mean of all data, which we call it μ.

$$\mu = \frac{\sum\limits_{i=1}^{Z} A_i}{Z} \tag{11}$$

Step 4. Calculate the between class matrix D:

$$D = \sum_{i=1}^{s} n_s(\mu_i - \mu)(\mu_i - \mu)^T \tag{12}$$

Step 5. Calculate in-class matrix I:

$$I = \sum_{j=1}^{s} \sum_{i=1}^{n_j} (A_{ij} - \mu_j)(A_{ij} - \mu_j)^T \tag{13}$$

where A_{ij} means the ith sample in the jth group.

Step 6. After that, we can calculate the matrix W:

$$W = I^{-1} D \tag{14}$$

Step 7. The eigenvalues (λ)eigenvectors (V) of W are then calculated.

Step 8. To reduce our calculating amounts, all of the eigenvectors are then used as a lower dimensional space (Vk).

Step 9. Project all the sample on to the lower dimension.

Note that in the process of categorization, R language is applied, which is a graphing calculator. This means that this part of criterion can be calculated on a graphing calculator and this part satisfies the requirement of "simple function" in the problem.

5.3.2 Binary Classification

When we are doing binary group classification, $s = 2$ in LDA. When all of the samples are projected on to the lower dimension, the deviation of all the groups will increase, enabling us to categorize each group.

An application of binary group classification is what we call exclude group classification, where distinguish one group from all of the groups. We assume that this group is G_1, and then we will do the binary group classification for G_1 and $\complement_U G_1$, where U is the whole set. This will distinguish G_1.

5.3.3 Multiclass Classification

In n group classification, we will use exclude group classification $n-1$ times.A detailed process of this is shown below:

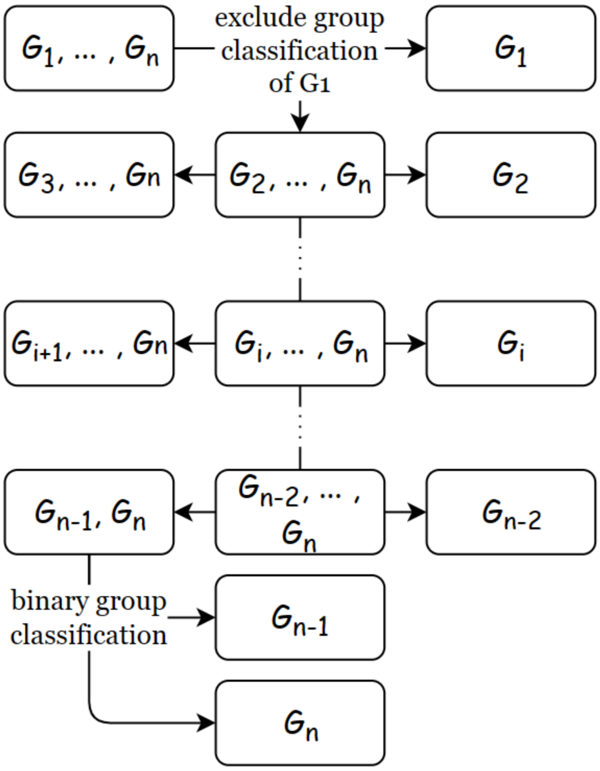


Figure 4: The process of multiclass group classification

We will also offer a detailed description of the circumstance when $n = 3$:

Assume that we need to distinguish G_1,G_2,G_3. We will then choose the two exclude group classification with the most accuracy, for example, we may exclude G_1 first, then exclude G_2, the rest is G_3. The process is shown in the following figure(Figure 4).

5.4 Evaluation

To evaluate the efficiency of our features and classification methods, an index P is defined, where t_{right} means all the correctly classified individuals and t_{all} means all the classified individuals:

$$P = \frac{t_{right}}{t_{all}} \tag{15}$$

Therefore, index P can be used to evaluate the accuracy of our classification.

6 Results

6.1 Task One

Step 1. Draw the of distribution of FPNr(the Figure 5). In this figure, it is obvious that feature FPNr can differentiate species5 effectively, as most of the individuals with FPNr value less than 11 are belong to the species5.

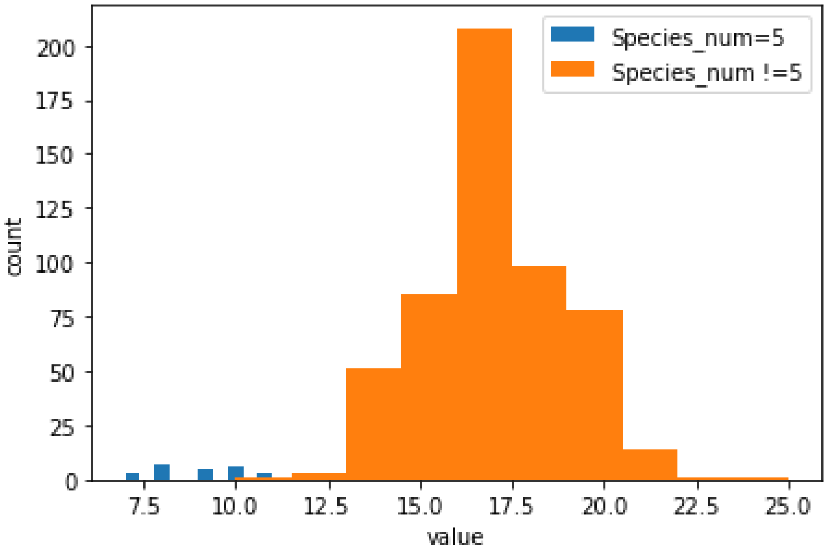


Figure 5: The distribution of FPNr

Step 2. Dot the data of FPNr on the axis

Step 3. Find a circle of radius r and calculate the frequency of all species in it.

Step 4. Conduct the cross-over analysis. The results of cross-over analysis in the single feature classification are represented in the table below for different values of r. Since we conduct the cross-over analysis for many times, we use the mean of result in the table.

Table 1: Specific data for $r = 2$

r=2	True species 5	True species 1-4,6-8
Classified as species 5	2.43	0.0
Classified as species 1-4,6-8	0.115	53.455

Table 2: Specific data for $r = 3$

r=3	True species 5	True species 1-4,6-8
Classified as species 5	2.165	0.265
Classified as species 1-4,6-8	0.115	53.455

Table 3: Specific data for $r = 4$

r=4	True species 5	True species 1-4,6-8
Classified as species 5	1.56	0.87
Classified as species 1-4,6-8	0.0	53.57

Step 5. Calculate P, which is the accuracy, and choose the r with the biggest P. In this task, $r = 2$

6.2 Task Two

Step 1. Normalize all the data

Step 2. Rank all the value of D_i of each feature in Figure 6 below. The bigger the D_i is, the more remarkable this feature is towards distinguishing G_1 and G_2.Therefore, FPNr and HFL are two most outstanding feature towards distinguishing G_1 and G_2.

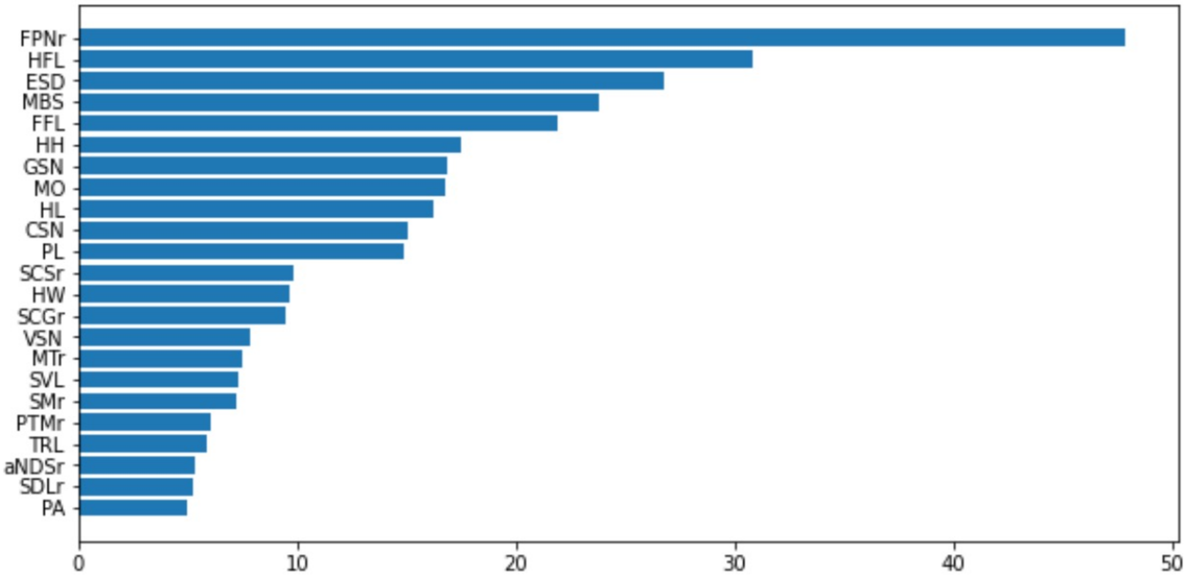


Figure 6: The D_i rank of features

Step 3. After using LDA, the coefficients of linear discriminants of feature FPNr and HFL have the greatest absolute value, which are −9.837 and 0.879, and the function of LDA is $z = -9.837FPNr + 0.879HFL$. This proves that our feature selection method is effective, as feature FPNr and HFL are also high in D_i

Step 4. Figure 7 below shows the result of LDA. The lizards of two groups, which are represented as triangles and dots, are clearly separated. This is also an evidence for the high effectiveness of our feature selection.

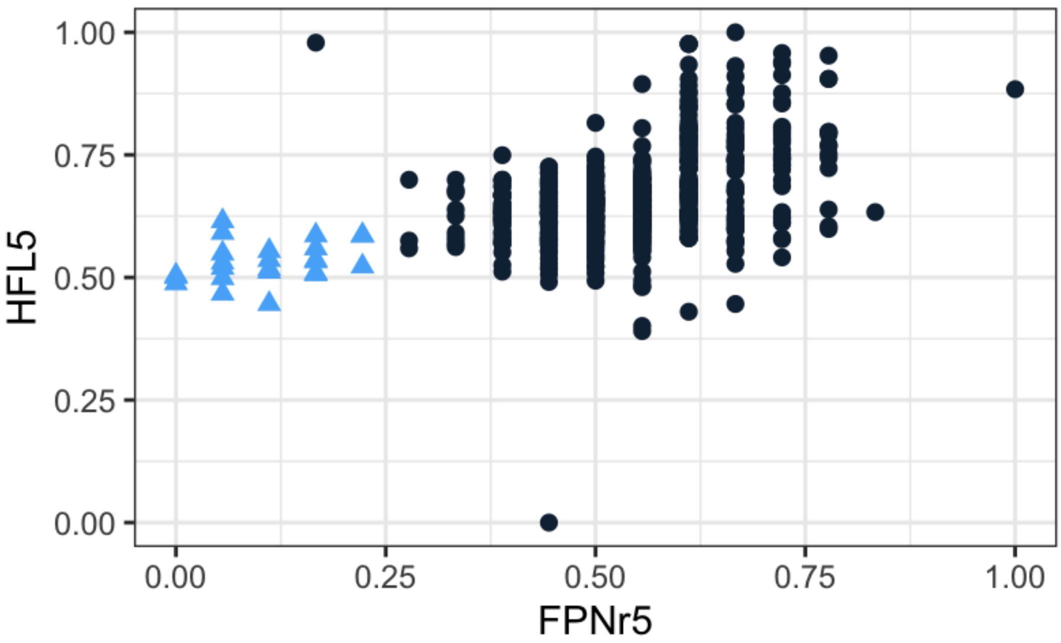


Figure 7: The classification result of task two

Step 5. Different pairs of features are chosen and their corresponding accuracy P, are shown in Figure 8 below. Therefore, the pair of FPNr and HFL is the most accurate pair.

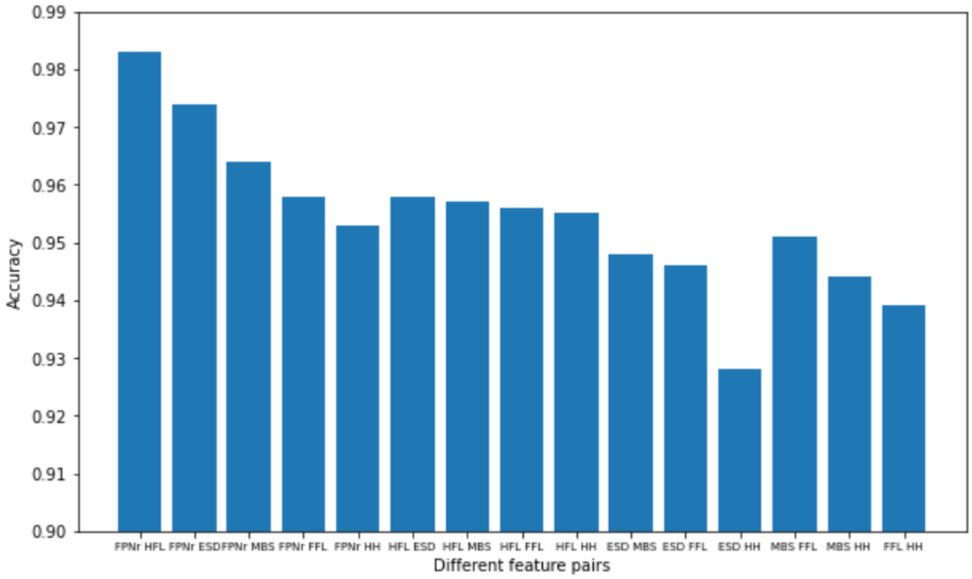


Figure 8: The accuracy of different feature pairs

6.3 Task Three

Step 1. Normalize all the data.

Step 2. Select features using binary feature selection. In this situation, the G_1 and G_2 represent male group and female group respectively. Figure 9 shows the ranking of determinant values D_i of different features

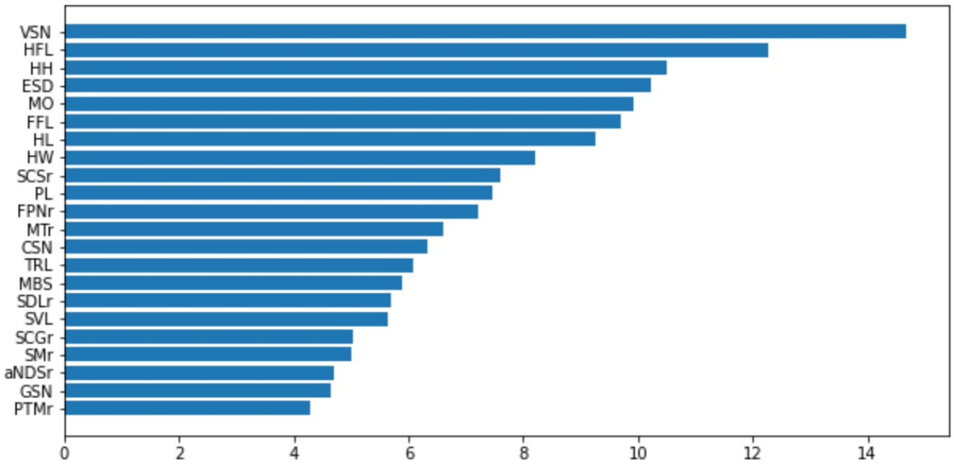


Figure 9: The D_i rank of features

The distribution of the data of different lizards, which are HFL paired with HH and VSN paired with HH specifically, are represented in Figure 10 and 11. We choose 10 percent of all the lizards and each dot represents a lizard's data:

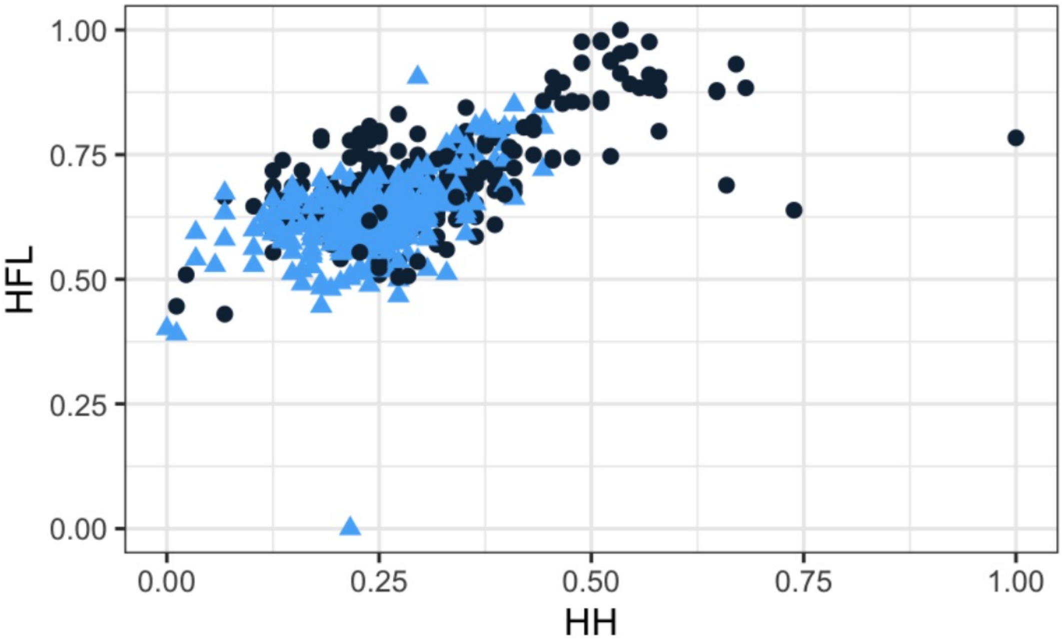


Figure 10: The figure about HFL and HH distribution

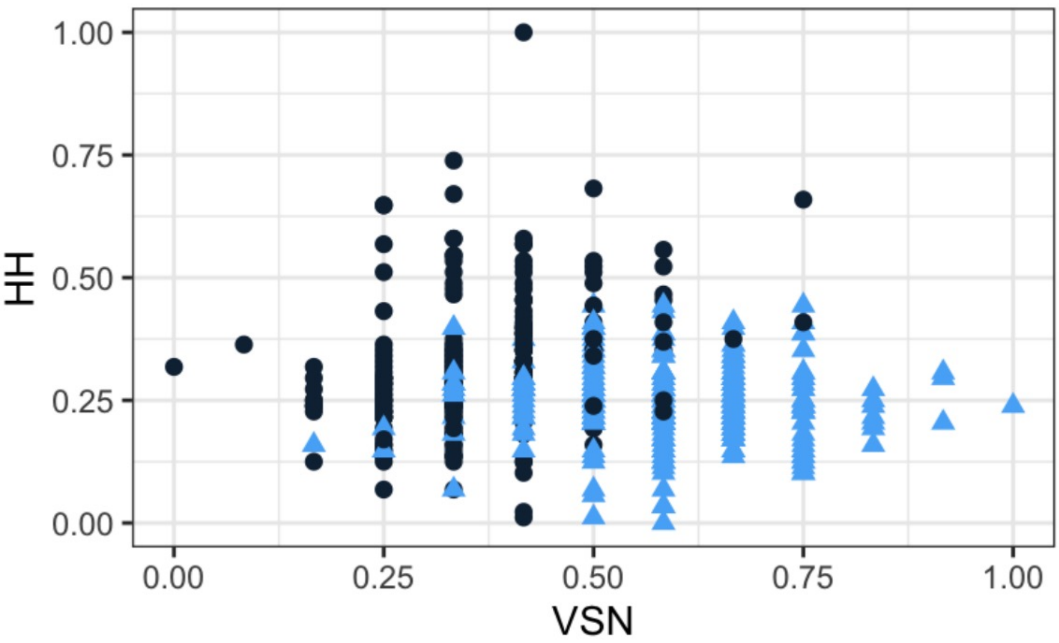
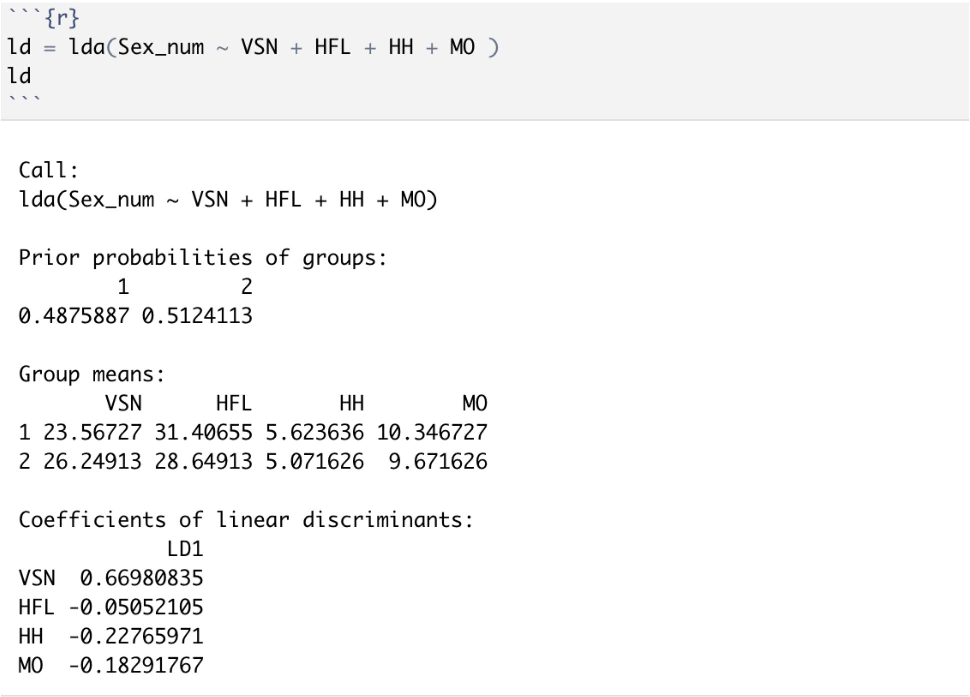


Figure 11: The figure about VSN and HH distribution

In the figure with HFL paired with HH, the two groups are mixed with each other so it is relatively difficult to distinguish the two groups. However,about the one paired VSH with HH, the distinctions between the two groups are more evident, which helps us distinguish them. According to the figure above, VSN is the most remarkable feature in distinguishing the sex.

When we choose the top four value, the data of LDA is VSN, HFL, HH and MO, as shown in Figure 12. The data indicates that:

$$z = 0.67VSN - 0.05HFL - 0.23HH - 0.18MO \tag{16}$$



~~~
```{r}
ld = lda(Sex_num ~ VSN + HFL + HH + MO )
ld
```
~~~

```
Call:
lda(Sex_num ~ VSN + HFL + HH + MO)

Prior probabilities of groups:
        1         2 
0.4875887 0.5124113 

Group means:
       VSN      HFL       HH        MO
1 23.56727 31.40655 5.623636 10.346727
2 26.24913 28.64913 5.071626  9.671626

Coefficients of linear discriminants:
            LD1
VSN  0.66980835
HFL -0.05052105
HH  -0.22765971
MO  -0.18291767
```

Figure 12: Coefficients of linear discriminants

The more the absolute value of the coefficient, the more effective that this feature has toward sex. This partly verifies our feature selection.

Step 3. Choose s features with the greatest D_i and use LDA to classify. We check the accuracy of the classification and draw the Figure 13, which shows the relationship between s and accuracy:

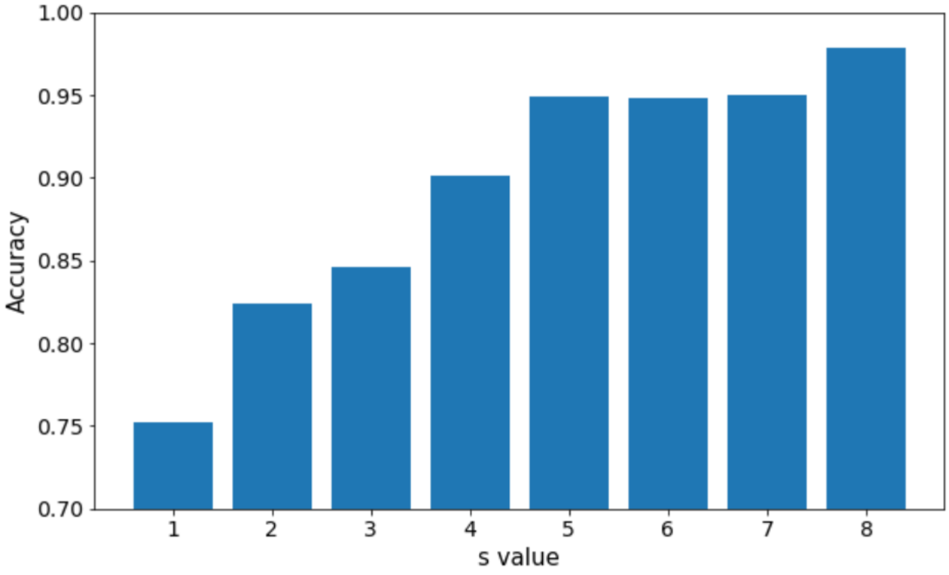


Figure 13: The accuracy of different sets of features

Therefore, approximately, accuracy rises to the highest when s equals to 8, and biologists can choose the value of s on their own, balancing the complexity of calculation and the accuracy of the result.

6.4 Task Four

In problem(a) and (b), the group number in LDA process will be 2. We will first select features using our previous feature selection process to decrease our calculation amount.

6.4.1 Problem A

After feature selection, we use the exclude group classification to distinguish species6 and species7.

We find out that between the species6 and the species7, feature HFL, FFl and ESD can distinguish them effectively. The figure of the distribution of the results of HFL and FFL are represented in Figure 14(the two largest D_i).

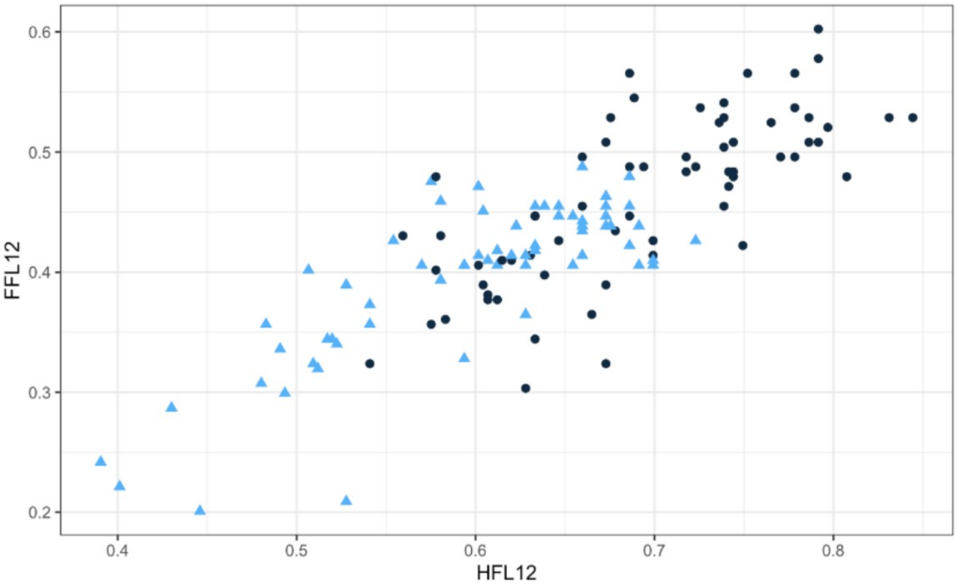


Figure 14: The figure about FFL and HFL distribution for task4a

We can see that the deviation between G_1 and G_2 is evident, allowing us to distinguish G_1 and G_2 effectively.

6.4.2 Problem B

After feature selection, we find out that feature HFL and GSN can distinguish the species1 and species2 effectively. The figure of the result of the data of different lizards are shown in Figure 15 below:

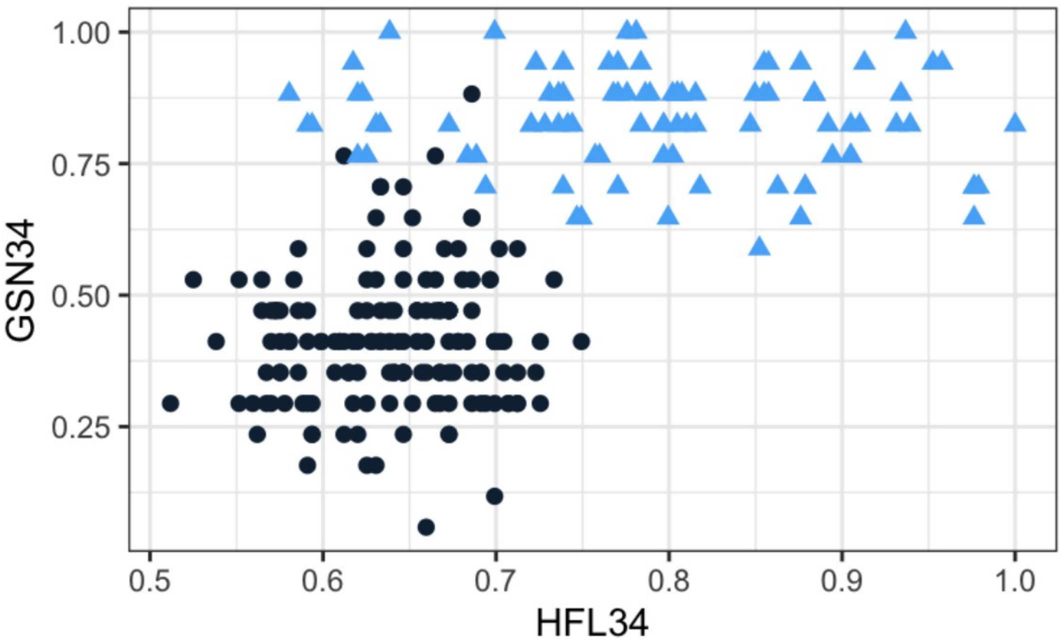


Figure 15: The figure about GSN and HFL distribution for task4b

6.4.3 Problem C

In problem(c), we will first distinguish species5 using the exclude group classification. After selecting species5, we will continue to use exclude group LDA process to distinguish the other two species. Feature MBS, SCGr and SDLr can distinguish species3 and species4 effectively and the graph of results are represented in Figure 16:

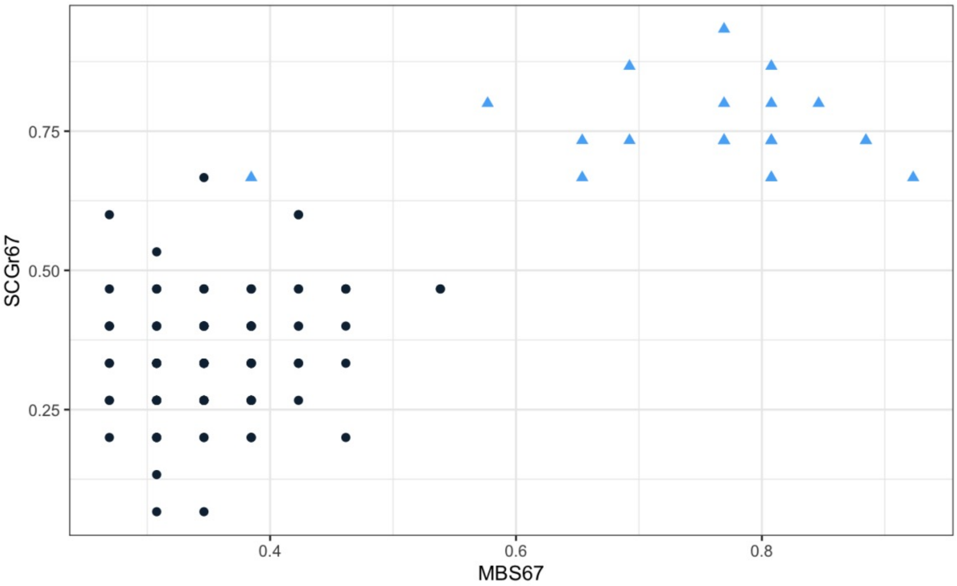


Figure 16: The figure about SCG and MBS distribution for task4c

All in all, the accuracy and the selecting features are compiled into a table:

Table 4: The accuracy of task four

Species	Feature1	Feature2	Feature3	Accuracy
6 and 7	MBS	SCGr	SDLr	0.9271
1 and 2	ESD	FFL	HFL	0.893
3 and 4	HFL	GSN	(NONE)	0.923

6.5 Task Five

The whole process of task five is shown in Figure 17 below.

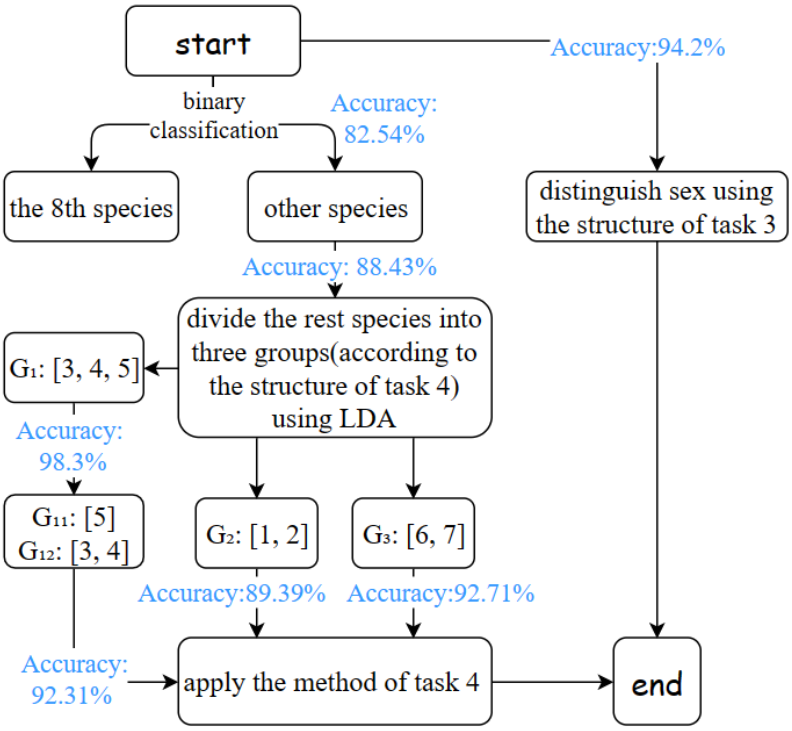


Figure 17: The process of Task 5

Case 1. Sex:

In order to distinguish sex, we can use the structure of task 3.

Case 2. Species:

In order to reduce our algorithm's complexity, we tried to use our basic structure of distinguishing species.

Step 1. We use a exclude group classification to exclude species8.

Step 2. Then a 3 group LDA is applied. The three groups are:

$$\begin{aligned} G_1 &= \{3, 4, 5\} \\ G_2 &= \{1, 2\} \\ G_3 &= \{6, 7\} \end{aligned} \tag{17}$$

The numbers in all of the sets represents the number of the species. LDA will distinguish the three groups.

Step 3. Then we will distinguish each species by applying the method of task 4, because we can see that the species in groups in this circumstance is the same as that of task 4.

Step 4. At last, we draw a graph about the index *P*, which means the accuracy of the classification, of the sixteen groups of species

Table 5: the accuracy of different kinds of lizards

Species	1	2	3	4	5	6	7	8
Male	0.615	0.615	0.624	0.624	0.676	0.637	0.637	0.778
Female	0.615	0.615	0.624	0.624	0.676	0.637	0.637	0.778

7 Strengths And Weaknesses

7.1 Strength

1. We invent a original way to select features in order to successfully reduce the calculation made by the computer while doing the LDA process. Furthermore, we verify this feature selection's credibility by using the result of the LDA process.

2. We use a simple and straightforward to distinguish species5 using the data of FPNr, and it is really accurate.

3. LDA process is really difficult when the group number is large, involving large amounts of matrix calculation. We successfully solve this problem by using the binary LDA process several times, decreasing the complexity of the LDA process.

7.2 Weakness

1. The judging radius *r* in task 1 that we choose is not accurate enough, it is easy to ignore several points when increasing a little.

2. We do not include the most appropriate feature for distinguishing sex because it is expected that sex is correlated with the ratios of some measured linear size, so we might ignore some essential features.

References

[1] Yu H, Yang J. A direct LDA algorithm for high-dimensional data—with application to face recognition[J]. Pattern recognition, 2001, 34(10): 2067-2070.

[2] Kaliontzopoulou A, Carretero M A, Llorente G A. Multivariate and geometric morphometrics in the analysis of sexual dimorphism variation in Podarcis lizards[J]. Journal of morphology, 2007, 268(2): 152-165.

[3] Guo G, Wang H, Bell D, et al. KNN model-based approach in classification[C]//OTM Confederated International Conferences" On the Move to Meaningful Internet Systems". Springer, Berlin, Heidelberg, 2003: 986-996.

[4] Izenman A J. Linear discriminant analysis[M]//Modern multivariate statistical techniques. Springer, New York, NY, 2013: 237-280.

[5] Patro S, Sahu K K. Normalization: A preprocessing stage[J]. arXiv preprint arXiv:1503.06462, 2015.

[6] https://usir.salford.ac.uk/id/eprint/52074/1/AI_Com__Tarek.pdf/