Agrupamiento Basado en Modelos en Frijoles Secos

Alberto Rodríguez

Universidad de Santiago de Chile. https://www.usach.cl/

Abstract. La base de datos "Dry Beans" corresponde a una serie de mediciones obtenidas a distintos frijoles no certificadas como de una única variedad. Se desarrolla un sistema de visión por computadora para distinguir siete variedades diferentes de frijoles secos. Se plantea utilizar la técnica de agrupamiento basada en modelos a través del paquete mclust del software R. Los resultados se comparan con lo encontrado en la literatura, comprobando la importancia de la mayoría de las variables para poder identificar a que clase pertenece. Esta investigación concluye con la obtención del modelo que mejor prediga la clase conforme al índice de BIC generado.

Keywords: Método de agrupamiento · Mclust · BIC.

1 Introducción

El frijol seco (*Phaseolus vulgaris L.*) es la legumbre más importante y producida en el mundo, por lo que desempeña un papel importante en la agricultura de Turquía [1]. La determinación de la mejor semilla es el principal problema para los productores de esta, por lo que es necesario determinar las características físicas como tamaño, color y diversidad. La clasificación manual de estas semillas de frijol es un proceso difícil y que consume mucho tiempo, por lo que se requiere métodos automáticos para clasificar y calificar.

Este artículo tiene la intención de usar el método de agrupamiento basado en modelos, el que permite el desarrollo de un modelo a través de la obtención de clústers de datos, con el objetivo de ver el agrupamiento que realiza el algoritmo respecto de los datos del dataset "Dry Beans", y que clases se pueden apreciar en cada grupo. Para realizar un contraste y aumentar la precisión dle estudio, se elabora un pre-procesamiento de los datos con correlaciones entre los atributos para descartar variables innecesarias. Se finaliza esta investigación con una discusión en base a los distintos modelos generados por el método de agrupamiento, como la obtención del mejor de estos, conforme al índice BIC generado.

2 Datos de la investigación

Esta conjunto de datos es obtenido por la Facultad de Tecnología de la Universidad de Selcuk. El objetivo del conjunto de datos es predecir según las características de cada frijol seco si este pertenece a una de las siete tipos de frijoles

secos. Se tomaron imágenes de 13.611 granos de las siete variedades, los conjuntos de datos describen características físicas de los frijoles y se tiene una variable resultado, la cual es una variable categórica la cual indica que tipo de frijol se tiene. Estas clases son:

- **Cali:** Es de color blanco, sus semillas son un poco gruesas y un poco más grandes que los frijoles secos y en forma de riñón.
- **Horoz:** Los frijoles secos de este tipo son largos, cilíndricos, de color blanco y generalmente de tamaño mediano.
- **Dermason:** Este tipo de frijoles secos, que son más planos, son de color blanco y un extremo es redondo y los otros extremos son redondos.
- **Seker:** Semillas grandes, de color blanco, la forma física es redonda.
- **Bombay:** Es de color blanco, sus semillas son muy grandes y su estructura física es ovalada y abultada.
- **Barbunya:** Fondo de color beige con rayas rojas o abigarrado, color moteado, sus semillas son grandes, la forma física es ovalada cerca de la redonda.
- **Sira:** Sus semillas son pequeñas, de color blanco, la estructura física es plana.

De los 13.611 frijoles, hay 1.322 (9,7% del total) que pertenecen a la clase Barbunya, 522 (3,8%) a Bombay, 1630 (11,9%) a Cali, 3546 (26%) a Dermason, 1928 (14%) a Horoz, 2027 (14,99%) a Seker y 2636 (19,3%) a Sira.

A continuación se enumeran y se describen los atributos del dataset, los cuales son todos numéricos. Se obtuvieron 12 características dimensionales y 4 de forma (todas medidas en milimetros).

- **Características dimensionales:**
  - **Área:** Zona de frijoles y el número de píxeles dentro de sus límites.
  - **Perímetro:** La circunferencia del frijol o longitud del borde.
  - **Longitud del eje mayor (L):** Distancia entre los extremos de la línea más larga.
  - **Longitud del eje menor (l):** La línea más larga que se puede trazar desde el frijol estando perpendicular al eje principal.
  - **Relación de aspecto:** Defina relación entre L y l.
  - **Excentricidad:** Excentricidad de la elipse que tiene los mismos momentos de la región.
  - **Área convexa:** N° de píxeles en el polígono convexo más pequeño que puede contener el área de la semilla de un frijol.
  - **Diámetro equivalente:** El diámetro de un círculo que tiene la misma área que el área de una semilla de frijol.
  - **Extensión:** La relación entre los píxeles del cuadro delimitador y el área del frijol.
  - **Solidez, Redondez, Compacidad:** Otras medidas del frijol
- **Características de forma:**
  - Shape Factor 1
  - Shape Factor 2
  - Shape Factor 3
  - Shape Factor 4

3 Pre-procesamiento

Tomando en cuenta los valores entregados por el dataset y con el objetivo de mejorar la precisión de la experiencia, se ha realizado un estudio de las variables. Primero se elimina la variable clase para no tener problemas posteriores en el agrupamiento (Este atributo lo utilizaremos después).

Segundo se verifica si el dataset se encuentra completo, es decir, que no hayas datos "perdidos" o algunas variables hayan sido rellenadas con un 0 o un "null". Al usar la función de R *apply*, se pudo verificar que a ninguna variable le faltan datos. Para complementar lo anterior también se busco que no existiera valores extremos (outliers) para esto se gráfico los datos de cada variable (ver **Figura 1**) y ver si existían outliers. Se puedo verificar que no tiene valores extremos.

3.1 Eliminación de variables

Eliminación del Área y Perímetro: A pesar de que la mayoría de los atributos son el resultados fórmulas que contienen como entrada otros atributos, es decir, atributos son dependientes de otros, estas son variables aleatorias y es importante ver si existe algún tipo de relación entre estas. Para esto evaluarlos utilizaremos una matriz de correlación.

Como se aprecia en la **Figura 2** existen muchas variables con varias relaciones entre ellas (Los colores azul y rojo mas oscuro indican que existe mayor relación), el coef. de correlación utilizado es Pearson ya que se trata solo de variables numéricas. Se puede observar que existe una relación muy grande entre el área y el perímetro, además estas tiene una relación muy grande con otras variables, así que se decide eliminar ambas

Eliminación de 4 variables más: Al realizar nuevamente una matriz de correlación (Ver **Figura 3**) Se puede apreciar que las variables "Compactness" y "Shape Factor3" tiene una fuerte relación cal igual que ConvexArea" y "Equiv-Diameter", por lo que se realiza un test de hipótesis para ambos, en donde:

- **H0:** No hay relación entre las variables
- **H1:** Hay relación entre las variables

Para realizar el test, se realiza un estudio del coef. de correlación de Pearson con 95% de confianza, el cual se resume en la siguiente tabla:

Table 1. Test de correlación entre Compacidad y Shape Factor 3.

Atributos	**Correlación**	**P-valor**	**Intervalo de confianza**
Compacidad / ShapeFactor3	0.9986	$< 2.2e\text{-}16$	[0.99864 0.99872]
ConvexArea / EquivDiameter	0.9852	$< 2.2e\text{-}16$	[0.98472 0.98571]

Con los resultados se cumple H1 y las cuatro variables esta relacionadas por lo que se eliminan.

Eliminación de Longuitud del eje menor y ShapeFactor1: Se realiza nuevamente una matriz de correlación (Ver **Figura 4**) y se encuentra relación entre los atributos "MinorAxisLength" y "ShapeFactor1" y se realiza el mismo test de hipótesis anterior obteniendo resultados similares, donde se cumple H1 y se eliminan ambas variables.

4 Método Utilizado

El método utilizado para realizar el agrupamiento de los datos existentes dentro del dataset corresponde al agrupamiento basado en modelos. Este modelo procesa la información y relaciones de los datos asumiendo alguna agrupación bajo algún modelo estadístico multivariado, esto con el objetivo de maximizar la verosimilitud y encontrar la distribución correspondiente de los datos usando el índice BIC pertinente, el cual nos permite tener el número de grupos adecuado a la complejidad del modelo.

5 Resultados de la investigación

Los resultados han sido obtenido luego de la realización de 3 pruebas, las cuales se diferencian según los atributos que se eliminan del dataset.

5.1 Prueba 1: Utilizando el dataset pre-procesado

Para la prueba 1, el dataset no fue modificado luego del pre-procesamiento. Tras aplicar el modelo sobre las variables del dataset, se han obtenido los siguientes 3 modelos con mejor BIC:

Table 2. Valores BIC Prueba 1.

	VVV, 9	VVV, 8	VEV, 9
BIC	-5347.9	-12379.9	-16296.8
BIC diff	0.0	-7032.2	-10949.1

5.2 Prueba 2: Eliminando variable ShapeFactor2

Al realizar una nueva matriz de correlación se observa que puede existir relación entre "ShapeFactor2" con 3 variables mas "Longitud del eje mayor", "Relación de aspecto" y "Excentricidad". Por lo que eliminamos esa variable y realizamos el método obteniendo los siguientes 3 modelos con mejor BIC.

Table 3. Valores BIC Prueba 2.

	VVV, 9	VVV, 8	VEV, 9
BIC	-56129.04	-64382.9	-65844.2
BIC diff	0.0	-8253.9	-9715.2

5.3 Prueba 3: Eliminando variable AspectRation y Eccentricity

Al igual que en la prueba anterior en la última matriz de correlación se puede ver que existe una relación entre la "Relación de aspecto" y la "Excentricidad" por lo que eliminamos ambas. Al aplicar el modelo se han obtenido los siguientes BIC

Table 4. Valores BIC Prueba 3.

	VVV, 9	VVV, 8	VEV, 9
BIC	-104702.9	-106177.7	-106402.6
BIC diff	0.0	-1474.7	-1699.6

5.4 El mejor BIC de la mejor prueba

La prueba 1 contiene el BIC más alto de las tres realizadas, donde la mejor configuración entregada por la función *MClust()* corresponde a VVV (Orientación, Forma y Volumen variables), y el mejor agrupamiento realizado es de 9 grupos. Para poder ver con mas detalles los otros BIC obtenidos, se recomienda observar la **Figura 5**. Además de estos resultados, también ha sido posible definir una configuración de los agrupamientos (Ver **Figura 6**) y una matriz de confusión del agrupamiento, definida a partir de la verificación de los sujetos de cada grupo.

Table 5. Matriz de confusión Prueba 1

Clase	1	2	3	4	5	6	7	8	9
Barbunya	1	1	50	14	0	109	10	0	1137
Bombay	0	0	0	0	520	0	0	0	2
Cali	0	23	1487	5	3	30	8	1	73
Dermason	2350	1	0	714	0	162	273	33	13
Horoz	5	1581	36	32	0	233	20	0	21
Seker	214	0	0	32	0	74	30	1665	12
Sira	199	19	36	916	0	97	1329	8	32

6 Discusión

De las tres pruebas realizadas, todas contienen la misma configuración VVV (Orientación, Forma y Volumen variables) y el mismo orden de agrupamiento, lo única que varia es el mejor BIC los cuales se diferencian bastante entre las pruebas, se escoge la prueba 1 por se la del BIC mayor. Con las pruebas 2 y 3 se puede apreciar que a medida que se eliminan mas variables el BIC va empeorando, esto se relaciona con lo dicho anteriormente, que como varios atributos son dependiente de otros, al ir eliminando algunos a pesar de estar relacionados va empeorando la agrupación de estos, aunque si se cuenta como muchos atributos el BIC puede ser muy bueno pero esto puedo causar sobreajuste.

De acuerdo a la matriz de confusión presentada en la Tabla 5, existen grupos que poseen muy una clase que esta muy por sobre las otra en cuanto a la cantidad de frijoles, ejemplo en el grupo 1 sobresale la clase Dermason, así como en el grupo 8 sobresale la clase Seker. Aunque, existen grupos que no tiene una cantidad elevada de una clase respecto a las otras como el grupo 4 y el grupo 6.

Viendo la distribución se puede inferir que tal como indica el estudio, las diferencias entre frijoles son características físicas tan pequeñas que un pequeño cambio en alguna variables puede clasificar un frijol en otra clase, entonces al ir eliminando variables hace mas difícil su clasificar entre los 7 tipos de variables diferentes.

7 Conclusiones

Con el objetivo de encontrar el modelo que mejor agrupe los datos, el pre-procesamiento anteriormente realizado, permite exponer la gran cantidad de correlaciones que hay entre los atributos, por lo que se tuvo que eliminar bastantes variables. El dataset de la prueba 1 el cual es el que salió del pre-procesamiento fue el que presento el mejor BIC de -5347. El segundo al cual se le extrajo la variable "ShapeFactor2" presentó el segundo mejor BIC de -56129.04. El tercer dataset al cual se le extrajeron 2 variables presentó el peor BIC de -104702.9. Las pruebas anteriores nos hacen concluir que a medida que se extraen variables empeora el agrupamiento, por lo que eliminar variables a pesar de que exista algún tipo de relación empeora el BIC, aunque una gran cantidad de variables nos puede llevar al sobreajuste, asi que igualmente se tiene que eliminar variables como se hizo en el pre-procesamiento.

Finalmente, al ver el agrupamiento realizado por el modelo, se puede concluir que cada clúster (a excepción del grupo 4 y 6) tiene mayormente una clase que las demás, por lo que puede notar "patrones" en las variables como para poder agruparlas en diferentes grupos.

References

1. Koklu, M., Ali Ozkan, I. (2020, mayo). Multiclass classification of dry beans using computer vision and machine learning techniques. https://www.sciencedirect.com/science/article/pii/S0168169919311573?via

8 Anexos

A A Análisis Estadístico

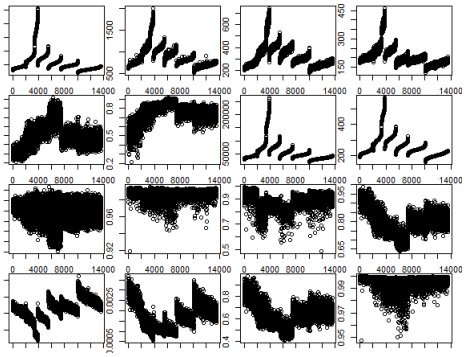

Fig. 1. Busqueda de Outliers en los datos de las variables

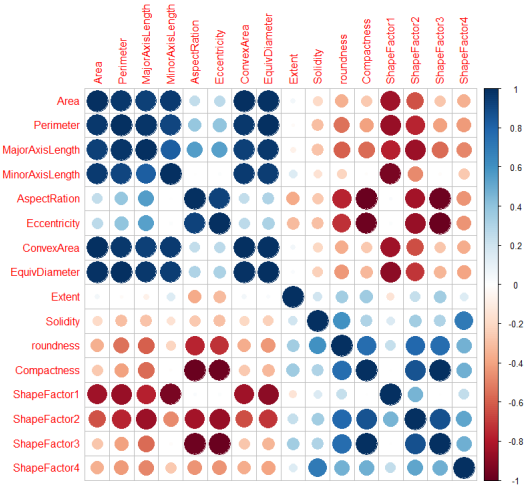

Fig. 2. Matriz de correlación con todas las variables

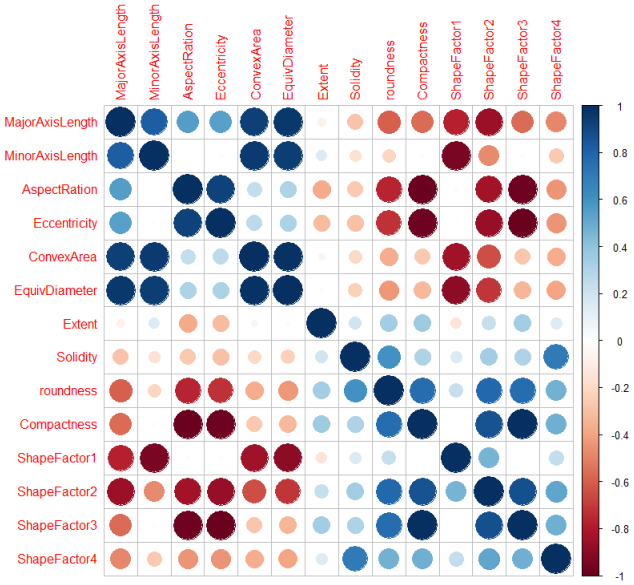

Fig. 3. Matriz de correlación con todas las variables

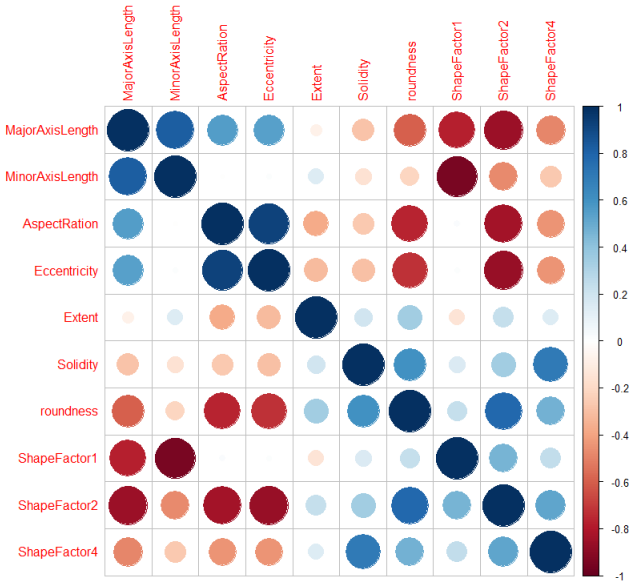

Fig. 4. Matriz de correlación con todas las variables

B B Aplicación MClust

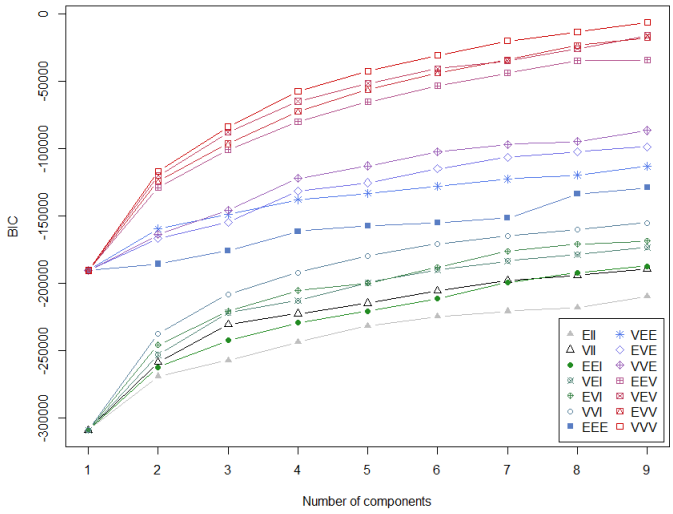

Fig. 5. Gráfico del ICL de los agrupamientos

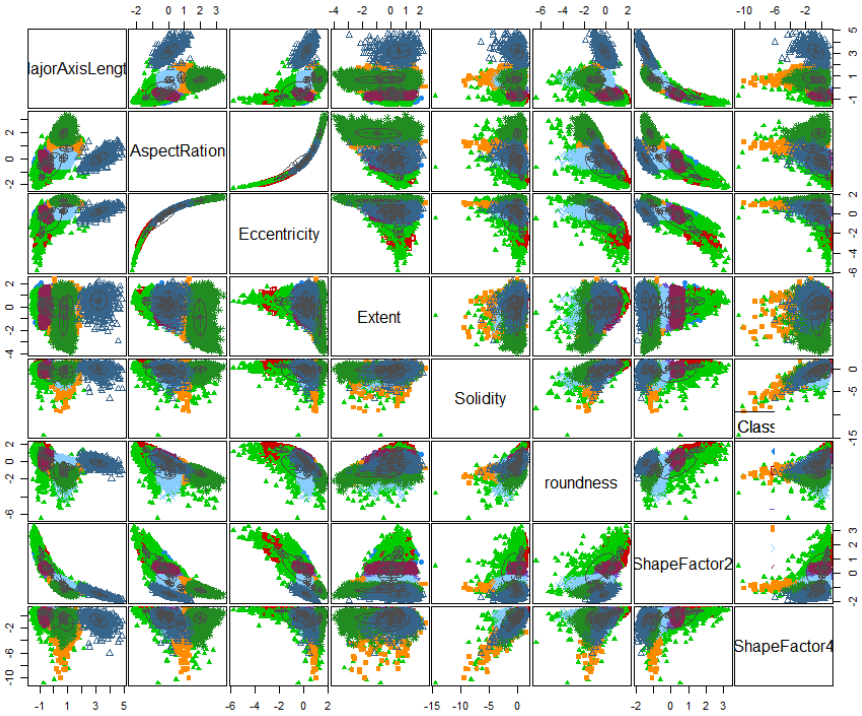

Fig. 6. Configuración de Agrupamientos