Introducción al reconocimiento de patrones 2013

José Luis Nunes
Matías Tailanián

Instituto de Ingeniería Eléctrica
Universidad de la República, Montevideo, Uruguay

Contents

Contents

Base de datos

Seguimiento realizado durante 9 meses sobre 891 vacas de 7 tambos diferentes.

Características Fenotípicas

- Edad.
- Condición corporal.
- Cantidad de partos.
- Anestro.
- Intervalo entre partos.
- Secado.
- Servicios.
- Concentración de progesterona.
- Cantidad de grasa en la leche.
- Cantidad de leche.

Características Genotípicas

Resumen

base de datos acotada y "limpia" con varias características fenotípicas que se quieren correlacionar con los genotipos de cada individuo.

Base de datos

Seguimiento realizado durante 9 meses sobre 891 vacas de 7 tambos diferentes.

Características Fenotípicas

- Edad.
- Condición corporal.
- Cantidad de partos.
- Anestro.
- Intervalo entre partos.
- Secado.
- Servicios.
- Concentración de progesterona.
- Cantidad de grasa en la leche.
- Cantidad de leche.

Características Genotípicas

Resumen

base de datos acotada y "limpia" con varias características fenotípicas que se quieren correlacionar con los genotipos de cada individuo.

Base de datos

Seguimiento realizado durante 9 meses sobre 891 vacas de 7 tambos diferentes.

Características Fenotípicas

- Edad.
- Condición corporal.
- Cantidad de partos.
- Anestro.
- Intervalo entre partos.
- Secado.
- Servicios.
- Concentración de progesterona.
- Cantidad de grasa en la leche.
- Cantidad de leche.

Características Genotípicas

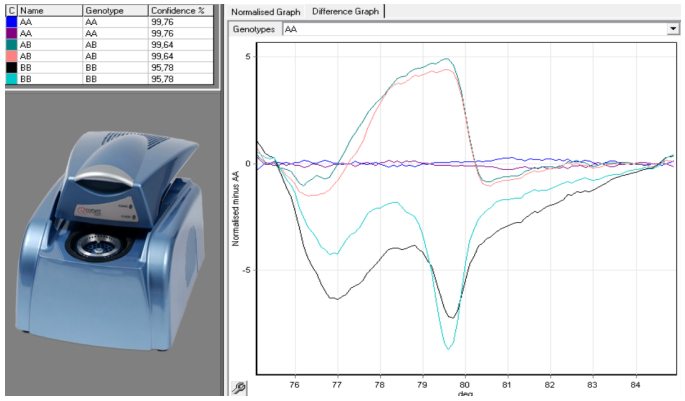


Figure: Determinación del genotipo

Clasificación en 3 clases: “AA”, “AB” y “BB”.

Resumen

base de datos acotada y “limpia” con varias características fenotípicas que se quieren correlacionar con los genotipos de cada individuo.

Base de datos

Seguimiento realizado durante 9 meses sobre 891 vacas de 7 tambos diferentes.

Características Fenotípicas

- Edad.
- Condición corporal.
- Cantidad de partos.
- Anestro.
- Intervalo entre partos.
- Secado.
- Servicios.
- Concentración de progesterona.
- Cantidad de grasa en la leche.
- Cantidad de leche.

Características Genotípicas

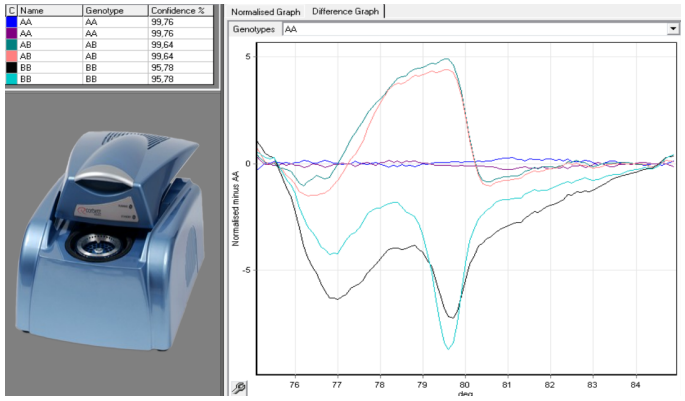


Figure: Determinación del genotipo

Clasificación en 3 clases: “AA”, “AB” y “BB”.

Resumen

base de datos acotada y “limpia” con varias características fenotípicas que se quieren correlacionar con los genotipos de cada individuo.

Contents

Etapa 1

Se abordará el problema como un trabajo de clasificación, tomando los genotipos como clases.

- Selección de características: método wrapper . Evalúa el set de atributos utilizando un esquema de aprendizaje y utiliza validación cruzada.
- Calsificadores
  - Árbol de decisión C4.5
  - Naive Bayes
  - k-NN

Resultados

	Tiempo [s]	$\sqrt{\text{MSE}}$	F-Measure	κ	Bien [%]
C4.5	0.84	0.45	0.331	0	49.83
Bayes	0.37	0.45	0.345	0.0044	49.60
k-NN	0.37	0.53	0.402	0.0043	45.23

```
a   b   c   <-- classified as
0 309   0 |   a = AA
0 444   0 |   b = AB
0 138   0 |   c = BB
```

- Resultado determinístico.
- El clasificador no funcionó adecuadamente.

Etapa 1

Se abordará el problema como un trabajo de clasificación, tomando los genotipos como clases.

- Selección de características: método wrapper . Evalúa el set de atributos utilizando un esquema de aprendizaje y utiliza validación cruzada.
- Calsificadores
  - Árbol de decisión C4.5
  - Naive Bayes
  - k-NN

Resultados

	Tiempo [s]	$\sqrt{\text{MSE}}$	F-Measure	κ	Bien [%]
C4.5	0.84	0.45	0.331	0	49.83
Bayes	0.37	0.45	0.345	0.0044	49.60
k-NN	0.37	0.53	0.402	0.0043	45.23

```
a   b   c   <-- classified as
0 309   0 |   a = AA
0 444   0 |   b = AB
0 138   0 |   c = BB
```

- Resultado determinístico.
- El clasificador no funcionó adecuadamente.

Etapa 1

Se abordará el problema como un trabajo de clasificación, tomando los genotipos como clases.

- Selección de características: método wrapper . Evalúa el set de atributos utilizando un esquema de aprendizaje y utiliza validación cruzada.
- Calsificadores
  - Árbol de decisión C4.5
  - Naive Bayes
  - k-NN

Resultados

	Tiempo [s]	$\sqrt{\text{MSE}}$	F-Measure	κ	Bien [%]
C4.5	0.84	0.45	0.331	0	49.83
Bayes	0.37	0.45	0.345	0.0044	49.60
k-NN	0.37	0.53	0.402	0.0043	45.23

```
a   b   c   <-- classified as
0 309   0 |   a = AA
0 444   0 |   b = AB
0 138   0 |   c = BB
```

- Resultado determinístico.
- El clasificador no funcionó adecuadamente.

Contents

Etapa 2

Dados los insatisfactorios resultados se decidió:

- Balancear las clases.
  Originalmente:
  - AA = 309
  - AB = 444
  - BB = 338
- Normalizar los descriptores.

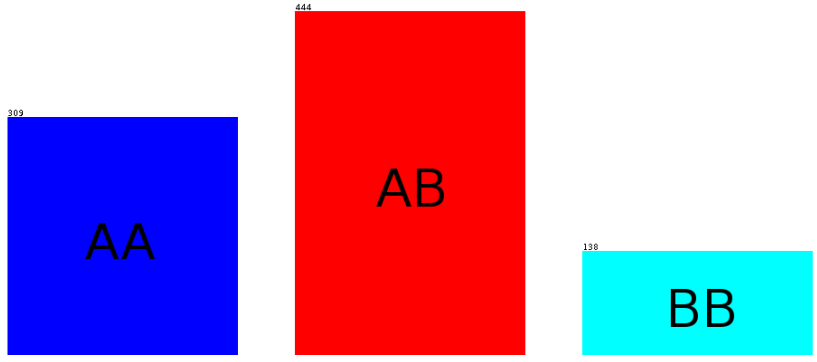


Figure: Desbalance entre clases

Resultados:

	Tiempo [s]	$\sqrt{\text{MSE}}$	F-Measure	κ	Bien Clasif [%]
C4.5	12.57	0.54	0.39	0.080	38.65
Bayes	11.77	0.50	0.30	0.044	36.23
k-NN	11.92	0.63	0.38	0.069	37.92

```
 a   b   c   <-- classif as C4.5
44  55  39 |   a = AA
51  54  33 |   b = AB
40  36  62 |   c = BB
```

Los resultados decayeron pero los algoritmos de clasificación respondieron acordemente a lo esperado.

Etapa 2

Dados los insatisfactorios resultados se decidió:

- Balancear las clases.
  Originalmente:
  - AA = 309
  - AB = 444
  - BB = 338
- Normalizar los descriptores.

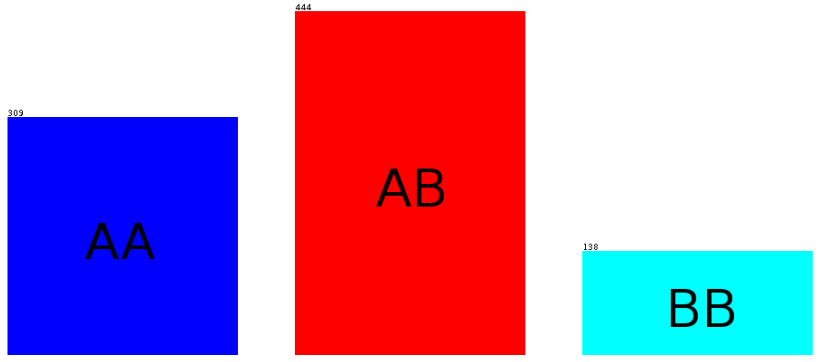


Figure: Desbalance entre clases

Resultados:

	Tiempo [s]	$\sqrt{\text{MSE}}$	F-Measure	κ	Bien Clasif [%]
C4.5	12.57	0.54	0.39	0.080	38.65
Bayes	11.77	0.50	0.30	0.044	36.23
k-NN	11.92	0.63	0.38	0.069	37.92

```
 a   b   c   <-- classif as C4.5
44  55  39 |   a = AA
51  54  33 |   b = AB
40  36  62 |   c = BB
```

Los resultados decayeron pero los algoritmos de clasificación respondieron acordemente a lo esperado.

Etapa 2

Dados los insatisfactorios resultados se decidió:

- Balancear las clases.
  Originalmente:
  - AA = 309
  - AB = 444
  - BB = 338
- Normalizar los descriptores.

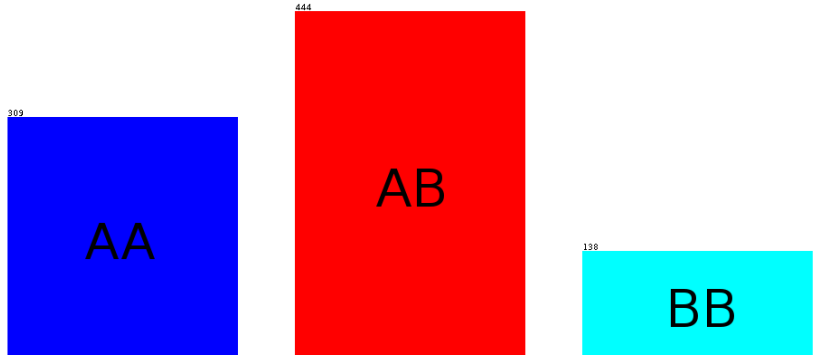


Figure: Desbalance entre clases

Resultados:

	Tiempo [s]	$\sqrt{\text{MSE}}$	F-Measure	κ	Bien Clasif [%]
C4.5	12.57	0.54	0.39	0.080	38.65
Bayes	11.77	0.50	0.30	0.044	36.23
k-NN	11.92	0.63	0.38	0.069	37.92

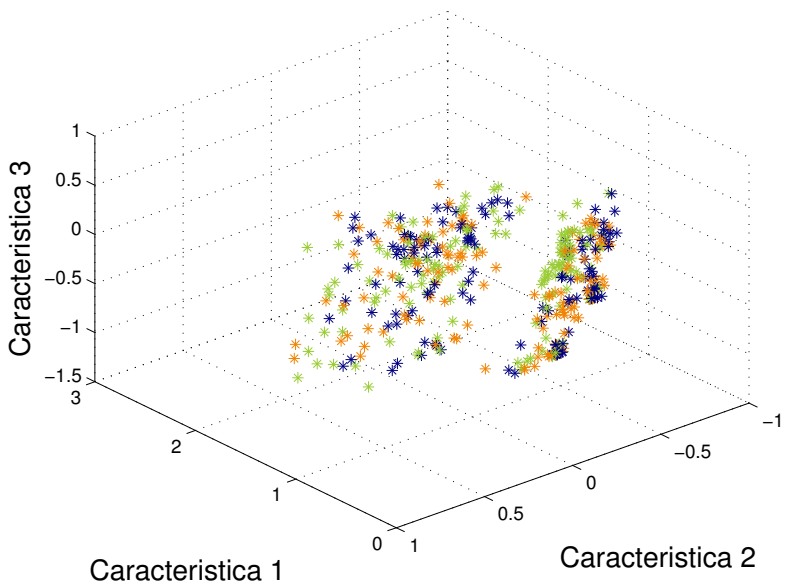
```
 a    b    c    <-- classif as C4.5
44   55   39 |   a = AA
51   54   33 |   b = AB
40   36   62 |   c = BB
```

Los resultados decayeron pero los algoritmos de clasificación respondieron acordemente a lo esperado.

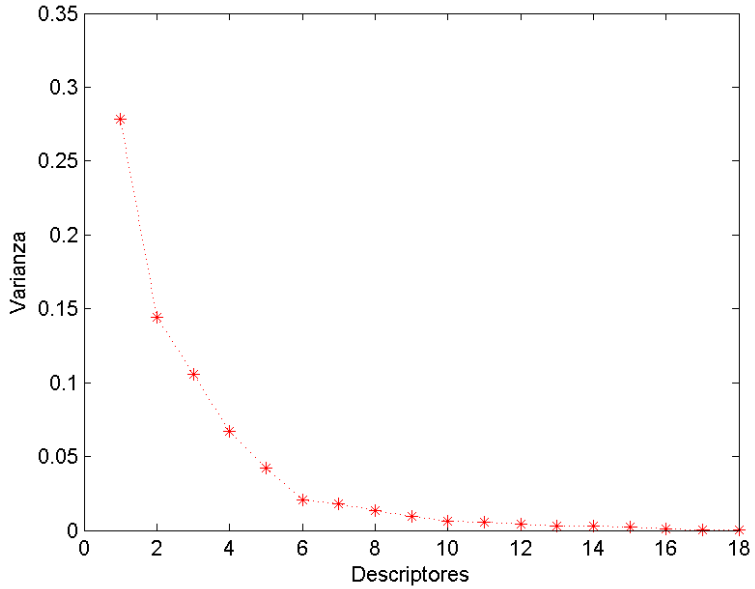
Contents

Etapa 3 — PCA

Resultados de aplicar PCA



(a) Características 1-2-3

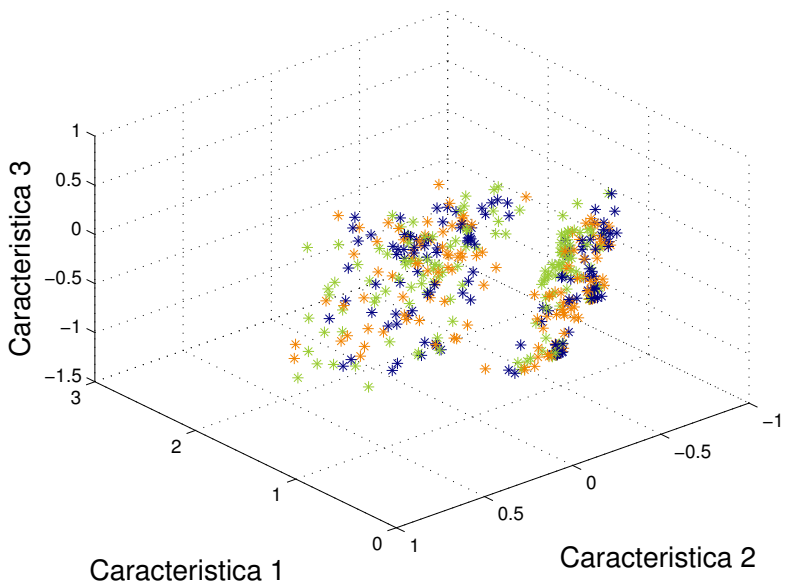


(b) Varianza vs Componentes

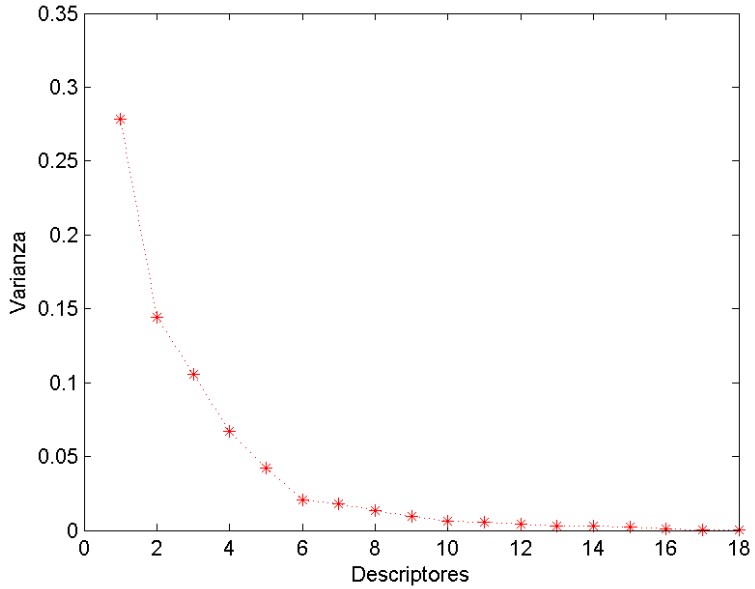
	Tiempo [s]	κ	Bien Clasif [%]
C4.5	0.03	0.143	42.75
Bayes	0.02	0.091	39.37
k-NN	0	0.149	43.24

Etapa 3 — PCA

Resultados de aplicar PCA



(c) Características 1-2-3



(d) Varianza vs Componentes

	Tiempo [s]	κ	Bien Clasif [%]
C4.5	0.03	0.143	42.75
Bayes	0.02	0.091	39.37
k-NN	0	0.149	43.24

Etapa 3 — LDA

Resultados de aplicar LDA

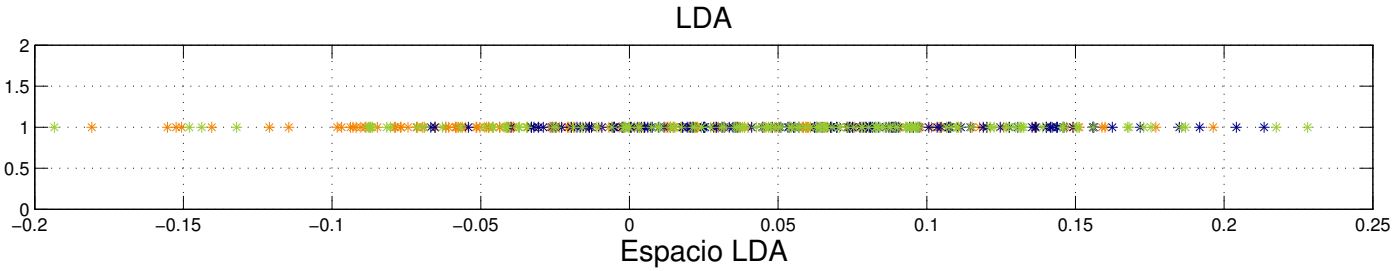


Figure: Proyección de los datos aplicando LDA

	Tiempo [s]	κ	Bien Clasif [%]
C4.5	0.03	0.1558	43.7198
Bayes	0	0.1667	44.4444
k-NN	0	0.1196	41.3043

Table: Resultados etapa 2

Etapa 3 — LDA

Resultados de aplicar LDA

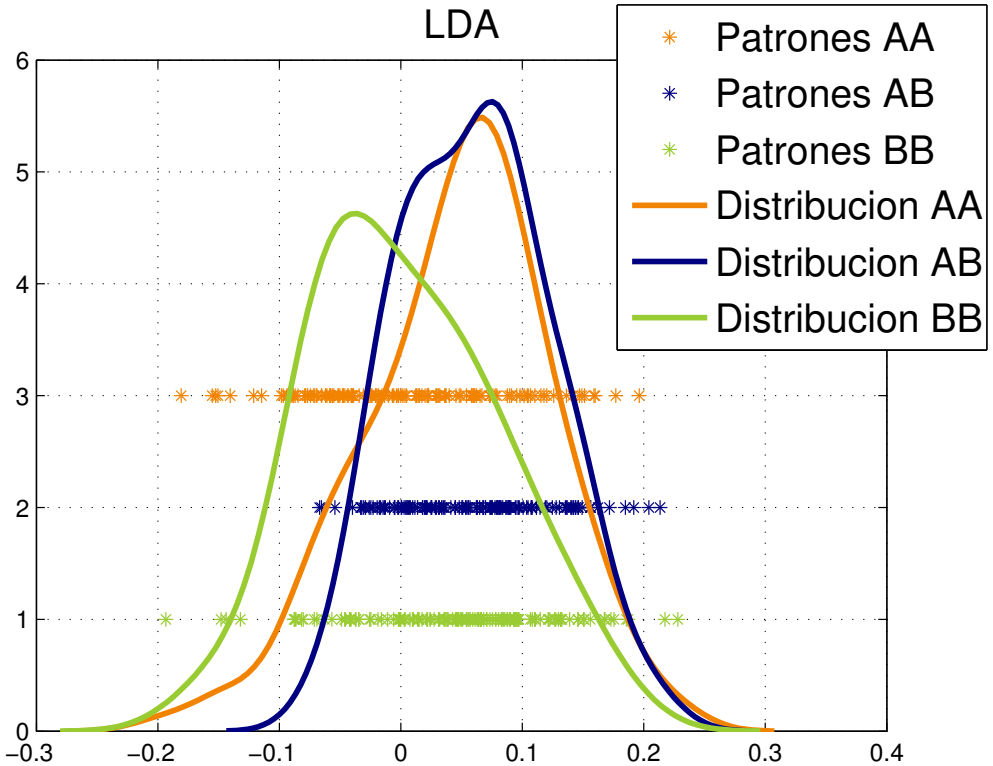


Figure: Estimación de la distribución de los datos

Se puede corroborar una vez más que las 3 clases son muy dificiles de separar, ya que presentan distribuciones realmente muy similares.

Etapa 3 — Diffusion Maps

Resultados de aplicar Diffusion Maps

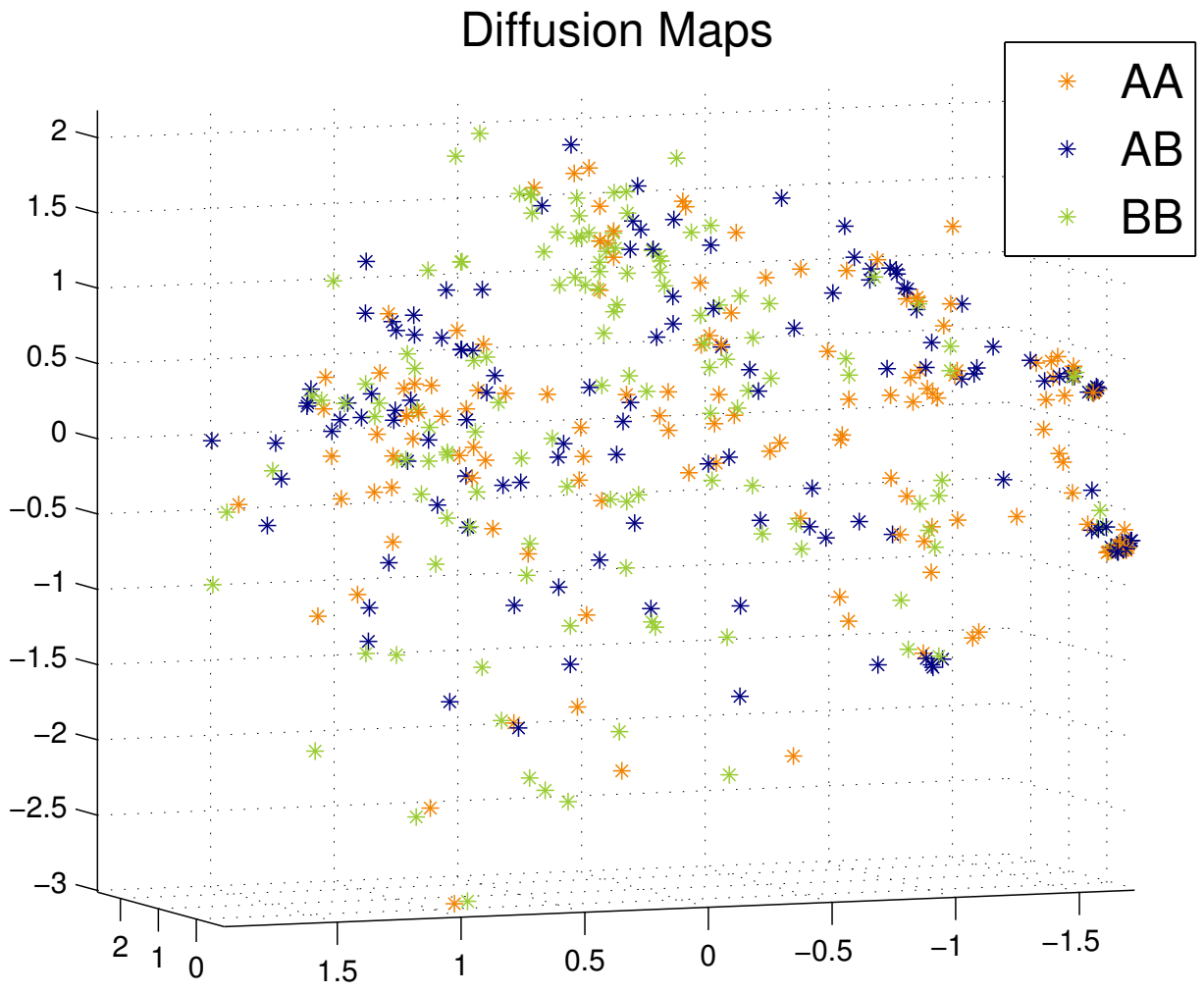


Figure: Diffusion Maps

Contents

Conclusiones

Conclusiones

- Pregunta ambiciosa, incluso para un genetista.
- Datos poco intuitivos, difíciles de interpretar
- Problema atacado con
  - Árbol de decisión C4.5
  - Naive Bayes
  - k-NN
  - PCA
  - LDA
  - Diffusion Maps
- En varias etapas
- Resultados contundentes

Trabajo a futuro

- REML

Conclusiones

Conclusiones

- Pregunta ambiciosa, incluso para un genetista.
- Datos poco intuitivos, difíciles de interpretar
- Problema atacado con
  - Árbol de decisión C4.5
  - Naive Bayes
  - k-NN
  - PCA
  - LDA
  - Diffusion Maps
- En varias etapas
- Resultados contundentes

Trabajo a futuro

- REML

¡Muchas gracias!

José Luis Nunes jlnunes@fing.edu.uy
Matías Tailanián mtailanian@fing.edu.uy