

Introducción al reconocimiento de patrones 2013

José Luis Nunes
Matías Tailanián

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

Contents

- 1 Base de datos
- 2 Primera etapa
- 3 Segunda etapa - Clases balanceadas
- 4 Tercera etapa - extracción de características
 - PCA
 - LDA
 - Diffusion Maps
- 5 Conclusiones

Contents

- 1 Base de datos
- 2 Primera etapa
- 3 Segunda etapa - Clases balanceadas
- 4 Tercera etapa - extracción de características
 - PCA
 - LDA
 - Diffusion Maps
- 5 Conclusiones

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

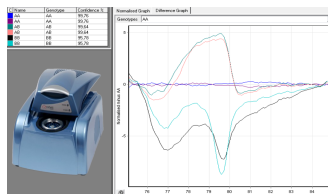


Figure: Determinación del genotipo

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

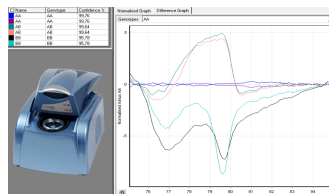


Figure: Determinación del genotipo

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

1 Base de datos

2 Primera etapa

3 Segunda etapa - Clases balanceadas

4 Tercera etapa - extracción de características

- PCA
- LDA
- Diffusion Maps

5 Conclusiones

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.

Se puede observar en la tabla anterior que se obtienen resultados muy similares para los 3 algoritmos. Para C4.5 y Bayes se obtiene un porcentaje de aciertos un poco menor al 50%, mientras que para k-NN los resultados son un poco inferiores.

La medida de error *kappa-statistic* (κ), es un indicador de la performance del algoritmo que tiene en cuenta las coincidencias por azar. Se calcula como

$$\kappa = \frac{P_0 - P_e}{1 - P_e}$$

donde P_0 es la proporción de coincidencias observadas y P_e la proporción de coincidencias esperadas en las hipótesis de independencia, es decir, coincidencias por azar. Se puede ver en la tabla ?? que se obtuvieron valores de κ realmente bajísimos, siendo este un indicador más de la mala performance alcanzada por los algoritmos.

Por otro lado resulta interesante analizar las matrices de confusión que resultan de estos algoritmos. Para el C4.5 se obtiene la siguiente matriz: Claramente el resultado obtenido no es el esperado. En este caso clasifica todos los patrones como pertenecientes a la clase "AB", y como esta clase representa casi el 50% de todas las muestras, el porcentaje de aciertos coincide. Este es un resultado determinístico, que más allá del porcentaje de aciertos, significa que el clasificador no funcionó adecuadamente. A su vez, analizando el árbol de decisión se puede ver que tiene una sola hoja.

Por otro lado las matrices de confusión para los algoritmos *Naive Bayes* y *k-NN* son las siguientes:

1	a	b	c	<— classified as	Bayes
2	6	300	3	a = AA	
3	8	435	1	b = AB	
4	1	136	1	c = BB	

1	a	b	c	<— classified as	k-NN
2	63	233	13	a = AA	
3	81	331	32	b = AB	
4	18	111	9	c = BB	

Si bien estos dos casos no se obtuvo un resultado determinístico como con C4.5, igualmente los resultados tienen un fuerte sesgo hacia la clasificación de los patrones como pertenecientes a la clase "AB".

Para mitigar el fenómeno de la salida determinística (y el sesgo) mencionado, el siguiente paso es atacar el problema del desbalance de clases. Para ello se realiza un sorteo aleatorio de las muestras pertenecientes a las 2 clases mayoritarias, de forma que las 3 clases tengan la misma cantidad de patrones.

Los resultados se presentan en la siguiente sección.

Contents

- 1 Base de datos
- 2 Primera etapa
- 3 Segunda etapa - Clases balanceadas
- 4 Tercera etapa - extracción de características
 - PCA
 - LDA
 - Diffusion Maps
- 5 Conclusiones

Las características seleccionadas son la condición corporal al momento del parto, a los 30, 45 y 90 días post parto, la cantidad de lactancias, la edad, el intervalo entre partos, progesterona y la cantidad de leche.

Los resultados de la segunda etapa, balanceando las clases, se muestran en la figura ??.

	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

Como primer ítem a mencionar se debe destacar el descenso en el porcentaje de aciertos, de algo más del 10%. Mientras que los resultados de la primera etapa arrojaban un porcentaje de aciertos de aproximadamente 49%, en esta etapa se nota un descenso hasta alrededor de los 38 puntos porcentuales. Aunque a priori parece un peor resultado, es interesante analizarlo con cuidado ya que por ejemplo el índice κ aumentó un orden de magnitud, aunque sigue siendo muy malo.

A diferencia de los resultados de la etapa anterior, para el caso del algoritmo C4.5 se obtiene un árbol no trivial con una cantidad total de 147 nodos y 74 hojas. Si bien la clasificación se encuentra por encima de una clasificación aleatoria, son resultados realmente muy malos.

El balanceo de clases logró solucionar el problema del resultado determinístico (o el fuerte sesgo) donde se clasificaba todos los patrones (o casi todos) como

pertenecientes a la clase "AB", y la matriz de confusión se muestra a continuación:

1	a	b	c	<— classified as	C4.5
2	44	55	39	a = AA	
3	51	54	33	b = AB	
4	40	36	62	c = BB	

1	a	b	c	<— classified as	Bayes
2	10	102	26	a = AA	
3	6	107	25	b = AB	
4	19	86	33	c = BB	

1	a	b	c	<— classified as	k-NN
2	45	52	41	a = AA	
3	54	49	35	b = AB	
4	31	44	63	c = BB	

Dado el terriblemente bajo porcentaje de aciertos obtenido en esta etapa, en la siguiente etapa se intentará realizar una extracción de características con métodos más sofisticados e intentar explicar el por qué de los resultados tan

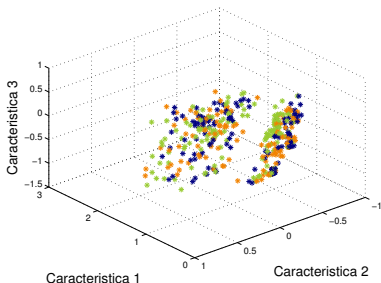
malos.

Contents

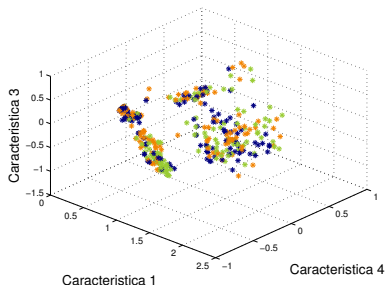
- 1 Base de datos
- 2 Primera etapa
- 3 Segunda etapa - Clases balanceadas
- 4 Tercera etapa - extracción de características
 - PCA
 - LDA
 - Diffusion Maps
- 5 Conclusiones

En la tercera etapa buscaremos realizar extracción de características con el fin de reducir la dimensionalidad y buscar características con mayor discriminación. Esto tiene como fin reducir los niveles de redundancia entre las características, visualizar características latentes significativas y generar para el futuro una mayor compresión en el proceso de generación de datos.

El algoritmo **PCA** (Análisis de componentes principales) tiene como fin encontrar la base de vectores que mejor exprese la distribución de los datos en el espacio completo. Es similar a encontrar las componentes ortogonales de un vector en un espacio, o lo que es igual, encontrar un conjunto de vectores que combinados en forma lineal representen los elementos. Estos elementos son los vectores propios de la matriz de covarianza correspondiente al espacio original. **PCA** tiene como fin encontrar un subespacio principal en el cual se maximice la varianza de los datos proyectados. Se ordenan las variables de acuerdo a la cantidad de varianza que concentran y se utilizan solamente las más significativas.

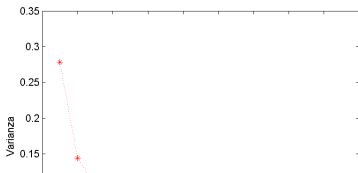


(a) Características 1-2-3



(b) Características 1-4-3

A simple vista resulta muy difícil reconocer algún tipo de estructura sobre los datos, con lo cual es de esperar que la clasificación no entregue mejores resultados de los ya vistos. La distribución de los datos en ambos subespacios (ver figura ??) resulta prácticamente randómica y es imposible identificar visualmente algún cluster por clases. (Esto es medio truco, pero sincero...)



En la figura 2 vemos la varianza en función de los componentes, podemos apreciar como la caída es abrupta y tiene sentido trabajar en el espacio de los

primeros tres componentes
que acumulan la
mayor cantidad de varianza.
Los resultados de aplicar
los clasificadores a los datos
procesados con PCA se

muestran en la tabla ??.

	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

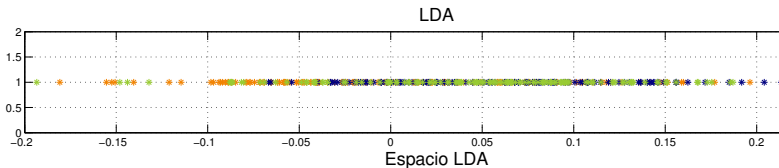
Las matrices de confusión para cada clasificador son:

1	a	b	c	<— classified as	C4.5
2	50	42	46	a = 1	
3	43	59	36	b = 2	
4	39	31	68	c = 3	
1	a	b	c	<— classified as	Bayes
2	19	93	26	a = 1	
3	22	85	31	b = 2	
4	18	61	59	c = 3	
1	a	b	c	<— classified as	k-NN
2	58	45	35	a = 1	
3	44	59	35	b = 2	
4	44	32	62	c = 3	

Como era de esperarse analizando la distribución de los datos, se obtuvieron resultados muy malos, aunque sensiblemente mejores que en la etapa anterior, llegando en el caso de k -NN a un porcentaje de aciertos de un poco más del 43%.

A su vez, la medida estadística κ aumentó un orden de magnitud respecto a la etapa anterior, y dos órdenes respecto a la etapa 1.

El algoritmo LDA (Análisis de discriminantes lineales) tiene como fin seleccionar una proyección que maximice separabilidad inter-clases. Busca una proyección de los datos en un espacio de menor (o igual) dimensión que las iniciales con el fin de que la discriminabilidad inter-clases sea lo más alta posible. Es una técnica supervisada ya que para poder buscar dicha proyección se debe entrenar el sistema con patrones etiquetados.



En la figura ?? se muestran los resultados de aplicar LDA a los datos. Nuevamente resulta imposible obtener algún resultado visualizando la distribución de los datos, pero de todas formas se intenta realizar una clasificación con los algoritmos C4.5, Naive Bayes y k -NN. Estos resultados se

muestran en la tabla ??

	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

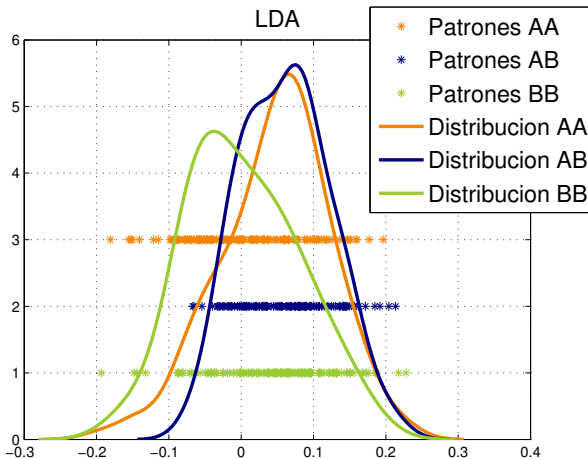
Como era de esperarse al analizar los patrones en el espacio transformado, nuevamente no es posible realizar una buena clasificación con ninguno de los clasificadores probados.

Las matrices de confusión para cada clasificador son:

1	a	b	c	<— classified as	C4.5
2	7	107	24	a = 1	
3	12	122	4	b = 2	
4	5	81	52	c = 3	
1	a	b	c	<— classified as	Bayes
2	10	85	43	a = 1	
3	7	94	37	b = 2	
4	3	55	80	c = 3	
1	a	b	c	<— classified as	k-NN
2	64	48	26	a = 1	
3	66	54	18	b = 2	
4	43	42	53	c = 3	

Para corroborar de otra forma que los datos son efectivamente inconcluyentes se presenta la figura ??, donde se puede ver con asteriscos los patrones representados en el espacio LDA y con líneas sólidas las distribuciones estimadas por clase. La representación de los patrones se realizó utilizando un

color y un nivel distinto para cada clase para ayudar a la claridad de visualización.

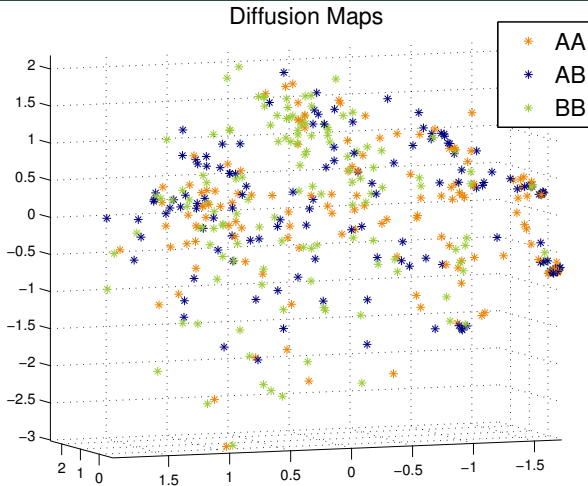


Se puede corroborar una vez más que las 3 clases son muy difíciles de separar, ya que presentan distribuciones realmente muy similares.

Con el objetivo de probar alguna técnica más sofisticada se utiliza Diffusion Map [?].

Diffusion maps es un algoritmo de *machine learning* que computa una familia de conjuntos de datos en un espacio embebido, usualmente de baja dimensión, cuyas coordenadas pueden ser calculadas de los vectores y valores propios de un operador de difusión de los datos. La distancia euclídea entre puntos en el espacio embebido es la “distancia de difusión”. A diferencia de otros métodos de reducción de dimensionalidad como PCA, este algoritmo es un método no lineal que se centra en descubrir *manifold* subyacente al muestreo de los datos. Integrando la similitud local de los datos a diferentes escalas, *diffusion maps* da una descripción global de los datos. Es robusto ante perturbaciones ruidosas y computacionalmente barato.

Los resultados de pasar a un espacio de 3 dimensiones (para poder visualizarlo) se muestran en la figura ??.



Como se puede apreciar en la figura, al igual que en los casos anteriores los patrones de las diferentes clases están realmente muy entremezclados, no

pudiéndose diferenciar ningún cluster.

Contents

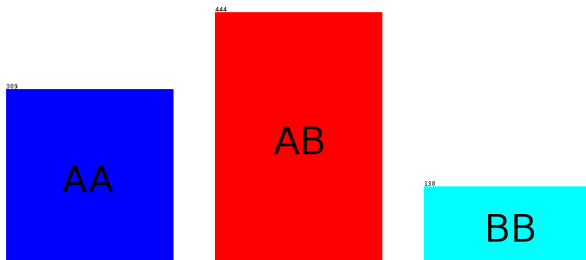
- 1 Base de datos
- 2 Primera etapa
- 3 Segunda etapa - Clases balanceadas
- 4 Tercera etapa - extracción de características
 - PCA
 - LDA
 - Diffusion Maps
- 5 Conclusiones

Cuestionarse si un determinado gen puede tener efectos en variables fenotípicas resulta una pregunta ambiciosa incluso para un genetista, y por esa razón es que desde un principio tenemos presente la dificultad del problema al que nos enfrentamos. A diferencia de otro tipo de estudios donde uno trata de elegir descriptores que sospecha que pueden tener una relación con la clasificación buscada, en nuestro caso buscábamos encontrar la relación entre un tipo de gen y sus marcadores moleculares respecto de un conjunto de variables fenotípicas, sin ningún tipo de conocimiento a priori. En este estudio nuestra parte consistía únicamente en realizar el análisis sobre los valores proporcionados. No tuvimos ninguna participación, como resulta obvio, en el diseño del experimento y la adquisición de datos. Es por eso que llegar a comprender el problema, las variables, los cuestionamientos e inclusive los objetivos, fue una tarea a la cuál se le debió dedicar un tiempo para nada despreciable dentro de los tiempos acotados del proyecto.

El problema fue atacado en su mayoría por las herramientas dadas en el curso, utilizando mayoritariamente el software weka y matlab. Se experimentó con una técnica un tanto más sofisticada como lo es "Diffusion Maps". A su vez, dado que este trabajo excede los objetivos de este curso ya que forma parte del proyecto de investigación en el que trabajamos, se intentará utilizar otro método para la clasificación: **Restricted maximum likelihood (REML)**, mediante la implementación en el software *Wombat* [?]. Es un método prometedor que es muy utilizado en problemas de este tipo.

Los resultados fueron contundentes, en la primera etapa, en la cual se trabajó con la base ya depurada pero sin ningún tipo de ajuste, los resultados no fueron

los esperados. Todos los clasificadores repondieron en la misma manera cometiendo el mismo error, clasificando a practicamente la totalidad de los patrones en una misma categoría: "AB". Observando la cantidad de patrones por clase, como se muestra en la figura ??, destaca la categoría "AB" superando ampliamente a las restantes. Es probable que eso lleve a que los clasificadores obtengan el mejor resultado (aproximadamente 50%) clasificando todos los patrones como "AB".



El siguiente paso fue trabajar o "masajear" la base de datos con el fin de balancear las clases y normalizar los descriptores. Esto tuvo dos efectos directos, en primer lugar los tres algoritmos de clasificación disminuyeron su performance clasificando en promedio un 37% de los patrones correctamente. Si bien el resultado es inferior al obtenido previamente, en esta ocasión el algoritmo trabajó y desarrollo sus clasificadores como se ve en los resultados de la matriz de confusión.

Una vez con la base trabajada y los clasificadores trabajando como es esperable se optó por aplicar los algoritmos básicos de extracción de características, PCA y LDA. En el caso de PCA se optó por trabajar en el subespacio de los primeros cuatro componentes cuya varianza acumulada supera al 80%. Este cambio se vio reflejado en un aumento promedio del 5% en la clasificación, pero más allá de la leve mejora esto nos lleva a pensar que no existe ningún tipo de distribución asociada a las clases. En la figura ?? se puede ver como los patrones presentan una distribución casi aleatoria respecto a las clases. Al aplicar LDA se redujo el espacio a una dimensión, la cual no presenta un poder de discriminación suficientemente superior a la observada anteriormente. Los resultados obtenidos son ligeramente superiores a los obtenidos mediante PCA salvo para el clasificador k-NN.

Finalmente los resultados demostraron que no es posible expresar, mediante los algoritmos aplicados, una correlación entre el gen estudiado y las variables fenotípicas. Naturalmente no podemos afirmar lo contrario, queda todavía un largo conjunto de herramientas estadísticas por aplicar, algunas de ellas con un enfoque mayor a las dadas en el curso hacia problemas biológicos.

También se comprobó como, que en general, cuando uno no puede visualizar la distribución de los datos o incluso las correlaciones entre las variables medidas de modo visual, previo a cualquier tipo de análisis. Es muy difícil que un algoritmo tenga un buen desempeño, los algoritmos vistos en el curso buscan automatizar ese tipo de análisis que uno puede realizar y hasta sospechar simplemente analizando los datos.