Facultad de Ingeniería de la
Universidad de la República

Introducción al Reconocimiento de Patrones

Curso 2013

Proyecto final

Autores:
Josué Luiz Ñuñez
Matías Tailanián

Tutores:
Federico Lecumberry
Ignacio Ramirez

11 de diciembre de 2013

Índice general

1. Introducción **2**

2. Base de datos **3**
2.1. Características fenotípicas 3
2.2. Características genotípicas - Determinación de polimorfismos (SNPs) 4
2.3. Resumen 4

3. Técnicas utilizadas **5**
3.1. Primera etapa 6
3.1.1. Naive Bayes 6
3.1.2. C 4.5 6
3.1.3. K-nn 6
3.2. Segunda etapa - masajeo de datos 6
3.2.1. Naive Bayes 6
3.2.2. C 4.5 6
3.2.3. K-nn 6
3.3. Tercera etapa - extración de características 6
3.3.1. PCA 6
3.3.2. LDA 9
3.4. NN? 10
3.5. SVM? 10

4. Conclusiones **11**

Introducción

El objetivo principal del proyecto es la investigación en técnicas que permitan contribuir con la predicción de fertilidad de rodeo lechero y la calidad de la carne integrando técnicas de reconocimiento de patrones sobre datos de alta dimensión.

A lo largo de los años se ha intentado relacionar ciertas características fenotípicas del ganado bovino con algunos indicadores genéticos. Es un problema actualmente abierto de gran interés mundial y Uruguay, siendo un país esencialmente ganadero, no puede estar ajeno. Desde hace varios años se viene desarrollando una línea de investigación relacionada con la genética molecular de la calidad de la carne bovina en el área de Genética de la Facultad de Veterinaria. Se ha realizado un abordaje desde diferentes puntos de vista, como la caracterización de genes conocidos en rodeos vacunos de nuestro país, o la búsqueda y análisis de nuevos genes asociados a la calidad de carne tanto bovina como ovina [1].

Obtener una correlación entre la información genética y la capacidad reprodutiva de bovinos, el impacto en la calidad de la leche y de la carne de los animales de nuestros rodeos tendría un alto impacto en la producción. A lo largo de los últimos años en Facultad de Veterinaria se han creado bases de datos con información fenotípica y genotípica relacionada conla calidad cárnica y la fertilidad bovina aplicada a la producción lechera. En esta oportunidad se trabajará con una base de datos confeccionada en nuestro país durante el año 2009, provista por la veterinaria Dra. Ana Meikle.

Base de datos

Para lo que sigue de este trabajo es importante conocer la base de datos con la que se trabajará, y realizar una breve descripción de las características relevantes.

La base de datos consta de información de un seguimiento realizado durante 9 meses sobre 891 vacas de 7 tambos diferentes y consiste en las características detalladas a continuación.

2.1. Características fenotípicas

- Edad.
- Condición corporal (BCS) desde 30 días previos al parto hasta 120 días posparto, con una frecuencia de al menos una vez cada 30 días. Se utilizó la escala de 5 puntos (1=flaca, 5= gorda).
- Cantidad de partos (Cantidad de lactancias). Es un indicador importante que tiene que ver con la historia y desgaste de la vaca.
- Anestro: es la cantidad de días que pasaron desde el parto hasta el reinicio. El reinicio se define como el día en que la progesterona aumenta a un nivel determinado, indicando que la vaca volvió a ciclar.
- Intervalo entre partos. Es la cantidad de días que pasan entre 2 partos consecutivos (válido solamente para vacas multíparas). Para maximizar la producción de leche lo que se busca es que la vaca quede peñada una vez al año, es decir un intervalo entre partos de 365 aproximadamente.
- Secado: la cantidad de días que pasan entre el último día que se le saca leche a la vaca y el parto. Es válido solo para vacas multíparas. Es un indicador de cuanto tiempo se dejó descansar a la vaca antes del parto. Cuanto más tiempo tiene, se prepara el cuerpo y llega en mejor forma.
- Servicios: Cantidad de inceminaciones realizadas para lograr la preñez.
- Concentración de progesterona hasta los 60 días posparto.

- Promedio de cantidad de grasa en la leche durante los primeros 100 días posparto.
- Promedio de cantidad de leche durante los primeros 100 días posparto.

2.2. Características genotípicas - Determinación de polimorfismos (SNPs)

Se extrajo ADN y se obtuvieron muestras de buena calidad. En la figura 2.1 se muestra el procedimiento realizado para caracterizar el gen IGF-I bovino en 3 clases: "AA", "AB" y "BB".

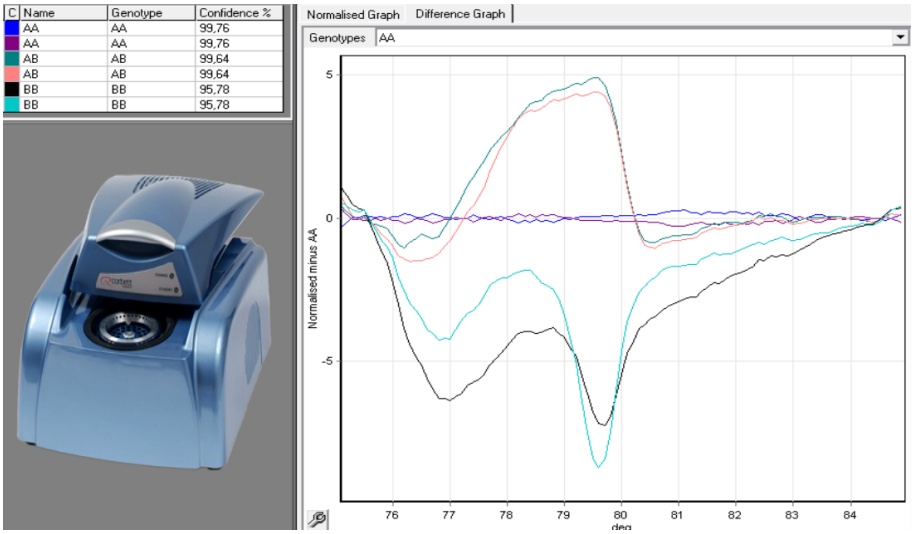


Figura 2.1: Determinación del genotipo

El eje de las ordenadas muestra la fluorescencia normalizada al genotipo "AA". Se puede observar que el genotipo "AB" se encuentra por encima del genotipo normalizado, mientras que el genotipo "BB" por debajo.

2.3. Resumen

En primer lugar fue necesario entender e interiorizarse con la base de datos, los conceptos y la terminología específica del problema. La lectura e interpretación de los datos no resulta para nada sencilla, por lo que se realizaron dos reuniones con la Dr. Ana Meikle, encargada de la investigación que llevó a cabo la confección de la base asociada a la producción de leche y fertilidad del ganado bovino. En ambas reuniones se buscó depurar la base con el fin de quitar información repetida y no relevante para el estudio. También se estableció un orden de relevancia en las características fenotípicas.

En resumen contamos con una base de datos acotada y "limpia" con varias características fenotípicas que se quieren correlacionar con los genotipos de cada individuo. En particular se abordará el problema como un trabajo de clasificación, tomando los genotipos como clases.

Técnicas utilizadas

La mayoría de los algoritmos de *machine learning* seleccionan los atributos apropiados para realizar sus decisiones. Por ejemplo los métodos de árbol de desición eligen en cada paso la característica que mejor separa en clases. Cuantas más características tengamos, en teoría tendremos más poder de discriminación, pero en la práctica no ocurre así. Agregar información irrelevante o características distractivas, confunde a los algorimos de machine learning. Está probado que el número de instancias necesarias para entrenamiento para producir un cierto nivel de performance crece exponencialmente con la cantidad de atributos irrelevantes.

Dado el efecto negativo de las características irrelevantes es muy común en muchos algoritmos de *machine learning* que se realice una etapa de selección de características donde se eliminen las irrelevantes. La mejor manera de realizar una selección de características es manual, adquiriendo un amplio conocimiento del problema en cuestión y logrando interpretar cada una de las características. Sin embargo es posible realizar una selección con algoritmos automáticos que resultan muy útiles. Reducir la dimensionalidad de los datos puede mejorar la performance de los algoritmos, además de reducir la complejidad del problema y bajar los requerimientos de capacidad de cómputo. Por esta razón se realizará una etapa de selección automática de características previa a la clasificación, como se explicará más adelante.

En una primera etapa de análisis sobre la base de datos se ataca el problema con los clasificadores **C4.5**, **Naive Bayes** y **K-NN**.

El algoritmo **C4.5** se utiliza para generar un árbol de desición que puede ser utilizado para clasificación. Por la naturaleza de los árboles, es un algoritmo no paramétrico, por lo que resulta robusto ante *outliers*.

Naive Bayes es un clasificador basado en la aplicación del teorema de Bayes. No particiona el espacio de instancias e ignora de forma robusta a las características irrelevantes. Asume por diseño que todas las características son independientes entre si, y paga un precio muy alto cuando hay características redundantes.

El algoritmo **k-NN** (k-nearest neighbors) es un algoritmo de clasificación super-

vizado no paramétrico que pedice la clase de las instancias basado en los k vecinos más cercanos. Es un algoritmo muy sensible a la propia estructura de los datos. Cuando $k \to \infty$, el algoritmo asegura una tasa de error no superior al doble de la tasa de error de Bayes (mínimo alcanzable dada la distribución de los datos).

Con esta batería de clasificadores se cubre un amplio espectro y se utilizan algunos de los algoritmos más utilizados para problemas de reconocimiento de patrones.

Para analizar los resultados se utilizan las implementaciones provistan en el software **Weka** [2].

3.1. Primera etapa

Se utiliza el clasificador compuesto AttributeSelectedClassifier que aplica una técnica de selección de características antes de entrenar al clasificador. Se logra entonces una reducción de la dimensionalidad. La estrategia elegida para la selección de características se realiza utilizando el enfoque wrapper , que evalúa el set de atributos utilizando un esquema de aprendizaje y utiliza validación cruzada para estimar la precisión del esquema de aprendizaje. El clasificador utilizado para estimar esta precisión es un árbol de decisión **C 4.5**.

En todos los casos se utilizó validación cruzada con 10 subconjuntos.

Una vez hecha la extracción de características, y así la reducción de dimensionalidad, se estudia el desempeño de diferentes clasificadores.

3.1.1. Naive Bayes

3.1.2. C 4.5

3.1.3. K-nn

3.2. Segunda etapa - masajeo de datos

3.2.1. Naive Bayes

3.2.2. C 4.5

3.2.3. K-nn

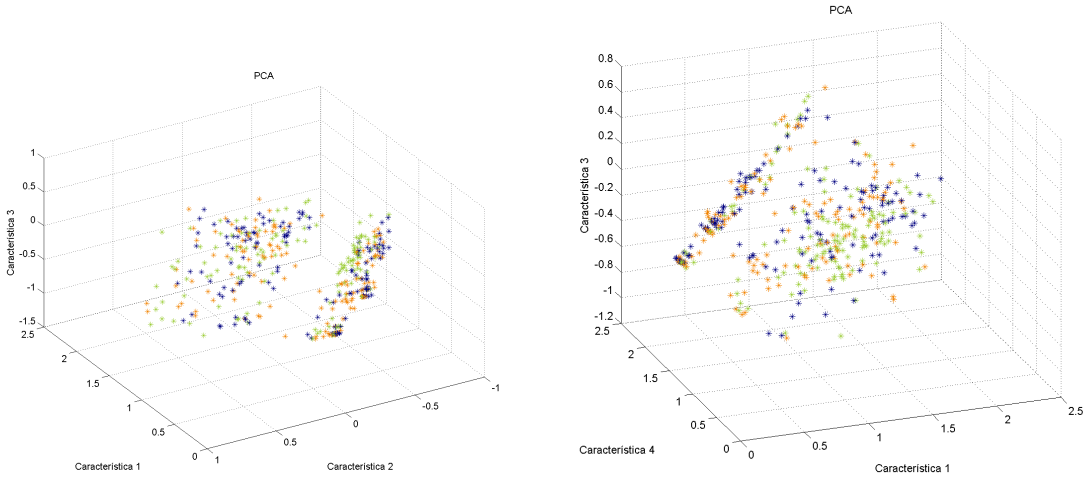
3.3. Tercera etapa - extración de características

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.

3.3.1. PCA

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.



(a) Características 1-2-3 (b) Características 1-4-3

Figura 3.1: Datos procesados con el algorítmo PCA

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 3.1) resulta prácticamente randómica y es imposible identificar visualmente algún cluster por clases. (Esto es medio trucho, pero sincero...)

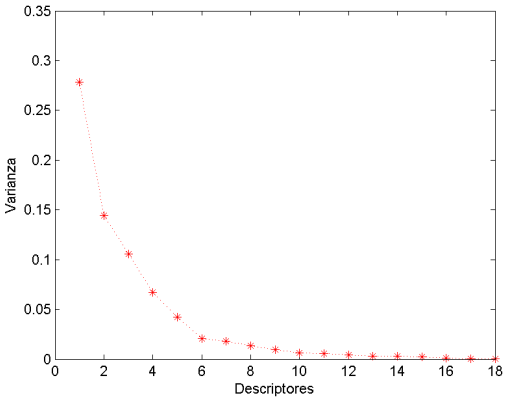


Figura 3.2: Varianza vs Componentes

En la figura 3.2 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.

Resultados de aplicar los clasificadores a los datos procesados con PCA: (Los escribo asi nomás para que queden registrados ya que no tengo las tablitas esas que a vos tanto te gustan...)

==========

Naive Bayes:

Time taken to build model: 0.02 seconds

=== Stratified cross-validation === === Summary ===

Correctly Classified Instances 163 39.372 % Incorrectly Classified Instances 251 60.628 % Kappa statistic 0.0906 Mean absolute error 0.4084 Root mean squared error 0.5064 Relative absolute error 91.8883 % Root relative squared error 107.4298 % Coverage of cases (0.95 level) 92.9952 % Mean rel. region size (0.95 level) 86.2319 % Total Number of Instances 414

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class 0,138 0,145 0,322 0,138 0,193 -0,010 0,540 0,367 1 0,616 0,558 0,356 0,616 0,451 0,055 0,563 0,365 2 0,428 0,207 0,509 0,428 0,465 0,232 0,672 0,526 3 Weighted Avg. 0,394 0,303 0,395 0,394 0,369 0,093 0,592 0,419

=== Confusion Matrix ===

a b c ¡– classified as 19 93 26 — a = 1 22 85 31 — b = 2 18 61 59 — c = 3

==========

C 4.5:

Number of Leaves : 35

Size of the tree : 69

Time taken to build model: 0.03 seconds

=== Stratified cross-validation === === Summary ===

Correctly Classified Instances 177 42.7536 Incorrectly Classified Instances 237 57.2464 Kappa statistic 0.1413 Mean absolute error 0.3913 Root mean squared error 0.5372 Relative absolute error 88.0331 Root relative squared error 113.9586 Coverage of cases (0.95 level) 78.5024 Mean rel. region size (0.95 level) 73.9936 Total Number of Instances 414

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class 0,362 0,297 0,379 0,362 0,370 0,066 0,553 0,396 1 0,428 0,264 0,447 0,428 0,437 0,165 0,601 0,425 2 0,493 0,297 0,453 0,493 0,472 0,192 0,614 0,416 3 Weighted Avg. 0,428 0,286 0,426 0,428 0,427 0,141 0,589 0,413

=== Confusion Matrix ===

a b c ¡– classified as 50 42 46 — a = 1 43 59 36 — b = 2 39 31 68 — c = 3

==========

k-NN:

IB1 instance-based classifier using 1 inverse-distance-weighted nearest neighbour(s) for classification

Time taken to build model: 0 seconds

=== Stratified cross-validation === === Summary ===

Correctly Classified Instances 179 43.2367 Incorrectly Classified Instances 235 56.7633 Kappa statistic 0.1486 Mean absolute error 0.387 Root mean squared error 0.6013 Relative absolute error 87.0681 Root relative squared error 127.5413 Coverage of cases (0.95 level) 48.0676 Mean rel. region size (0.95 level) 39.694 Total Number of Instances 414

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0,420 0,319 0,397 0,420 0,408 0,100 0,553 0,399 1
0,428 0,279 0,434 0,428 0,431 0,149 0,578 0,444 2
0,449 0,254 0,470 0,449 0,459 0,198 0,619 0,419 3
Weighted Avg. 0,432 0,284 0,434 0,432 0,433 0,149 0,584 0,421

=== Confusion Matrix ===

a b c ¡– classified as
58 45 35 — a = 1
44 59 35 — b = 2
44 32 62 — c = 3

==========

3.3.2. LDA

El algorítmo 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 disciminabilidad 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.

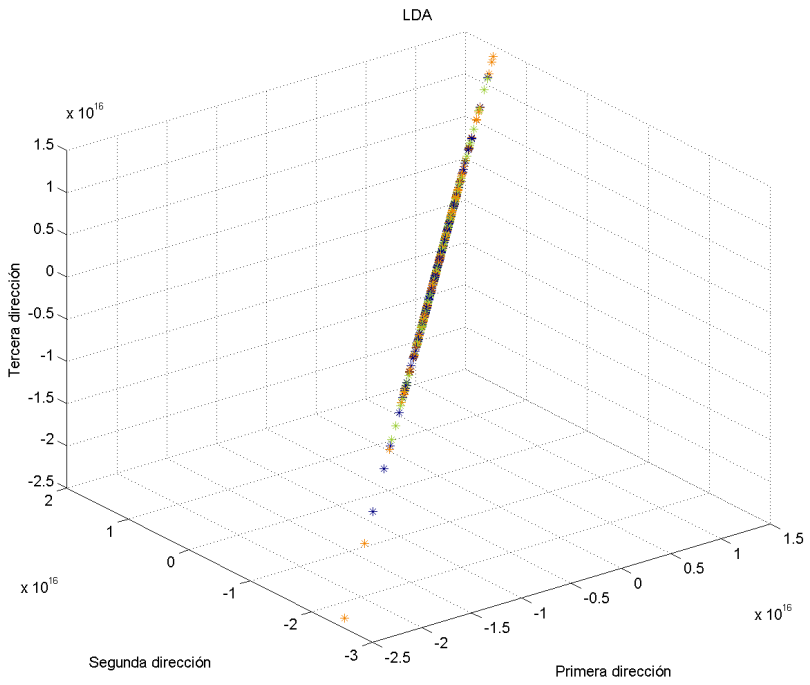


Figura 3.3: Proyeccion de los datos aplicando LDA

Nuevamente resulta imposible obtener algun resultado visualizando la distribución de los datos, de todas formas probaremos nuestros clasificadores sobre la base proyectada.

Resultados de aplicar los clasificadores a los datos procesados con LDA: (Los escribo así nomás para que queden registrados ya que no tengo las tablitas esas que a vos tanto te gustan...)

===

Naive Bayes:

Correctly Classified Instances 141 34.058 Incorrectly Classified Instances 273 65.942 Kappa statistic 0.0109 Mean absolute error 0.4438 Root mean squared error 0.4728 Relative absolute error 99.8449 Root relative squared error 100.2839 Coverage of cases (0.95 level) 100 Mean rel. region size (0.95 level) 100 Total Number of Instances 414

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class 0,246 0,283 0,304 0,246 0,272 -0,038 0,490 0,324 1 0,601 0,536 0,359 0,601 0,450 0,062 0,531 0,349 2 0,174 0,170 0,338 0,174 0,230 0,005 0,490 0,336 3 Weighted Avg. 0,341 0,330 0,334 0,341 0,317 0,009 0,504 0,336

=== Confusion Matrix ===

a b c ¡– classified as 34 75 29 — a = 1 37 83 18 — b = 2 41 73 24 — c = 3

===

3.4. NN?

3.5. SVM?

Conclusiones

Bibliografía

[1] Eileen Armstrong Reborati, *Detección y análisis de genes asociados a la calidad de la carne en bovinos*, Tesis de doctorado, Madrid, 2011.

[2] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11.

[3] Meyer, K. (2007). WOMBAT – A tool for mixed model analyses in quantitative genetics by REML, J. Zhejiang Uni.