Selección genética para la mejora de la raza ovina manchega mediante técnicas de Minería de Datos

M. Julia Flores, José A. Gámez, Juan L. Mateo y José M. Puerta

Departamento de Informática & SIMD - $i^3\mathcal{A}$
Universidad de Castilla-La Mancha
Campus Universitario s/n. Albacete. 02071
{julia,jgamez,juanlmc,jpuerta}@info-ab.uclm.es

Resumen

La estimación del mérito o valor genético (*Breeding value* en inglés) juega un papel muy importante en el Esquema de Selección de la Raza Ovina Manchega (ESROM), el cual comenzó hace quince años con el objetivo de mejorar las cifras en cuanto a la producción de la oveja manchega. En el esquema ESROM el valor genético se estima cada semestre empleando el modelo de evaluación genética BLUP. En este trabajo estudiamos el uso de técnicas de minería de datos para obtener una clasificación de dicho valor genético (modelo predictivo). Igualmente, se intentará aplicar técnicas de aprendizaje de redes Bayesianas que capten el modelo y las relaciones entre los distintos factores influyentes en la determinación del mérito genérico (modelo descriptivo).

Nuestro objetivo no es por supuesto el de sustituir el uso de BLUP para el esquema de selección ESROM, sino al contrario, pretendemos aprender tanto con técnicas supervisadas como no supervisadas de aquellos resultados obtenidos mediante la metodología BLUP. Además, a partir de los modelos aprendidos podremos extraer información preliminar sobre el valor genético de un animal, que puede resultar de gran utilidad. Se emplearán técnicas de clasificación con el fin de identificar buenos (subconjuntos de) predictores.

1. Introducción

La oveja manchega es la raza autóctona de la región de Castilla-La Mancha, y los dos principales productos que de ella se obtienen (queso y cordero manchegos) representan más del 50 % de la producción animal de dicha región. Por la implicación económica que esto supone y con el propósito de mejorar la producción de oveja manchega surgió un esquema de selección denominado ESROM. Este proceso de mejora se basa en la selección de los reproductores para obtener la siguiente generación en función del mérito genético de todos los animales de la población. Las autoridades regionales lo iniciaron hace quince años, con la finalidad de mejorar la producción de leche del ganado ovino. Uno de los principales puntos incluidos en el esquema ESROM es el valor genético [1], y su uso en la sustitución de animales dentro del rebaño. El esquema ESROM estima el BV empleando el *modelo animal* BLUP, el cual es un complejo método basado en relacionar los distintos rasgos/características mediante ecuaciones y resolviéndolas simultáneamente teniendo en cuenta toda la información disponible. El BV estimado nos permite ordenar a los animales genealógicamente y tomar decisiones como qué animales mejorarán la tendencia genética del rebaño, qué animales interesa (o cuáles no) incorporar al catálogo o mercado, qué hembras son las mejores candidatas para ser tratadas mediante inseminación artificial, etc, ... ESROM también anima a los ganaderos a seleccionar sus sustituciones en el rebaño basándose en el valor genético del animal.

El mérito genético de un animal es un

[1] Lo abreviaremos como *BV* del término anglosajón.

Actas del III Taller Nacional de Minería de Datos y Aprendizaje, TAMIDA2005, pp.223-232
ISBN: 84-9732-449-8

valor numérico que en el ESROM representa la desviación del animal con respecto al BV medio de las ovejas (hembras) manchegas nacidas en 1990 (el año de referencia). La estimación de este valor genético[2] mediante BLUP es un proceso complejo que ESROM realiza cada seis meses en un centro especializado. Además, el valor genético de un animal es dinámico, puesto que puede cambiar de una medición a la siguiente debido a cambios en los datos de producción del propio animal, a cambios en los datos de sus parientes, a datos relativos a su rebaño, etc, ...

A pesar de que el BV sea un dato numérico, muchas decisiones se toman en el percentil en el que el animal se sitúa con respecto a su propio rebaño o al del rebaño entero. En función de esta decisión, se emplean distintos umbrales: por encima del 50 %, por encima del 70 %, etc. Por este motivo, el objetivo de este trabajo es el de trabajar en la clasificación del BV dentro del esquema ESROM utilizando técnicas de aprendizaje automático [7] y minería de datos [4], cuya aplicación a la agricultura ha ganado gran interés en los últimos años [8, 3]. Obviamente, no perseguimos reemplazar el uso de la metodología BLUP, sino estudiar las posibilidades de predecir el valor genético de un animal mediante una aproximación guiada por los datos. Este otro enfoque empleará (con mucha diferencia) menos información que BLUP, y es más sencillo, pudiéndose utilizar en cuanto la información de un determinado animal esté lista, sin necesidad de esperar a conocer la evaluación genética semestral completa.

Para alcanzar este objetivo el artículo se ha estructurado en cuatro apartados, además de esta introducción. En el apartado 2 describimos los conjuntos de datos empleados en este trabajo. El apartado 3 explica el proceso de clasificación que se ha llevado a cabo. El apartado 4 está dedicado al uso de redes Bayesianas como método de modelado del valor genético. Finalmente, el apartado 5 presenta las principales conclusiones y futuras mejoras.

2. Procesamiento de los datos

Las bases de datos que hemos manejado en este trabajo nos han sido facilitadas por expertos de AGRAMA[3]. Los datos proporcionados contienen información adquirida entre 1989 y 2003, que dan lugar a un total de 9894 registros y 25 variables las cuales se pueden distribuir en 4 familias de datos distintas como muestra la tabla 1.

En nuestro conjunto de datos, hemos añadido una variable más: PedigreeIndex que se trata en realidad de una variable construida o derivada a partir de otros datos[4].

Nuestro interés radica en conseguir una estimación del BV de las ovejas en las primeras etapas. Todos los registros en la base de datos se refieren a primíparas, puesto que únicamente tras el primer parto y tras las lactaciones correspondientes es cuando se evalúa a las ovejas por vez primera, y por tanto, sería muy útil tener, tan pronto como sea posible, una aproximación de su mérito genético con el fin de poder tomar decisiones sobre ellas cuanto antes.

A partir de esta base de datos original se ha elaborado una nueva mediante la discretización de la variable BV para transformar lo que sería un problema de regresión en uno de clasificación. Se han seguido las recomendaciones de los expertos de AGRAMA, los cuales conocen el problema a fondo, a la hora de determinar la discretización. Así, sobre la variable BV se ha efectuado una discretización en 4 *bins* (grupos) de igual frecuencia dando lugar a la variable BV4= $\{f1, f2, f3, f4\}$. Denominaremos 4labels a la base de datos que proviene de sustituir la variable BV por su discretización BV4.

De esta base de datos inicial, se derivan otras dos bases de datos de naturaleza diferente, y que serán usadas para fines también

[2] Estrictamente hablando, deberíamos llamarlo *valor genético estimado* o EBV (en inglés), sin embargo, por simplicidad en la notación, seguiremos empleando BV, ya que es evidente que estamos trabajando con estimaciones.

[3] Asociación Nacional de Criadores de Ganado Ovino Selecto de Raza Manchega.

[4] Se trata simplemente de la media de los valores genéticos de sus progenitores: $\frac{\mathsf{BVMother}+\mathsf{BVFather}}{2}$

distintos: predictivo (apartado 3) y descriptivo (apartado 4).

En el primer caso lo que intentamos valorar es cómo de acertada es nuestra selección, cómo de bueno es el valor genético, incluso antes de que podamos calcular el correspondiente valor mediante BLUP. De este modo cuando una oveja nace, se le asignará el índice de pedigree como valor genético, ya que al no haber tenido su primer parto, no existe control lechero y, por tanto, su valor genético no se puede estimar todavía (por lo que BLUP no es aplicable). Las bases de datos en las que no se encuentren estas variables de lactación se las llamará predecir. Sin embargo, también se analizará la clasificación suponiendo que estos datos están disponibles y a estas bases de datos se las denominará blup. Nuestra intención será la de predecir el BV usando estas variables, comprobando después cómo de acertada resulta esta predicción. En esta tarea de clasificación en ningún caso se tendrá en cuenta la variable BVReliability, que se corresponde con la fiabilidad de la medición del BV. Para distinguir estos caso tendrán el sufijo nr. La razón para ello es que nosotros justamente lo que pretendemos es clasificar tratando de predecir el valor de la variable objetivo (BV). Por ello, no parece razonable emplear información sobre la fiabilidad de la medición de un valor que en realidad estamos tratando de predecir.

Es importante indicar que en este caso el objetivo no es analizar la aproximación al índice BLUP, sino buscar un error menor que si empleáramos únicamente la variable relativa al pedigree.

La segunda base de datos nos servirá para encontrar un modelo que extraiga las relaciones entre las distintas variables del problema, por ello incluye todos los datos disponibles, exceptuando la variable (calculada) PedigreeIndex. Precisamente en este caso nuestra intención es la de extraer las dependencias existentes entre las distintas variables y, cuando las haya, medir igualmente la relevancia de las mismas. De este modo, si incluyéramos, el índice de pedigree (PedigreeIndex) estaríamos introduciendo implícitamente una relación entre los valores genéticos de sus progenitores. Perderíamos así funcionalidad en nuestro objetivo de conseguir un modelo descriptivo, ya que de existir esta relación entre las variables BVFather y BVMother lo que deseamos es que nuestro modelo aprendido sea capaz de captar dicha relación así como de cuantificarla. Para distinguir las bases de datos en las que que no se encuentre la variable PedigreeIndex se les añadirá el sufijo np, mientras que en las que si esté pondremos p.

En este caso, trataremos de aplicar como técnica de minería de datos las redes Bayesianas con el fin de aproximar la valoración hecha por BLUP, pero proporcionando una herramienta más descriptiva e intuitiva para el usuario. De nuevo, esto dará lugar a un estudio de la precisión de esta valoración. [5] presenta un trabajo previo usando para el mismo fin Naive Bayes y c4.5.

Del estudio de las variables dentro del conjunto de datos vemos que las variables del entorno (salvo TypeOfBirth) son nominales con un gran número de posibles valores (de 64 a 127). Este tipo de variables puede introducir una cantidad considerable de ruido en el proceso de clasificación/aprendizaje, de modo que los hemos preprocesado mediante la implementación en ELVIRA[2] del método propuesto en [1]. Este método reduce el número de valores para una variable nominal a $|C| + 1$ etiquetas, siendo $|C|$ el número de clases (y añadiendo un grupo extra *desconocido*). Tras este proceso el número de valores de las variables va de 2 a 34 presentando en 4labels-nr una media de 7.48, mientras que en 4labels-np la media es de 7.04.

3. Minería de datos predictiva

En esta sección abordaremos el problema de la obtención del Valor Genético o *Breeding Value* (BV) desde el punto de vista de la clasificación. Más concretamente, la tarea consiste en buscar un buen conjunto de variables que permitan precisar un valor del parámetro BV aceptable antes de que éste sea calculado de manera real, ya que este proceso es muy costoso.

Para realizar esta tarea se pretende analizar

Datos BV:	
BVFather, ReBVF, BVMother, ReBVM, BVMaternalGM, ReBVMGM, BVParentalGM, ReBVPGM, BVMaternalGF, ReBVMGF, BVParentalGF, ReBVPGF, BVReliability, **BV**	
Datos Lactación:	
AvLactNorm, AvLact120	
Datos Entorno:	Datos lactación madre:
TypeOfBirth	NLactM
StockFarm	AvLactNormM
FatherStockFarm	MaxLactNormM
MotherStockFarm	AvLact120M, MaxLact120M

Cuadro 1: Variables dentro de las conjunto de datos, siendo **BV** la variable objetivo.

la bondad de una variación del algoritmo de selección de variables Filter+Wrapper propuesto en [5], al que llamaremos FW, usado con varios clasificadores ampliamente conocidos y utilizados. Para ello se ha tenido en cuenta una selección manual de variables de las bases de datos usadas para cada uno de los clasificadores. Principalmente se tiene como caso base para la comparación la efectividad obtenida por un clasificador teniendo en cuenta tan solo la variable PedigreeIndex. Esta variable se obtiene por construcción sobre BVFather y BVMother, y como se mostraba en [5] constituye un predictor bastante fiable de la variable clase BV.

Los clasificadores citados son Naive-Bayes, KDB con 1,2 y 3 padres, TAN y una algoritmo llamado SemiNaiveBayes que hace su propia selección de variables, junto con construcción por producto cartesiano, basada en el algoritmo propuesto por Michael Pazzani [9] pero sólo con selección forward.

3.1. Algoritmo FW

Básicamente, el algoritmo FW en su definición inicial consiste en realizar un ranking con las variables mediante una métrica, que en este caso es *Symmetrical Uncertainty*. Esta métrica viene dada por la siguiente expresión:

$$SU(C,X) = \frac{2 \cdot (H(C) - H(C|X))}{H(C) + H(X)}$$

De esta manera se trata con la información mutua de cada variable (X) con respecto a la clase (C), pero con una proyección de los valores en el rango [0,1]. Con esto se consigue una normalización para poder comparar variables con distinto número de estados.

A continuación, el algoritmo utiliza este orden para ir añadiendo variables al clasificador, el cual es evaluado mediante validación cruzada y se detiene al no obtener un mejor resultado. Sin embargo, se puede presentar el caso en que haya dos variables contiguas en el orden establecido que están fuertemente correlacionadas y por tanto es muy posible que detenga el algoritmo a pesar de que pueda haber más variables con un orden inferior que aporten nueva información al clasificador. Para solucionar este problema se utiliza el parámetro *lookahead* que establece cuántas variables se pueden seguir valorando a partir de un punto en el que no se obtiene mejora con la variable actual.

A partir de este algoritmo, en este trabajo, se ha introducido una modificación según la cual, una vez se ha conseguido una lista de variables seleccionadas, ésta se recorre en orden inverso tratando de encontrar alguna que al eliminarla se mejore la clasificación. Esto está justificado ya que, a pesar de que el orden obtenido por la métrica SU es bastante bueno, es posible que haya variables con una valoración muy alta de la métrica que fueron introducidas en los primeros pasos pero que la información aportada para el clasificador sea mínima y haya otras variables que se añadieron posteriormente que engloben a aquélla. Esto quiere decir que una variable con una valoración para la métrica usada puede suponer que aporta una mínima información nueva, con respecto a las anteriores, sobre la variable clase pero es tan poca que se pueden presentar otras que aun con una valoración de la métrica inferior complementen mejor la información del resto de las variables. Por tanto, en este caso, la evaluación del clasificador mejoraría al eliminar la variable.

Aquí también se tiene en cuenta el parámetro *lookahead* de la misma manera, es decir, que aunque al probar con una variable no se mejora con su eliminación se puede avanzar en la búsqueda tantas variables como in-

dique el valor de este parámetro.

3.2. Desarrollo y análisis

Una vez planteados los objetivos y las herramientas de las que haremos uso pasamos a describir los pasos realizados comentando los resultados obtenidos.

Como ya se expuso, la efectividad del algoritmo FW se analizará tomando como referencia el clasificador formado tan sólo por la variable PedigreeIndex. Los resultados que se obtienen se muestran en la tabla 2. Evidentemente al ser un clasificador con una sola variable predictora todos los clasificadores tienen el mismo valor.

Por ahora sólo nos interesa la primera columna en la que se presenta la precisión del clasificador para las bases de datos originales en las que la variable clase cuenta con 4 estados. En principio, dado que las bases de datos sólo difieren en el número de variables, el resultado debería ser el mismo en ambos casos sin embargo la diferencia es muy escasa y se explica por el orden de las instancias en cada una de ellas.

Como primera aproximación a la aplicación del algoritmo FW se realizaron tres tandas distintas. La primera con el parámetro *lookahead* igual a cero y sin el procesamiento backward, la segunda con *lookahead* igual 5 y con procesamiento backward y la tercera también con procesamiento hacia atrás pero con *lookahead* igual a infinito. Los resultados para los seis clasificadores utilizados con estas configuraciones se muestran en la parte (a) de la tabla 3.

Como era de esperar al introducir el procesamiento backward y también aumentar el nivel de búsqueda tras fallo (*lookahead*) se obtiene una precisión mayor para los clasificadores, a pesar de que ésta no es muy significativa. También hay que notar que el requerimiento temporal al aumentar la precisión de la búsqueda no es demasiado elevado en relación con tiempo total de ejecución.

Para todos los clasificadores, salvo para Naive-Bayes, con cualquier base de datos se mejora el valor del clasificador de referencia. Incluso en el caso más básico, es decir, sin contar con la propia variable construida PedigreeIndex ni con la información de la lactación de la propia oveja, se mejora la clasificación en medio punto. En el caso más favorable, utilizando toda la información, la ganancia es de casi dos puntos y medio, también hay que notar que esta ganancia se consigue con el algoritmo SemiNaiveBayes que es el más costoso computacionalmente.

Tras analizar estos resultados se observó, según las matrices de confusión en cada caso, que la clasificación fallaba de manera muy notable en el tercer estado de aquellos en los que se había discretizado la variable clase. Esto propició la idea de utilizar una modificación de esta discretización. A partir de este punto el resto de ejecuciones del algoritmo FW se realizaron con *backprocessing* y con *lookahead* igual a infinito. En primer lugar se probó uniendo los estados centrales de forma que se distinguía entre un valor genético bajo, medio y muy alto. Esto significa que se categoriza la variable BV en los intervalos de percentiles de 0 a 50 % de 50 a 80 % y superior a 80 %.

Otra posibilidad fue hacer una clasificación binaria de forma que se unen los tres primeros estados y sólo se distingue entre un valor genético muy alto o no. Aquí la distinción está en valorar si se supera el percentil 80 % o no.

Los resultados del algoritmo FW para estas discretizaciones están en el apartado (b) y (c) de la tabla 3. En las columnas 2 y 3 de la tabla 2 se muestran los resultados base para estos dos casos. En ambos casos la mejora sobre la clasificación con la variable PedigreeIndex es similar al anterior, no se obtienen mejoras significativas y en cambio se pierde información al tener menos estados entre los que distinguir la variable clase.

Como último caso de prueba se intentó modificar la clasificación para sólo tener en cuenta los casos a los que se le asigna al menos un 70 % de probabilidad ignorando el resto y no teniéndolos en cuenta para obtener la precisión de la clasificación. Ahora se trabaja con las discretizaciones originales.

También con esta modificación se observaba en las matrices de confusión que en los estados centrales es donde más instancias se ig-

	4labels	3labels	2labels	conf. 0.7
predecir	72.4784	78.2088	91.3887	89.1419
blup	72.4178	78.1989	91.2774	89.0162

Cuadro 2: Resultados para un clasificador que sólo cuenta con la variable PedigreeIndex.

		NB	KDB-1	KDB-2	KDB-3	TAN	SNB
no bp	predecir-np	$70{,}3253_{2}$	$72{,}5489_{3}$	$72{,}6400_{3}$	$72{,}6400_{3}$	$\mathbf{72.7713}_{6}$	$72{,}3165^{2}_{1}$
	predecir-p	$72{,}8421_{4}$	$73{,}2160_{4}$	$73{,}3779_{4}$	$\mathbf{73.5091}_{6}$	$73{,}1655_{4}$	$73{,}1959^{2}_{1}$
	blup-np	$72{,}1853_{3}$	$75{,}1870_{6}$	$75{,}0861_{6}$	$74{,}6718_{7}$	$\mathbf{75.3185}_{7}$	$74{,}4390^{3}_{1}$
	blup-p	$74{,}8637_{5}$	$75{,}8441_{5}$	$76{,}1573_{6}$	$75{,}1265_{8}$	$75{,}8947_{4}$	$\mathbf{76.8651}^{3}_{1}$
bp-5	predecir-np	$70{,}3253_{2}$	$72{,}5489_{3}$	$72{,}6400_{3}$	$72{,}6400_{3}$	$\mathbf{72.8419}_{5}$	$72{,}3165^{2}_{1}$
	predecir-p	$72{,}8421_{4}$	$73{,}2160_{4}$	$73{,}3880_{3}$	$\mathbf{73.5091}_{6}$	$72{,}9836_{4}$	$73{,}3273^{3}_{1}$
	blup-np	$72{,}1853_{3}$	$75{,}1870_{6}$	$75{,}0861_{6}$	$74{,}6718_{7}$	$\mathbf{75.5207}_{7}$	$74{,}4997^{3}_{1}$
	blup-p	$74{,}8637_{5}$	$75{,}8441_{5}$	$76{,}1573_{6}$	$75{,}1265_{8}$	$75{,}9958_{4}$	$\mathbf{76.4304}^{3}_{1}$
bp-inf	predecir-np	$70{,}7094_{4}$	$72{,}7106_{4}$	$\mathbf{73.1049}_{5}$	$72{,}8118_{5}$	$72{,}9532_{8}$	$72{,}3569^{2}_{1}$
	predecir-p	$72{,}8421_{4}$	$73{,}3879_{3}$	$\mathbf{73.6710}_{5}$	$73{,}5193_{5}$	$73{,}0845_{5}$	$73{,}1959^{2}_{1}$
	blup-np	$72{,}3470_{5}$	$75{,}1870_{6}$	$75{,}0861_{6}$	$74{,}6920_{8}$	$\mathbf{75.4498}_{6}$	$74{,}6614^{3}_{1}$
	blup-p	$75{,}7433_{4}$	$75{,}8441_{5}$	$76{,}1573_{6}$	$76{,}2179_{6}$	$75{,}8846_{5}$	$\mathbf{76.8954}^{3}_{1}$

(a) Resultados con las bases de datos originales

3labels	predecir-np	$77{,}2991_{3}$	$78{,}2796_{6}$	$78{,}5625_{3}$	$\mathbf{78.6636}_{5}$	$78{,}3502_{3}$	$78{,}5727^{3}_{1}$
	predecir-p	$78{,}4918_{3}$	$78{,}8961_{3}$	$79{,}3812_{5}$	$\mathbf{79.5731}_{4}$	$78{,}8658_{3}$	$79{,}0276^{2}_{1}$
	blup-np	$78{,}8658_{3}$	$80{,}8672_{4}$	$81{,}1400_{4}$	$\mathbf{81.1907}_{4}$	$81{,}1703_{7}$	$80{,}7053^{3}_{1}$
	blup-p	$80{,}9074_{4}$	$81{,}4634_{6}$	$\mathbf{82.1002}_{6}$	$82{,}0495_{8}$	$81{,}4533_{6}$	$82{,}0091^{3}_{1}$

(b) Resultados con la variable clase en tres estados

2labels	predecir-np	$91{,}2270_{2}$	$91{,}4492_{6}$	$91{,}4997_{4}$	$\mathbf{91.6515}_{4}$	$91{,}4493_{5}$	$91{,}3281^{2}_{1}$
	predecir-p	$91{,}7626_{5}$	$\mathbf{92.0659}_{3}$	$91{,}9951_{3}$	$91{,}9951_{3}$	$92{,}0557_{5}$	$91{,}9546^{2}_{1}$
	blup-np	$92{,}2074_{4}$	$\mathbf{93.0667}_{7}$	$92{,}8845_{5}$	$92{,}7632_{7}$	$93{,}0565_{7}$	$92{,}7835^{4}_{2}$
	blup-p	$93{,}2686_{2}$	$93{,}2687_{3}$	$\mathbf{93.4302}_{3}$	$93{,}3898_{4}$	$93{,}3293_{3}$	$93{,}3697^{2}_{1}$

(c) Resultados con la variable clase en dos estados

0.7	predecir-np	$91{,}8270_{4}$	$90{,}3320_{3}$	$90{,}0072_{3}$	$90{,}0072_{3}$	$90{,}7400_{5}$	$\mathbf{92.9503}^{3}_{2}$
	predecir-p	$89{,}9109_{3}$	$89{,}6036_{4}$	$89{,}2037_{3}$	$89{,}2037_{3}$	$89{,}4803_{3}$	$\mathbf{90.8471}^{3}_{1}$
	blup-np	$95{,}2283_{3}$	$92{,}6744_{4}$	$90{,}6824_{3}$	$90{,}3085_{3}$	$92{,}4532_{5}$	$\mathbf{95.6181}^{4}_{3}$
	blup-p	$92{,}6239_{4}$	$90{,}3871_{4}$	$90{,}8715_{2}$	$90{,}8715_{2}$	$90{,}3001_{5}$	$\mathbf{95.4598}^{3}_{2}$

(d) Resultados forzando la clasificación con una confianza del 70 %

Cuadro 3: Resultados para el algoritmo FW para los distintos clasificadores: (a) con las bases de datos originales en las que la clase tiene 4 estados variando los parámetros del algoritmo *backprocessing* y *lookahead*, (b) con las bases de datos en las que la variable clase tiene 3 estados (3labels), (c) con las bases de datos en las que la variable clase tiene 2 estados (2labels) y (d) con las bases de datos originales pero contabilizando sólo aquellas instancias que se categorizan con al menos un 70 % de probabilidad.
En cada caso, el subíndice especifica la cardinalidad del conjunto de variables seleccionadas por el clasificador. Para el algoritmo SemiNaiveBayes el superíndice especifica el número real de variables usadas.

noran por no superar el porcentaje fijado de confianza al clasificar. Aquí se obtienen unos resultados de clasificación muy buenos ya que la precisión del clasificador pasa del 90 %, o por lo menos se queda muy cerca, en todos los casos. Sin embargo esta mejora cuantitativa se ve empañada por el elevado número de casos que se dejan sin clasificar. En porcentaje, no se podrían clasificar entre un 38 % y un 56 % de las instancias. Es decir, casi en la mitad de los casos no se podría afirmar nada sobre la oveja de estudio aunque en un 90 % de la otra mitad se acertaría.

Un aspecto muy llamativo de esta última tabla es que en aquellos conjuntos de datos en los que se ha introducido la variable PedigreeIndex todos los clasificadores producen peores resultados, cuando hasta ahora, y parece lo más lógico, esta variable mejora el poder de clasificación. La razón que explica este hecho se debe a que siempre esta variable obtiene el primer lugar en el ranking por la métrica SU y por tanto siempre se elige como parte de la selección. Pero esta elección condiciona el resto por el hecho de que algunas de las variables que se seleccionaron con las bases de datos sin PedigreeIndex no se elegirán en este caso al tener mucha correlación con esta variable o con otras que ya se han elegido y no mejoran el resultado del clasificador en la etapa Wrapper del algoritmo. En pocas palabras, es evidente que el algoritmo encuentra un óptimo local.

Desde un punto de vista general, después de presentar los resultados numéricos en cada caso, podemos analizar los conjuntos de variables seleccionados. En primer lugar hay que señalar que en el conjunto más grande de variables, la cardinalidad de ocho, lo que supone un tercio de las variables iniciales aunque en la mayoría de los casos es menor, entorno a 4 o 5. En este aspecto, el algoritmo SemiNaiveBayes obtiene un clasificador todavía más simple en el que hay tan solo de 1 a 3 variables finales, aunque esta reducción obedece a que este algoritmo hace su propia construcción de variables. Pero aún así, el número de variables reales tomadas no pasan de cuatro.

Dejando atrás el aspecto numérico, se observa de los conjuntos de variables seleccionados que las que aparecen con más frecuencia son BVFather, BVMother y PedigreeIndex en aquellos casos donde ésta se pueda elegir. En un principio se podía aventurar que estos tres atributos son los que más información proporcionan para estimar el valor genético de un animal. Sin embargo además de éstos aparecen a continuación, en orden de frecuencia, otros como StockFarm y AvLac120. Además, estas cinco variables, que en este apartado se destacan como las más informativas sobre el valor genético de una oveja, en el apartado 4 se verá que también se resaltan desde otro punto de vista más general.

4. Minería de datos descriptiva

En esta sección nos dedicaremos a analizar mediante redes Bayesianas las asociaciones entre las variables del dominio estudiado en este trabajo. Para ello hemos realizado aprendizaje automático de redes Bayesianas mediante el uso de técnicas de búsqueda local (ascensión de colinas) y utilización de una métrica Bayesiana, BDeu [6], para medir la adecuación de cada red candidata a los datos obtenidos.

Los datos de partida son los descritos en la sección segunda en el cuadro 1, teniendo en cuenta todas las variables y dos posibles discretizaciones de la variable BV en 4 y 3 particiones (bins). El aprendizaje se ha realizado mediante el paquete ELVIRA utilizando dos métodos de búsqueda local: uno sin ningún tipo de restricciones de partida sobre las relaciones entre las variables y por consiguiente se permiten todos los operadores locales habituales, inserción de un arco no existente, borrado de un arco existente e inversión de un arco, todo ello sin introducir ciclos dirigidos en las redes candidatas. El segundo método es imponiendo una serie de restricciones de existencia de relaciones (arcos) entre las variables correspondientes a los datos BV, básicamente, se imponen las relaciones de parentesco entre los animales tanto para las variables de confianza de valoración genética y la propia valoración genética. Las relaciones, impuestas como arcos orientados, serán de padres a hijos, de esta manera la variable BV tendrá como padres los

valores respectivos de BV de su padre y de su madre; y éstas a su vez, las de sus respectivos padres (abuelos del animal en cuestión). Este árbol ancestral también se impone con las variables ReBV de confianza de la valoración genética respectivas. Estos dos árboles genealógicos se pueden observar en la figura 2 mediante los arcos de trazo más grueso.

Una vez impuestas las relaciones anteriores, se inicia un proceso de búsqueda local igual que el descrito previamente, pero con la restricción, en el espacio de búsqueda, de que siempre se tienen que tener en cuenta en la medición de la adecuación de los datos a las redes candidatas, dicho de otra forma, toda red candidata tendrá siempre como subgrafo el que aparece en la figura 2 con arcos más gruesos.

En las figuras 1 y 2 se pueden observar las redes resultantes para el aprendizaje sin restricciones y con restricciones respectivamente. Estas dos redes corresponden a la discretización de la variable BV en 3 intervalos. Al ser un problema real, y ante la ausencia de tener una red Bayesiana que represente el dominio de forma exacta con la que podamos comparar los resultados obtenidos, hemos optado por medir la distancia de Kullback-Leiber (KL) entre las distribuciones conjuntas representadas por las dos redes Bayesianas aprendidas y la red Bayesiana vacía de enlaces, esto es, todas las variables son marginalmente independientes y con la distribución empírica representada por la base de datos original[5]. Los resultados de estas medidas se pueden observar el la tabla 4.

De las medidas anteriores se pueden obtener algunas conclusiones: en primer lugar hemos de observar que las medidas KL de las redes Bayesianas con y sin restricciones a los datos son prácticamente iguales, lo que quiere decir que ambas representan los datos de forma similar. Si ambas tienen una adecuación a los datos similar es preferible, entonces, el modelo en donde las relaciones se puedan interpretar por parte de los expertos de forma causal y, por tanto, en este caso preferiríamos las redes que imponen las restricciones de parentesco por su natural interpretación. Sin embargo, si preferimos el modelo más simple este puede ser el modelo sin restricciones por el hecho de tener cuatro arcos menos, de todas formas, esta diferencia no es significativa y la complejidad de las tablas de probabilidad en ambos modelos en similar, de hecho, los modelos con restricciones tienen una medida de KL con respecto a la red vacía de enlaces menor que los modelos sin ellas.

Redes	KL Datos	KL vacía
4-interv. Sin Rest.	13,67	14,16
4-interv. Con Rest.	13,66	13,61
3-interv. Sin Rest.	13,62	14,62
3-interv. Con Rest.	13,63	13,59

Cuadro 4: Medidas de Kullback-Leiber (KL)

En las redes aprendidas podemos destacar el conjunto de relaciones asociadas a la variable BV, variable objetivo para la predicción. Los nodos no coloreados en cada una de las redes de las figuras 1 y 2 corresponden al manto de Markov asociado a la variable BV. Este conjunto lo incluyen los padres, hijos y padres de éstos, conjunto que hace condicionalmente independiente BV del resto de variables. Este conjunto está muy relacionado con el proceso de selección de variables descrito en la sección de predicción de este trabajo, viéndose en cada caso, que los subconjuntos de variables que mejores resultados obtienen en la clasificación incluyen, cuando no todas, subconjuntos de variables de la unión de ambos mantos de Markov.

5. Conclusiones

En este trabajo hemos tratado de mostrar dos vías para ayudar predecir el valor genético de la oveja manchega como complemento al sistema BLUP. Estos dos puntos de vista han sido por un lado, mediante clasificación, encontrar un subconjunto de atributos relativos a cada oveja que permitan clasificarla de una forma sencilla y rápida. Por otro la-

[5] Es equivalente a medir la log-verosimilitud de los datos dados los modelos representados en las redes Bayesianas

Figura 1: Red Bayesiana sin imponer restricciones de parentesco

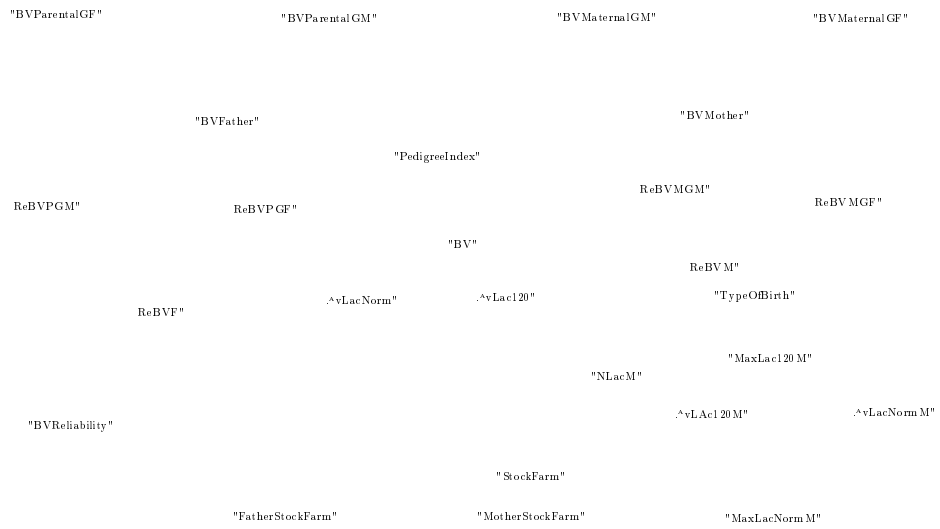

Figura 2: Red Bayesiana imponiendo restricciones de parentesco

do, mediante redes bayesianas se ha intentado mostrar las relaciones existentes entre las variables, de manera que, de una forma muy intuitiva, de manera visual cualquier persona pueda analizar y comprender el dominio del problema mucho mejor.

Pensamos que hemos logrado los objetivos marcados desde un principio, sobretodo porque los resultados de cada método se corroboran mutuamente. Con esto se ha podido establecer un subconjunto de variables que proporcionan una base para el trabajo de los ganaderos de clasificar las ovejas cuando no es posible esperar a los resultados del test BLUP ya que éste se realiza cada seis meses.

Como se ha mostrado en el apartado 3, la fiabilidad de la clasificación que se puede proporcionar se acerca al 75 % para el caso normal en el que todavía no hay datos de lactancia de la propia oveja, siendo un poco superior cuando se tienen estas mediciones. Este último escenario también se puede presentar cuando una vez conocidos los datos de lactación se pretende saber si el cambio de algún otro factor influirá en la clasificación de la oveja. En ambos casos la mejora con respecto al esquema anterior en el que sólo se contaba con la variable PedigreeIndex se ha dejado patente.

Como trabajos futuros se pretende seguir con el estudio del dominio con otros modelos de predicción tanto en tareas de clasificación como en tareas de regresión.

Referencias

[1] P. Berka and I. Bruha. Discretization and grouping: Preprocessing steps for data mining. In *Proc. of Principles of Data Mining and Knowledge Discovery PKDD'98*, LNAI 1510, pages 239–245. Springer Verlag, 1998.

[2] Elvira-Consortium. Elvira: An environment for creating and using probabilistic graphical models. *Proc. of 1st European Workshop on Probabilistic Graphical Models*, pages 222–230. 2002.

[3] I. Farkas (Ed.). Special issue: Artificial intelligence in agriculture. *Computers and Electronics in Agriculture*, 39(1-3), 2003.

[4] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17:37–54, 1996.

[5] J.A. Gámez. Mining the esrom: A study of breeding value prediction in manchego sheep by means of classification techniques plus attribute selection and construction. Technical Report DIAB-05-01-3, Universidad de Castilla-La Mancha, 2005.

[6] D. Heckerman, D. Geiger, and D.M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–244, 1995.

[7] Tom M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[8] H. Murase (Ed.). Special issue: Artificial intelligence in agriculture. *Computers and Electronics in Agriculture*, 29(1-2), 2000.

[9] M. J. Pazzani. Searching for dependencies in Bayesian classifiers. Departament of Information and Conputer Science University of California, 1997