Predicción del valor genético en ovejas de raza manchega usando técnicas de aprendizaje automático

Jose A. Gámez
Dept. de Informática
Universidad de Castilla-La Mancha
Campus Universitario s/n
02071, Albacete
jose.gamez@uclm.es

Antonio Salmerón
Dept. de Estadística y Matemática Aplicada
Universidad de Almería
La Cañada de San Urbano, s/n
04120, Almería
antonio.salmeron@ual.es

Resumen

En este trabajo se presenta un estudio sobre la predicción del valor genético en ovejas de raza manchega usando técnicas basadas en regresión y métodos probabilísticos. El objetivo del trabajo es doble: por un lado obtener buenas predicciones y por otro lado testear si los métodos probabilísticos basados en redes gausianas son competitivos con los métodos clásicos basados en regresión. Para ello hemos probado con tres técnicas de cada tipo y además se ha realizado una fase de selección de variables para identificar predictores simples. Como resultado de los experimentos podemos ver que los métodos basados en regresión obtienen buenos resultados pero los basados en redes gausianas (con las suposiciones aquí realizadas) no son competitivos en la mayoría de los casos.

1. Motivación

La *oveja manchega* es la raza ovina autóctona en Castilla-La Mancha y sus dos principales productos (queso y cordero manchego) representan más del 50 % de la producción final animal en la región. Esta importancia en la economía de la región junto con el objetivo de ganar competitividad frente a razas foráneas llevó a las autoridades de la región a implantar hace 15 años el *Esquema de Selección de la Raza Ovina Manchega* (ESROM). El principal objetivo del ESROM es la mejora de las cifras productivas de la raza ovina manchega, principalmente en cuanto a producción lechera, mediante la progresiva mejora genética de los animales incluidos en los rebaños de la región. Para lograr esta mejora el ESROM provee a los ganaderos con una serie de instrumentos: generación de rankings de animales atendiendo a su mérito genético, creación de bolsas de sementales, procesos de inseminación artificial, etc, ...

El factor clave en el ESROM, es por tanto, el mérito genético de los animales, o mejor dicho la estimación de dicho mérito genético (*breeding value o BV*). En el ESROM el mérito genético de un animal es estimado usando el *modelo animal* de la metodología BLUP (Best Linear Unbiased Prediction), que consiste en un modelo complejo basado en relacionar diferentes características mediante ecuaciones lineales y resolver el sistema considerando de forma simultánea toda la información disponible. En el sistema de ecuaciones planteado, el método BLUP intenta tener en cuenta aquellos factores que no son mérito del animal (alimentación, higiene, instalaciones, ...) para que los resultados sean comparables entre toda la cabaña. El valor BV estimado permite situar los animales en el ranking genealógico y tomar decisiones sobre qué animales constituyen una mejora genética para el rebaño, qué animales deben (o no) ser añadidos al catálogo de sementales, qué ovejas son candidatas para ser usadas como madres en los programas de in-

seminación artificial, etc, ... Las autoridades y asociaciones pertenecientes al ESROM *recomiendan* a los ganaderos hacer la reposición de animales en sus rebaños basándose en los rankings obtenidos a partir del valor genético estimado.

El valor genético (BV) de un animal es un valor numérico que en el caso del ESROM representa la desviación del mérito genético del animal con respecto al BV medio de las ovejas de raza manchega nacidas en 1990 (conocido como el año de referencia). La estimación del valor genético usando BLUP es un proceso complejo que en el ESROM se lleva a cabo cada seis meses en un centro especializado. Además, el valor genético de un animal es un valor dinámico puesto que puede cambiar de una medición a otra debido a cambios de producción del propio animal, en sus parientes, en el rebaño en general, etc ...

En este trabajo nos proponemos llevar a cabo un estudio de la estimación del valor genético usando técnicas de aprendizaje automático basadas en regresión y adicionalmente realizar un caso de estudio con este problema para estudiar predictores basados en redes Gaussianas. Distinguiremos dos tareas distintas:

- Estimación del valor genético en ovejas recién nacidas. El mérito genético de una oveja no es estimado usando el método BLUP hasta que ha tenido un primer parto y se ha controlado el período de lactación posterior al parto. Hasta ese momento se usa el índice de pedigree: $\frac{BV_{padre}+BV_{madre}}{2}$ como valor genético estimado. Nuestra primera tarea consiste en estudiar posibles mejoras a esta estimación mediante el uso de las técnicas antes mencionadas y/o el uso de otras variables.

- Estimación del valor genético a partir de BLUP. Obviamente no intentamos sustituir al método BLUP para realizar la estimación del valor genético, sino por el contrario aprender de forma supervisada de los valores obtenidos por BLUP, e inducir modelos que con mucha menos información (variables predictoras) de la usada por BLUP y sin necesidad de esperar a la evaluación global de la cabaña cada seis meses, puedan proporcionar estimaciones razonables, con el objetivo de poder ser usadas en la toma de decisiones tempranas.

En cuanto al estudio de los predictores basados en redes Gaussianas, el punto de partida será usar un método de tipo Naïve Bayes. En [4] se hace una experimentación con un Naïve Bayes (NBc) con clase continua de la que se deduce que al contrario del caso de la clasificación, el método NBc se ve muy afectado por la hipótesis de independencia realizada y no obtiene buenas predicciones. En este trabajo estudiaremos (sobre el caso de estudio propuesto) los resultados obtenidos por otros modelos basados en redes Gaussianas y compararemos con los obtenidos por NBc.

Para llevar a cabo nuestro objetivo dividimos el trabajo en 4 secciones además de esta introducción. En la sección 2 describimos los conjuntos de datos utilizados. En la sección 3 se describen las técnicas usadas para realizar la predicción, así como los experimentos diseñados con ellas. A continuación aplicamos selección de variables en la sección 4 para descubrir predictores de similar rendimiento a los ya obtenidos, pero más simples. Por último en la sección 5 presentamos nuestras conclusiones y vías de expansión de este trabajo.

2. Conjuntos de datos

Los conjuntos de datos usados en este trabajo han sido obtenidos a partir de las bases de datos de AGRAMA (*Asociación nacional de Ganaderos de ovejas de RAza MAnchega*), que contienen registros sobre animales entre los años 1989 y 2003. Después de un proceso de preparación de datos (descrito en [6]) y siguiendo las indicaciones de los expertos de AGRAMA hemos obtenido un conjunto de datos con aproximadamente 10000 registros y 22 variables (todas ellas numéricas) (Tabla 1). Dados los objetivos que hemos propuesto en el apartado anterior, todos los registros de nuestro conjunto de datos corresponden a ovejas primiparas, es decir, ovejas a las que sólo se les ha controlado una lactación.

- Datos de valoración genética: incluyen la valoración genética de padres y abuelos, así como un valor que indica la fiabilidad de la estimación (proporcionado por BLUP). Se incluye aquí la variable a predecir **BV**, cuya distribución puede observarse en la figura 1.

- Datos de lactación materna: incluye el número de lactaciones controladas a la madre, así como datos medios y máximos en lactaciones de 120 días y de otras duraciones pero normalizadas en cuanto a grasa, ...

- Datos de lactación: mismos datos que para la madre excepto que no aparece la variable de número de lacataciones por ser siempre uno, ni las variables relativas a máximo por coincidir con la media.

- Otros datos: en este caso se ha usado el número de hijos en el parto en el qué nació el animal (entre 1 y 6).

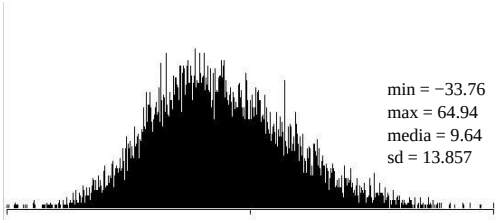

Figura 1: Histograma de la variable BV

Para el proceso de estimación mediante técnicas de aprendizaje automático hemos considerado cuatro casos, dividiendo para cada uno de ellos el conjunto de datos en uno de entrenamiento ($2/3 = 6926$ registros) y otro de test ($1/3 = 2968$ registros).

- Tarea 1: No se consideran las variables de lactación. Se distingue si se incluye o no la variable índice de pedigree como variable predictora. Nótese que esta variable puede entenderse como un proceso construcción de atributos propio de minería de datos [6].

- Tarea 2: Se consideran todas las variables, e igualmente se distingue si se incluye o no la variable índice de pedigree como variable predictora.

3. Técnicas utilizadas

Hemos considerado 6 modelos diferentes para predecir el valor de la variable BV, tres de ellos basados en regresión y otros 3 en un enfoque bayesiano.

3.1. Técnicas basadas en regresión

Las técnicas basadas en regresión que hemos usado en este trabajo son la regresión lineal (LR), los árboles de regresión (RT) y los árboles de modelos (MT).

En el modelo de regresión lineal (LR), si Y es la variable a predecir (o variable dependiente) y $X_1, \ldots, X_n$ son las variables predictoras (o variables independientes), el valor de Y se obtiene ajustando un modelo $\hat{y} = a_1 x_1 + \cdots + a_n x_n$, donde $a_1, \ldots, a_n$ se ajustan a partir de la muestra por mínimos cuadrados. La limitación de este modelo viene dada por la suposición de linealidad en la dependencia entre la variable dependiente y las predictoras, además de la suposición de independencia entre éstas.

Los árboles de regresión (RT) [2], se basan en la discretización del dominio de las variables predictoras representado por una estructura de árbol, en el que cada nodo interior representa una variable predictora y cada arco un intervalo de valores de la variable contenida en el nodo del que sale. En cada nodo hoja, se almacena un número real que representa el valor estimado para la variable a predecir dado que el valor de las variables predictoras se encuentra en la región determinada por el camino que va desde la raíz del árbol hasta la hoja en cuestión. Los árboles de regresión son capaces de representar relaciones no lineales entre las variables predictoras y la dependiente, aunque tiene el inconveniente de que el tamaño de los árboles puede ser muy grande si se pretende obtener una precisión razonable.

Los árboles de modelos (MT) [8] son una

Datos de valoración genética:	Datos de lactación:	Datos de lact. materna:
BVFather, ReBVF, BVMother, ReBVM,	AvLactNorm	NLactM
BVMaternalGM, ReBVMGM,	AvLact120	AvLactNormM
BVParentalGM, ReBVPGM,	Otros:	MaxLactNormM
BVMaternalGF, ReBVMGF,	TypeOfBirth	AvLact120M
BVParentalGF, ReBVPGF, **BV**		MaxLact120M

Cuadro 1: Variables en el conjunto de datos. **BV** es la variable objetivo.

combinación del modelo de regresión lineal y de los árboles de regresión, diferenciándose de éstos en que en cada hoja se almacena una función de regresión lineal en lugar de un único número. De esta forma, se mejora el poder de ajuste del modelo y se disminuye el número de reglas necesarias, si bien las reglas son más complejas al consistir el consecuente en una ecuación de regresión.

3.2. Técnicas probabilísticas

Las técnicas probabilísticas de predicción se basan en la estimación de la distribución de probabilidad de la variable dependiente dado el valor de las variables predictoras. La predicción se hace obteniendo un valor de la variable a partir de la distribución estimada (normalmente la media o la moda). En este trabajo hemos considerado tres modelos basados en la suposición de normalidad de la distribución conjunta de todas las variables involucradas en el modelo, viniendo las diferencias entre los tres modelos determinadas por las distintas suposiciones de independencia entre las variables. Las independencias entre las variables se codifican mediante una red bayesiana (ver, p.e. [7]), que es un grafo dirigido acíclico donde cada nodo representa una variable aleatoria y para cada nodo se especifica una distribución de probabilidad para la variable que contiene dados sus padres. En particular, trabajaremos con las llamadas *redes bayesianas gaussianas* [3], en las que la distribución conjunta es normal multivariante $N(\mu, \Sigma)$, donde μ es el vector de medias y Σ la matriz de covarianzas.

En una red bayesiana Gaussiana, la densidad condiciondada de cada variable dados sus padres es

$$f(x_i|\pi_i) \sim N\left(\mu_i + \sum_{j=1}^{i-1} \beta_{ij}(x_i - \mu_j), v_i\right) ,$$

donde β_{ij} es el coeficiente de regresión de X_j en la ecuación de regresión de X_i sobre sus padres, Π_i, $v_i = \Sigma_i - \Sigma_{i\Pi_i}\Sigma_{\Pi_i}^{-1}\Sigma_{i\Pi_i}^T$ es la varianza condicional de X_i dado $\Pi_i = \pi_i$, Σ_i es la varianza marginal de X_i, $\Sigma_{i\Pi_i}$ es el vector de covarianzas entre X_i y las variables de Π_i y Σ_{Π_i} es la matriz de covarianzas de Π_i. Obsérvese que β_{ij} mide la fuerza de la relación entre X_i y X_j, de forma que si $\beta_{ij} = 0$, entonces X_j no es un padre de X_i.

Hemos considerado tres variantes de redes bayesianas gaussianas:

- **Gaussian Naive Bayes (GNB)**: Se basa en el modelo llamado Bayes ingenuo o Naive Bayes, en el cual se supone que todas las variables predictoras son independientes si se conoce el valor de la variable a predecir. Esto quiere decir que los únicos arcos de la red son los que van desde la variable dependiente a cada una de las variables predictoras, lo que implica que todos los β_{ij} serán iguales a cero salvo cuando X_j sea la variable a predecir. En [4] se hace un estudio basado en este modelo pero usando kernels para modelar las funciones de densidad. La conclusión es que si bien la hipótesis de independencia no resta (mucha) precisión al modelo NB en clasificación, si hace que esta precisión se resienta al realizar predicción numérica.

- **Gaussian TAN (GTAN)**: Se basa en el modelo Tree Augmented Network (TAN),

que es un Naive Bayes donde se admite que cada variable padre tenga un padre más aparte de la variable dependiente. Esto quiere decir que para cada variable X_i, dos valores β_{ij} serán distintos de cero, el correspondiente a la variable a predecir y otro más. La construcción del modelo se hace en dos pasos: (1) se construye una factorización de la distribución conjunta entre las variables predictoras (condicionada a la clase) mediante un modelo gráfico en forma de árbol usando el algoritmo propuesto en [5], y (2) se aumenta el modelo gráfico construido añadiendo una estructura tipo NB, es decir se añade la variable clase y arcos desde ella a todas las variables predictoras.

- **Gaussian Full network (GFULL)**: Se basa en no realizar ninguna suposición de independencia entre las variables, y por tanto la matriz de covarianzas se estima íntegramente a partir de los datos.

3.3. Experimentos

Se han probado los métodos descritos anteriormente para las tareas citadas en la sección 2. La tabla 2 contiene los resultados obtenidos para la tarea 1, mientras que la tabla 3 se refiere a la tarea 2. Se ha probado con distintos casos: uso de todas las variables o uso de distintos grupos (valor genético de los padres, valor genético más lactación, etc...). Concretamente, en ambas tablas BVp se refiere a los datos de valoración genética de los padres, BVAll a todos los datos de valoración genética, lM se refieren a los datos de lactación de la madre, lact a los datos de lactación del individuo y ped al índice de pedigree. En la primera columna se indica el conjunto de variables usado y el número de variables contenidas en el mismo. Cada celda de las demás columnas contiene, en primer lugar, la correlación lineal entre los valores reales y los estimados y en segundo lugar el error cuadrático medio estandarizado de las estimaciones. En la primera fila se indica la correlación y el error producido al usar el índice de pedigree como predictor. Como dato adicional podemos indicar que un predictor constante que siempre devuelva la media aritmética como valor predicho obtiene una correlación de 10e-7 y un error de 13.99.

Como análisis de resultados podemos indicar lo siguiente:

- En todos los casos los mejores resultados son obtenidos por los árboles de modelos, excepto cuando sólo se usa la variable pedigree, que son los métodos probabilísticos los que obtienen mejor resultado.
- Los métodos probabilísticos (al menos bajo las suposiciones aquí realizadas) no son competitivos, obteniéndose los mejores resultados en los casos más sencillos, cuando sólo las variables de valoración genética de los padres son usadas o cuando sólo se usa la variable pedigree. Sí que se aprecia en general que el considerar dependencias mejoran los resultados, siendo habitual que GFULL obtenga mejor resultado que GTAN y éste que GNB, aunque no siempre como puede verse en la primera línea de la tabla 3. En cualquier caso, hay que tener en cuenta que los datos usados no cumplen la hipótesis de normalidad conjunta. Esta hipótesis se ha contrastaso usando el test de Shapiro-Wilk multivariante [9].
- En cuanto al caso de estudio, en lo relativo a la tarea 1, en los mejores casos se ha mejorado la estimación realizada por el índice de pedigree en más de tres puntos en cuanto a correlación y un punto de reducción en el error cuadrático medio estandarizado, lo que corresponde a bajar de 53 a 40 en error cuadrático medio. Creemos por tanto que se trata de una mejora a tener en cuenta. No obstante, la única pega podría ser que se usan muchas variables en la estimación.
- En la tarea 2 se han obtenido correlaciones superiores al 93 %, lo que creemos que es una buena aproximación teniendo en cuenta que se usa mucha menos información que en el método BLUP. De nuevo la pega puede ser que se usan muchas variables (de entre las disponibles).

pedigree		0.8560 / 7.2792					
variables		LR	RT	MT	GNB	GTAN	GFULL
BVp	2	0.8741	0.8749	0.8815	0.8781	0.8795	0.8795
		6.7058	6.6928	6.5183	7.0034	7.0671	7.0671
ped	1	0.8560	0.8532	0.8574	0.8804	0.8804	0.8804
		7.2386	7.3042	7.2071	7.0034	7.0034	7.0034
BVall	12	0.8793	0.8794	0.8884	0.66	0.61	0.57
		6.5751	6.5792	6.3384	18.876	15.231	14.559
BVall+ped	13	0.8730	0.8747	0.8839	0.765	0.549	0.553
		6.8299	6.7908	6.5516	22.549	12.869	13.876
BVall+lM	17	0.8826	0.8791	0.8897	0.69	0.666	0.690
		6.4891	6.5868	6.3023	15.900	15.006	14.550
BVall+lM+ped	18	0.8761	0.8743	0.8862	0.771	0.515	0.585
		6.7524	6.8003	6.4877	25.414	16.847	13.016
All	18	0.8828	0.8791	0.8899	0.185	0.016	0.159
		6.4858	6.5870	6.2985	15.953	15.606	13.540
All+ped	19	0.8763	0.8741	0.8862	0.649	0.714	0.781
		6.7475	6.8061	6.4885	15.516	11.235	8.691

Cuadro 2: Resultados para la tarea 1.

pedigree		0.8560 / 7.2792					
variables		LR	RT	MT	GNB	GTAN	GFULL
BVp+l	4	0.9101	0.9058	0.9191	0.910	0.138	-0.09
		5.8311	5.9897	5.5444	6.339	74.828	102.32
ped+l	3	0.9084	0.9028	0.9103	0.918	-0.035	0.101
		5.8469	6.0207	5.7902	6.046	75.369	74.104
BVall+l	14	0.9141	0.9070	0.9246	0.386	0.387	0.426
		5.7048	5.9523	5.3570	59.760	60.098	61.028
BVall+l+ped	15	0.9163	0.9122	0.9237	0.489	0.485	0.412
		5.6019	5.7367	5.3590	60.059	60.037	59.69
BVall+l+lM	19	0.9244	0.9069	0.9313	0.35	0.331	0.284
		5.3659	5.9562	5.1248	63.075	61.977	59.466
BVall+l+lM+ped	20	0.9265	0.9121	0.9359	0.493	0.494	0.425
		5.2633	5.7402	4.9273	64.71	60.522	58.808
All+l	20	0.9245	0.9068	0.9327	0.223	0.076	0.256
		5.3628	5.9574	5.0731	15.596	15.404	13.540
All+l+ped	21	0.9266	0.9121	0.9359	0.681	0.817	0.816
		5.2590	5.7401	4.9275	15.027	8.815	8.492

Cuadro 3: Resultados para la tarea 2.

4. Selección de variables

En este trabajo no pretendemos hacer un estudio del problema de la selección de variables aplicado a este problema, sino únicamente ver si es posible encontrar de forma *rápida* algunos subconjuntos de variables que puedan usarse como buenos predictores. Por ello, hemos elegido una combinación de métodos de filtrado (*filter*) con métodos de envoltura (*wrapper*) que evalúa a lo sumo $O(2n)$ subconjuntos de variables, siendo n el número de variables predictoras en el conjunto de datos. El algoritmo tiene dos fases, creación de un ranking (filter) y evaluación de subconjuntos basada en dicho ranking (wrapper).

4.1. Creación de un ranking

La creación de un ranking entre las variables suele estar basado en medir de alguna forma la relación entre cada variable predictora y la variable objetivo. Una de las opciones más frecuentes es usar la cantidad de *información mutua (IM)* como medida. La IM es una medida de interdependencia entre variables, y para dos variables se define como:

$$I(X,Y) = H(X) + H(Y) - H(X,Y),$$

donde $H(X)$ es la *entropía* de Shannon y para variables numéricas se define como:

$$H(X) = -\int p(x) log_2(p(x)) dx.$$

Si asumimos que X e Y son variables aleatorias Gaussianas, entonces la IM puede calcularse como una transformación del coeficiente de correlación ρ [10]:

$$I(X,Y) = -\frac{1}{2} log(1-\rho^2).$$

También en el caso de variables aletorias Gaussianas, en [10] se define la información mutua entre dos variables X e Y condicionadas a una tercera Z como

$$I(X,Y|Z) = -\frac{1}{2} log(1 - corr^2(X,Y|Z))$$

con $corr(X,Y|Z)$ el coeficiente de correlación parcial. Es esta medida la que se ha usado en el primer paso de la construcción del modelo GTAN.

En este trabajo (al igual que en otros muchos) realizaremos la suposición de que las variables se modelan con distribuciones gaussianas. La tabla 4 muestra el ranking obtenido calculando $I(X_i, BV)$ para toda variable predictora X_i. Se indica con * las dos variables que no son usadas en la tarea de predicción en ovejas recién nacidas.

Variable	Mut. Inf.
PedigreeIndex	0.6690
BVFather	0.2972
BVMother	0.2936
BVParentalGF	0.0998
AvLac120 (*)	0.0887
BVMaternalGM	0.0684
BVParentalGM	0.0607
ReBVF	0.0589
AvLacNorm (*)	0.0522
ReBVPGM	0.0361
MaxLac120M	0.0303
ReBVPGF	0.0297
BVMaternalGF	0.0257
MaxLacNormM	0.0242
AvLAc120M	0.0195
ReBVM	0.0184
ReBVMGM	0.0181
NLacM	0.0178
AvLacNormM	0.0148
ReBVMGF	0.0119
TypeOfBirth	0.0016

Cuadro 4: Ranking producido por IM.

4.2. Selección de un subconjunto

La selección de variables basada en un ranking suele conducir a la inclusión de las k primeras variables del ranking en el subconjunto seleccionado. Sin embargo, esto plantea (al menos) dos problemas:

- La elección de k.
- Si el ranking se ha obtenido (como es nuestro caso) midiendo de forma individual cada variable predictora con la variable objetivo, es habitual que se añadan

variables redundantes al subconjunto seleccionado.

El primer problema puede solventarse probando con subconjuntos que incluyen hasta la variable k, $k = 1, \ldots, n$, evaluando (de modo wrapper) cada subconjunto con el método de aprendizaje seleccionado y eligiendo el valor de k que haya producido el mejor resultado (mínimo error, mayor correlación, etc...). Este es el procedimiento seguido por ejemplo en el método *Selective Ranking Naive Bayes* [1], que tiene la ventaja de evaluar $O(n)$ subconjuntos, pero que en general no solventa el problema de incluir variables redundantes entre sí en el subconjunto seleccionado.

Aquí proponemos un esquema parecido al anteriormente descrito pero que sólo añade un atributo al subconjunto si mejora el resultado (evaluación wrapper) respecto al subconjunto actual. Además, la condición de parada se modifica de forma que el algoritmo se detiene si l atributos consecutivos en el ranking no han sido añadidos al subconjunto. El parámetro l puede verse como un parámetro de *lookahead*, ya que nos permite rechazar hasta l atributos *redundantes* en busca de un nuevo atributo significativo. Adicionalmente, como alguno de los atributos añadidos puede pasar a ser redundante al añadirse otros con posterioridad, dotamos al algoritmo de una fase *backward* que evalúa (modo wrapper) la bondad del subconjunto actual sin cada atributo, eliminándolo en caso de no empeoramiento. Por tanto, la complejidad máxima del método es $O(2n)$, pero raramente se alcanza, puesto que no es habitual que todos los atributos sean incluidos en la fase *forward* del algoritmo.

4.3. Experimentos

En esta subsección indicamos los experimentos realizados para selección de variables. Hemos considerado las dos tareas: recién nacidos (tabla 5) y aproximación a BLUP (tabla 6). En cada caso hemos probado con y sin la inclusión de la variable índice de pedigree. En todos los casos se ha usado el valor de *lookahead* ($l = 5$), que se ha mostrado como satisfactorio en nuestros experimentos preliminares. El primer bloque de cada tabla hace referencia al algoritmo de selección anteriormente descrito. Si observamos la tercera línea de cada caso vemos que, en general, para los métodos basados en regresión se añaden muchos atributos. Un anális detallado de la situación nos muestra que en ocasiones se añaden atributos que sólo mejoran en la sexta o séptima cifra decimal la correlación (métrica usada para medir la bondad de cada subconjunto en nuestro caso) obtenida sin ellos. Debido a esto hemos modificado el criterio de inclusión exigiendo una mejora mínima. En este caso (segundo bloque de las tablas) vemos que exigiendo un uno por mil de mejora en la correlación, las métricas apenas se resienten y sin embargo, el número de variables incluidas disminuye drásticamente. No ocurre lo mismo con los métodos probabilísticos, que directamente seleccionan muy pocas variables y, por tanto, no hay diferencia al poner el umbral del 1 por mil.

Concretamente se han obtenido buenos predictores *simples* con los siguientes subconjuntos de variables:

• Tarea 1.- Árboles de modelos: Subconjunto (BVFather, BVMother, ReBVF). Este predictor mejora al índice de pedigree en 3.5 puntos en correlación y sólo usa 3 variables. Llama la atención el uso de la fiabilidad del valor paterno y no el del materno, así como el no usar ninguna variable de lactación materna.

• Tarea 2.- Árboles de modelos: Subconjunto (BVFather, BVMother, AvLact120, BVParentalGM, ReBVF, ReBVM, AvLact120M). Obtiene una correlación de más del 93 % con respecto a BLUP usando sólo 7 variables predictoras que además son *razonables*, valores genéticos de los padres junto con la fiabilidad de estas mediciones más datos de lactación del propio animal y la madre. Además se usa la valoración genética de la abuela materna.

• Métodos probabilísticos. En este caso se aprecia que la selección de variables aquí propuesta mejora significativamente su comportamiento. En términos de error y correlación, los modelos probabilísticos siguen quedando ligeramente por detrás de la regresión y los árboles de modelos, pero superan a los árboles

pedigree		0.8560 / 7.2792					
pedigree		LR	RT	MT	GNB	GTAN	GFULL
false	corr	0.8827	0.8782	0.8923	0.8781	0.8888	0.8805
	rmse	6.4858	6.6076	6.2303	7.0034	6.9954	7.1833
	#att	14	5	10	2	3	3
true	corr	0.8761	0.8756	0.8864	0.8804	0.8904	0.8904
	rmse	6.7517	6.7680	6.4823	6.9478	6.7008	6.7008
	#att	10	9	9	1	2	2
false	corr	0.8808	0.8789	0.8896	0.8781	0.8888	0.8888
	rmse	6.5358	6.5922	6.3033	7.0034	6.9954	6.9955
	#att	6	3	3	2	3	3
true	corr	0.8747	0.8760	0.8861	0.8804	0.8904	0.8904
	rmse	6.7863	6.7574	6.4895	6.9478	6.7008	6.7008
	#att	6	4	4	1	2	2

Cuadro 5: Resultados para predecir-todos con FW

pedigree		0.8560 / 7.2792					
pedigree		LR	RT	MT	GNB	GTAN	GFULL
false	corr	0.9243	0.9078	0.9321	0.8934	0.9033	0.9143
	rmse	5.3687	5.9314	5.0925	6.8884	6.4091	6.0798
	#att	14	6	16	3	3	3
true	corr	0.9263	0.9125	0.9344	0.9157	0.9210	0.9229
	rmse	5.2683	5.7297	4.9804	6.1590	5.8694	5.8385
	#att	16	10	12	2	2	3
false	corr	0.9230	0.9079	0.9315	0.8934	0.9033	0.9143
	rmse	5.4124	5.9261	5.1160	6.8884	6.4091	6.0798
	#att	6	4	7	3	3	3
true	corr	0.9252	0.9119	0.9360	0.9157	0.9210	0.9229
	rmse	5.3082	5.7486	4.9224	6.1590	5.8694	5.8385
	#att	6	6	8	2	2	3

Cuadro 6: Resultados para contrablup con FW

de regresión. Llama también la atención que añaden un número muy pequeño de variables al comparar con los métodos de regresión. En cuanto a los predictores seleccionados, en la tarea 1 básicamente se usa sólo el pedigree, o cuando esta variable no está incluida, se seleccionan las dos valoraciones de los padres complementadas con MaxLactNormM o ReVPGM dependiendo del predictor. En la tarea 2, el pedigree o las dos valoraciones de los padres son complementadas con la variable de lactación AvLact120, resultando en un predictor muy destacado también en otros trabajos [6].

Por último, indicar que computacionalmente el proceso es muchísimo menos costoso que con los árboles de regresión o árboles de modelos.

5. Conlusiones

En este trabajo se ha evaluado el uso de técnicas basadas en regresión y métodos probabilísticos para la predicción numérica del valor genético en ovejas de raza manchega. Se han contemplado dos casos: predicción en caso de ovejas recién nacidas y predicción en ovejas después de su primer parto. En el primer caso

el objetivo era comparar contra la medida al uso, el índice de pedigree, habiéndose obtenido mejoras de más de tres puntos en el coeficiente de correlación y usando predictores simples basados en sólo tres o cuatro variables. En el segundo caso el objetivo era obtener una buena estimación del valor genético con respecto a la proporcionada por BLUP pero usando mucha menos información y pudiéndose realizar en cuanto los datos estén disponibles, sin tener que esperar a la evaluación semestral de toda la cabaña. En este sentido se han obtenido correlaciones de más del 93 % con respecto a BLUP, lo que consideramos que es un buen resultado.

Por otra parte nos proponíamos estudiar las prestaciones de los métodos basados en redes Gaussianas frente a los basados en regresión. En este sentido lo único que podemos decir es que con las suposiciones (normalidad de las variables) aquí realizadas, ni siquiera el hecho de considerar dependencias entre las variables ha contribuido a obtener unos buenos resultados, quizás precisamente a la violación de la hipótesis de normalidad conjunta. Sin embargo, después de aplicar los métodos de selección de variables FW el rendimiento de estos modelos mejora sustancialmente usando además un número de variables realmente bajo. Creemos, no obstante, que es necesario trabajar tanto en el modelo estimado como en la técnica de predicción usada para que estos métodos mejoren.

Referencias

[1] I. Inza A. Pérez, P. Larrañaga. Supervised classification with gaussian networks. filter and wrapper approaches. In *Tendencias de la minería de datos en España (Eds. Giráldez, Riquelme, Aguilar)*, pages 379–390, 2004.

[2] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and regression trees.* Wadsworth, 1984.

[3] E. Castillo, J.M. Gutiérrez, and A.S. Hadi. *Expert systems and probabilistic network models.* Springer-Verlag, New York, 1997.

[4] E. Frank, L. Trigg, G. Holmes, and I.H. Witten. Technical note: Naive Bayes for regression. *Machine Learning*, 41:5–25, 2000.

[5] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29:131–163, 1997.

[6] 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, Departamento de Informática. Universidad de Castilla-La Mancha, January 2005.

[7] Finn V. Jensen. *Bayesian networks and decision graphs.* Springer, 2001.

[8] J.R. Quinlan. Learning with continuous classes. In *Proceedings of the 5th Australian Joint Conference on AI*, pages 343–348, 1992.

[9] J.P. Royston. Some techniques for assessing multivariate normality based on the Shapiro-Wilk W. *Applied Statistics*, 32:121–133, 1983.

[10] J. Whittaker. *Graphical Models in Applied Multivariate Statistics.* Wiley, 1990.