RECONOCIMIENTO DE PATRONES DE CURVAS DE LACTANCIA MEDIANTE RED NEURONAL Y ANALISIS DISCRIMINANTE, AL PRIMER TERCIO DE LACTANCIA A CONTROLES DE VACAS LECHERAS DE LA IX REGION.

Claudio Sebastián Cárdenas Mansilla.
Académico encargado del área de estadística, Instituto de Formación Continua, Universidad de los Lagos, Sede Castro, El mirador, Pasaje los Franciscanos Nº3, Castro. Fono: 65-322470- 322471, ccard007@gmail.com

ABSTRACT

Recognition of lactation pattern curves through neural networks and discriminate analysis, during the first third of lactation in dairy cows from the IX Región.

Key words: Lactation Curves, Neural Networks, Discriminant Analysis.

In this study the efficiency of the Artificial Neural Models and Linear Discriminant Analysis was demonstrated, to recognize and classify models or forms of lactation curves. The optimal neural model, using the Multi-level Perceptron with a hidden layer was obtained when five nodes were used (Artificial Neuron) in the inner layer. Although a higher performance of the neural models was observed compared to the linear discriminant analysis, this was slight and there was no difference in the statistical test of proportional difference (P-Value = 0.7224). This can be illustrated in the percentage of achievements in the classification for both methodologies; 97.87% for the neural model and 95.7% for the linear discriminant models, both in the training sample and validation sample.

RESUMEN

Palabras Claves*:* Curvas de Lactancia, Redes Neuronales, Análisis Discriminante

En este trabajo se demostró la eficiencia de los modelos Neuronales Artificiales y Análisis Discriminante Lineal, en procesos de reconocimiento y clasificación de patrones o formas de curvas de lactancia. El modelo neuronal óptimo, utilizando el Perceptron Multinivel con una capa oculta, se obtuvo cuando su utilizaron cinco nodos (Neurona Artificial) en la capa intermedia. También, aunque, se observó una mayor rendimiento de los modelos neuronales en comparación con el análisis discriminante lineal, este fue leve no observándose una diferencia significativa en la prueba estadística de diferencia de proporciones (P-Value = 0.7224), esto se refleja en el porcentaje de aciertos en la clasificación para ambas metodologias; 97.87% para el modelo neuronal contra un 95,7% para los modelos discriminantes lineales, esto tanto en la muestra de aprendizaje (Training), como en la de validación.

INTRODUCCIÓN

La necesidad de aumentar la eficiencia del sistema de producción de leche hace conveniente conglomerar los animales en grupos productivos. Una evaluación importante en cualquier sistema de producción, es la detección de unidades productivas que presenten bajo rendimiento o anomalías, que afectan la eficiencia global del sistema. Esto, en la producción lechera, implica el reconocimiento oportuno de las curvas de lactancia que presentan comportamientos que se escapan a las tendencias consideradas normales (Cárdenas, 2003).

Recepción de originales: 3 de enero de 2008

La primera etapa en todo diseño de reconocimiento, consiste en el establecimiento de las clases: en lo que se podría denominar como la definición del universo de trabajo del sistema. En la mayoría de los casos ésta es directa y trivial. No obstante, puede ocurrir que las clases sean desconocidas a priori, situación que se presenta en ciertos campos de la biología en donde no está clarificado el universo de clases. En tales situaciones, se recurre a técnicas denominadas de conglomeración o clustering (Maravall, 1993). Una vez superado esto, resultaría pertinente aplicar un método clásico de reconocimiento de patrones llamado; Análisis Discriminante, que es una técnica estadística que permite estudiar las diferencias entre dos o más grupos de objetos con respecto a varias variables, simultáneamente, la cual, es una técnica de clasificación y asignación de individuos a grupos, a los cuales, se les conoce sus características (Cuesta,1992). Pero en el último tiempo aparece otro instrumento aplicable al reconocimiento de patrones, esta es, La teoría de los Modelos Neuronales Artificiales, los cuales, se inspira en la estructura y funcionamiento del Sistema Nervioso, donde la neurona es el elemento fundamental (Hilera y Martínez, 1995), considerando siempre que los elementos básicos de un sistema neuronal biológico son las neuronas, que se agrupan en conjuntos compuestos por millones de ellas organizadas en niveles constituyendo un sistema de funcionalidad propia. Según esto, el elemento esencial de partida será la neurona artificial que se organizará en capas, varias de estas constituirán una red Neuronal Artificial (Martín del Brio y Sanz, 2002).

El objetivo general de este trabajo fue evaluar el grado de discrepancia en la capacidad de reconocer y clasificar curvas de lactancia de las redes neuronales artificiales y análisis discriminante lineal con la finalidad de reconocer patrones productivos al primer tercio de lactancia.

MATERIALES Y MÉTODOS

El trabajo se realizó a partir de información de lactancias recopilada en el rebaño lechero de la Estación Experimental Maipo de la Universidad de la Frontera y de los controles lecheros de la actual Estación Experimental de la misma Universidad; el Fundo Maquehue.

La muestra se elaboró en forma selectiva con relación a la tendencia de producción de leche, seleccionando sólo lactancias completas. La muestra estuvo conformada por 108 lactancias, esta muestra está conformada por controles diarios lo cual es equivalente a 300 controles por lactancia, las cuales se distribuyen desde el año 1997 hasta el año 2001. Además, se contó con la Información de control lechero oficial recopilada por la SOFO, en la Novena región, donde se tomó una muestra de 80 lactancias. Esta muestra estuvo conformada por lactancias controladas mensualmente (cada 30 días aproximadamente), correspondiente a los años 1998 al 2001. Software utilizado en el diseño de redes El programa computacional usado en el diseño y la ejecución de las diferentes estructuras neuronales evaluadas fue Neural Connection 2.0, propiedad intelectual de SPSS Inc. y Recognition Systems Inc (Nconnection 1997). En este trabajo se siguió un procedimiento, de identificación y caracterización de tipologías presentes en lactancias de vacas lecheras de la IX región, basado en un análisis cuantitativo multidimensional y de estadística multivariante, con el propósito de generar un conjunto de patrones productivos que sea confiable y representativo de la población de individuos en estudio, que en una etapa posterior permitió evaluar y validar la eficiencia en el reconocimiento de patrones y clasificación de curvas de lactancia, de modelos de redes neuronales y análisis discriminante lineal.

Se seleccionaron las mediciones iniciales que presentan controles con mayor nivel de estabilidad de la producción inicial, con lo cual, se pudo determinar que a partir del día Nº8 de la lactancia se normaliza la distribución de la producción. Se identificaron cinco tipologías o grupos productivos de lactancia en función de los cuales se modelaron los patrones o formas de curvas de lactación. Dos de los cinco grupos identificados presentaron comportamientos normales de producción, lo cual correspondió a un 42.01% del total de la muestra. Los tres conglomerados restantes presentaron comportamientos productivos atípicos (57.99% de la muestra).

RESULTADOS Y DISCUSION

Validación de los modelos de clasificación obtenidos Mediante Análisis Discriminante Lineal

En el proceso de validación de las metodologías se utilizó el 25% de la información disponible, en términos de números de lactancias. Aquí pudo observarse la buena calidad y la eficiencia clasificadora de los modelos estimados mediante Análisis Discriminante Lineal. Los modelos evaluados corresponden a los obtenidos mediante la metodología de estimación; Por Etapas o Stepwise y Simultánea a través de los cuales se estimaron cuatro funciones de discriminación y además se obtuvieron las variables con la mayor capacidad discriminatoria. Estos modelos obtuvieron un porcentaje de clasificación correcta de un 95.7% de aciertos, tanto, para los datos de la muestra utilizada para entrenar el modelo y la usada para validarlo. Los modelos derivados mediante el Método Simultáneo (Introducir todas las variables) alcanzaron 90.8% de aciertos en la muestra de entrenamiento y un 91.5% en la muestra de validación. Es importante mencionar que se observaron discrepancia en los métodos utilizados, si bien en ambos casos, las ecuaciones de clasificación solo estuvieron conformadas por 4 predictores (días de controles), solamente hubo coincidencia en dos de las cuatro variables independientes; esto es, en las variables Pro0 (día 8) y Pro1(día 38), de tal manera que los modelos estimados mediante el método simultáneo adoptaron las variables Pro01 (día 23) y Pro2 (día 68), mientras tanto que las funciones discriminatorias estimadas a través del método por etapas consideraron las variables Pro23 (día 83) y Pro4 (día 128). Es evidente que los modelos estimados por uno u otro método difieren en lo referente al intervalo de tiempo utilizado para la clasificación de lactancias y en el periodo de tiempo requerido para clasificar, es así, como los modelos estimados a través del Método Simultáneo discriminan mejor utilizando mediciones de menor amplitud de tiempo (15 – 15 – 30 días, lapso entre mediciones) y son capaces de clasificar una curva de lactancia a los 60 días de lactación. Por su parte, los modelos derivados mediante el Método por Etapas discriminan mejor utilizando mediciones de mayor amplitud (30 – 45 – 45 días, lapso entre mediciones) y requieren en forma integra del periodo de estudio (120 días).

Para valorar las contribuciones de los nueve predictores se emplearon las cargas discriminantes, las cuales representan la correlación entre la variable independiente y la puntuación D discriminante, además de los coeficientes estandarizados de las funciones discriminantes.

Cuadro 1. Matriz de estructura o cargas discriminantes de las variables ordenadas por el tamaño de correlación en la función, con el método por etapas.
Table 1. Matrix structure or discriminating burdens of the variables ranked by the size of the correlation function, the method by stages.

Variables	Función 1	Función 2	Función 3	Función 4
Pro23	**0.727***	-0.035	-0.506	-0.462
Pro3 [a]	**0.683***	-0.013	-0.422	-0.138
Pro34 [a]	**0.681***	0.115	-0.426	0.229
Pro12 [a]	**0.615***	-0.361	0.522	-0.131
Pro2 [a]	**0.582***	-0.117	-0.169	-0.441
Pro0	0.237	**0.832***	0.434	-0.251
Pro01 [a]	0.466	0.230	**0.848***	-0.026
Pro1	0.514	-0.265	**0.812***	0.076
Pro4	0.535	0.326	-0.403	**0.667***

[a]. Variable que no es empleada en el análisis
*. Mayor correlación absoluta entre cada variable y cualquier función discriminante

En el Cuadro 1, se revela el hecho de que las variables Pro3 (día 98) y Pro34 (día 113) cuentan con el segundo y tercer mayor valor y no fueron incluidas. Este hecho destaca que la solución por etapas fue afectada debido a la redundancia entre variables altamente correlacionadas. En la solución simultánea se presenta un fenómeno similar producto de que muchas variables no superaron los criterios mínimos de tolerancia.

Sensibilidad y especificidad

Según los objetivos de este estudio, sólo se presentan las medidas de valor diagnóstico de **sensibilidad** P*r*(P/A) y **especificidad** P*r*(no-P/N) alcanzadas por los modelos discriminantes estimado mediante el procedimiento por etapas, debido a la mayor eficiencia clasificadora alcanzada por las funciones discriminantes estimadas mediante este método. Los resultados obtenidos alcanzaron niveles de **sensibilidad** y **especificidad** del 96.55% y 100% respectivamente, por tanto la Proporción de casos Atípicos (A), según el criterio de referencia, que son identificados correctamente alcanza el 96.55 de cada cien veces y la proporción de casos normales (N), según el criterio de referencia e identificados correctamente por el modelo, se presenta en la totalidad de los eventos.

Validación de los modelos de Redes Neuronales Artificiales obtenidos

Al analizar la metodología de las redes neuronales artificiales, utilizando el Perceptron Multinivel y siguiendo el procedimiento de análisis propuesto pudo observarse que el modelo óptimo, para esta metodología de obtención de patrones y entrenamiento propuestas y además utilizando solamente una capa oculta de unidades de procesamiento, se obtuvo con 5 nodos en la capa oculta, donde se observó un 97.87% de aciertos tanto en los datos de la muestra de entrenamiento como en la de validación. Posterior a esto, se presenta el típico comportamiento de sobre-entrenamiento manifestado por diferentes autores, es decir, el modelo se vuelve muy específico a la información presentada en la muestra de entrenamiento **(T)**, aprendiendo incluso el ruido presente en los datos, como consecuencia se eleva el **Error de generalización (Error-G)**, lo cual provoca una declinación progresiva en el porcentaje de acierto de los distintos modelos

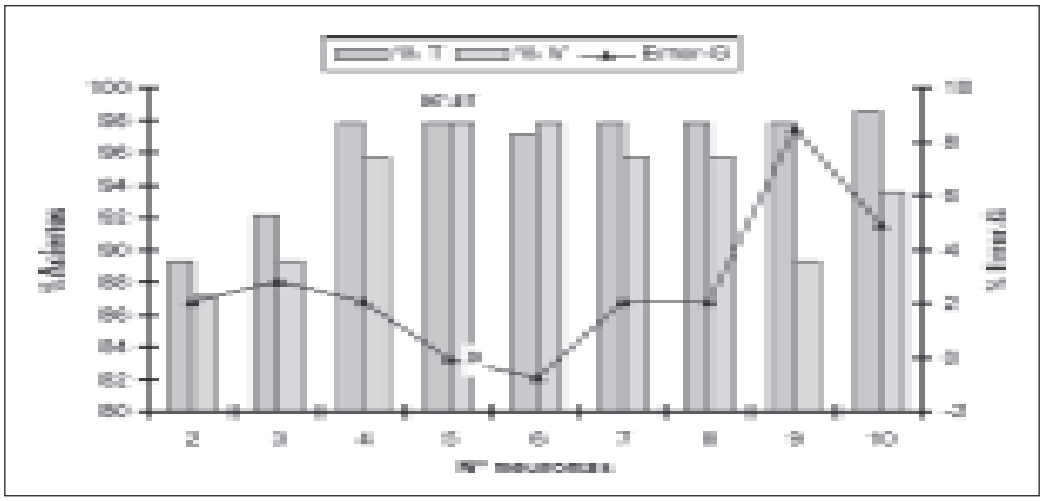

Figura 1. Optimo de unidades de procesamiento en la capa oculta basado en el porcentaje de aciertos obtenidos para la muestra de entrenamiento (T) y validación (V) y error de generalización (Error-G) de los modelos.
Figure 1. Best of processing units in the hidden layer based on the percentage of hits obtained for the sample of training (T) and validation (V) and fallacy of composition (Error-G) of the models.

Cuadro 2. Prueba de diferencia entre dos proporciones, para aciertos clasificatorios entre Red Neuronal y Análisis Discriminante
Table 2. Evidence of difference between two ratios for hits qualifiers between Neural Network Analysis and Discriminant analysis.

Z	**0.5985**
Significancia Bilateral (P-value)	0.7224

estimados a la hora de clasificar los vectores presentados en la muestra de validación **(V).**

Evaluación de la sensibilidad y especificidad de la Red Neuronal

Las medidas de valor diagnóstico de Sensibilidad P*r*(P/A) y Especificidad alcanzadas por el modelo neuronal estimado en este estudio alcanzaron los niveles máximos, es decir, Sensibilidad y Especificidad alcanzaron valores de 100%. Según esto, la Proporción de casos Atípicos, según el criterio de referencia que son identificados correctamente y proporción de casos normales según el criterio de referencia e identificados correctamente por el modelo es de un 100%.

A continuación se presentan los resultados del contraste estadístico basado en la prueba de diferencia de proporciones (Walpole y Myers, 1992:Pag; 347) que se realizó sobre los proporciones de aciertos obtenidos en la muestra utilizada para validar al modelo.

El nivel más bajo de significancia en el cual el valor observado del estadístico de prueba es significativo es considerablemente elevado, lo cual lleva a concluir que existe suficiente evidencia estadística para pensar que no existen diferencias significativas en las proporciones de aciertos clasificatorios obtenidos tanto por las Redes Neuronales como los modelos lineales de clasificación estimados mediante Análisis Discriminante.

CONCLUSIONES

La metodología de análisis discriminante permite identificar controles de producción, utilizados como predictores, que presentan mayor capacidad de discriminación.

La eficiencia de la técnica estadística multivariante de clasificación utilizada en este estudio se ve afectada por la redundancia de información, es decir controles de producción utilizados como predictores altamente colineales. Este hecho genera funciones de discriminación que sólo utilizan los predictores (controles) con mayor capacidad discriminante, esto afecta la eficiencia clasificadora de los modelos discriminantes estimados, debido a que las funciones sólo discriminan en base a los controles utilizados en el análisis y no considera los excluidos, en consecuencia se pierde continuidad en el perfil del patrón debido a la reducción de información.

El Modelo Neuronal Artificial óptimo, para la metodología de análisis seguida en este estudio, se obtuvo al incluir cinco unidades de procesamiento (nodos), en la capa oculta (hidden layer), con lo cual se obtuvo un 97.87% de aciertos, tanto en la muestra utilizada en el entrenamiento como en la usada para validar el modelo.

Existe suficiente evidencia estadística, según los datos de esta muestra (P-Value = 0.7224), para concluir que no existen diferencias significativas en las proporciones de aciertos clasificatorios obtenidos tanto por las Redes Neuronales como los modelos lineales de clasificación estimados mediante Análisis Discriminante.

AGRADECIMIENTOS

Quiero agradecer a mis Profesores de la Universidad de la Frontera, Dr. Horacio Miranda y Dra. Sonia Salvo, por su consejo y orientación científica en el desarrollo este trabajo y además por haber y seguir siendo para siempre mis maestros.

BIBLIOGRAFIA

CARDENAS, C. 2003. Reconocimiento de Patrones y Clasificación de Curvas de Lactancia Mediante Redes Neuronales y Análisis Discriminante Lineal Aplicadas al Primer tercio de Lactancia a Controles de Vacas Lecheras de la IX Región. Tesis Ingeniero Agrónomo. Universidad de la Frontera. Temuco, Chile. 98 p.

CUESTA, M. 1992. Análisis Discriminante. In: Vallejo, G. (ed.). Análisis Multivariantes Aplicados A las Ciencias del Comportamentales. Universidad de Oviedo. Oviedo, España. 286 p.

HILERA, J.; MARTÍNEZ, V. 1995. Redes Neuronales Artificiales; fundamentos, modelos y aplicaciones. Primera Edición. RA – MA Editorial. Madrid, España. 390 p.

MARAVALL, D. 1993. Reconocimiento de Formas y Visión Artificial. Primera Edición. RA – MA Editorial. Madrid, España. 433 p.

MARTÍN DEL BRIO, B.; SANZ, A. 2002. Redes neuronales y sistemas difusos. Segunda edición. Ed. RA-MA. Madrid, España. 399 p.

SPSS INC./ RECOGNITION SYSTEMS INC. 1997. Neural Conection 2.0 Use‘s Guide. Chicago, USA. 267 p.

WALPOLE, R.; MYERS, R. 1992. Probabilidad y Estadística. Cuarta edición (tercera edición en español). Editorial McGRAW-HILL. Ciudad de México, México. 797 p.