available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/compag

Mining the ESROM: A study of breeding value classification in Manchego sheep by means of attribute selection and construction

*M. Julia Flores**, José A. Gámez, Juan L. Mateo*

Departamento de Sistemas Informáticos & SIMD-i^{3}A, Escuela Politécnica Superior, Universidad de Castilla-La Mancha, 02071 Albacete, Spain

ARTICLE INFO

Article history:
Received 21 February 2005
Received in revised form
20 April 2005
Accepted 2 August 2007

Keywords:
Manchego sheep
Selection scheme
Breeding value
Classification algorithms
Data mining
Attribute selection
Attribute construction

ABSTRACT

Manchego sheep breeding represents an important factor in the economy in the region of Castilla-La Mancha, Spain. For this reason, the selection scheme for Manchego sheep (ESROM) was created to improve milk production in ewes belonging to the Manchego breed. This scheme relies on the use of several tools that depend on ewes' genetic merit, which is calculated by using a sophisticated linear regression model. This paper presents a study about how the use of data mining techniques can help to approximate the genetic qualities of a ewe, before the *official* 6 months assessment is carried out, and by using less input. This study focuses on two well-known data mining tasks: pre-processing and classification. In the pre-processing stage, state-of-the-art algorithms and new proposals are used to identify relevant subsets of features by means of selection and construction. By using these subsets of highly predictive variables, different classifiers are trained, which in turn, are used to assess the genetic quality merit of any given ewe. As a result, original and constructed relevant variables have been identified for the target problem, this being a valuable result in itself. Furthermore, from simulated tests, reliable classification rates have been obtained when using the identified classifiers to approach ESROM tasks.

1. Introduction

In the area of Castilla-La Mancha, Spain, sheep breeding represents one of the key components in the regional economy. There are two main final products from Manchego sheep: (1) Manchego cheese[1] and (2) Manchego lamb,[2] both of them being of excellent quality.

Due to the crisis suffered during recent years in the sheep meat market, milk production has attained a leading role in sheep breeding, and foreign breeds having higher milk production represent a menace to the Manchego breed, even though higher milk production does not always mean greater net profit. In response to this potential threat, several public organizations and regional authorities have opted for the improvement of production data in Manchego sheep,

* *Corresponding author*. Tel.: +34 967599200; fax: +34 967599224.
E-mail addresses: julia@dsi.uclm.es (M.J. Flores), jgamez@dsi.uclm.es (J.A. Gámez), juanlmc@dsi.uclm.es (J.L. Mateo).

[1] Controlled by a Guarantee of Origin (CRDOQM, 2004) since 1984 which requires it to be made only from milk obtained from Manchego ewes raised in the region of Castilla-La Mancha (which in 2001 represents 40% of the total ewe milk in Castilla-La Mancha).

[2] Controlled by a Specific Guarantee (CRDECM, 2004) since 1995.

0168-1699/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.compag.2007.08.004

especially in potentially hazardous areas. To achieve this goal the selection scheme for Manchego sheep (ESROM) was created in 1987.

The ESROM selection scheme (SS), which is similar to other selection schemes developed for other breeds, includes a series of activities whose joint purpose is the *genetic improvement of the breed in terms of milk production*. This scheme is being used by several organizations. So far the results of this scheme have been successful, with each ewe born from artificial insemination producing 25 l more milk as compared to before using ESROM (ITAP, 2001).

The SS has four main tools:

(1) *Genealogical ranking*. A register of all the ewes in a stock farm submitted to the milk controls performed by the SS.
(2) *Stud catalogue*. Males included in the SS having a very high genetic merit (computed from their daughters' genetic merit).
(3) *Milk production control*. A report of the lactation of each ewe containing all the data pertaining to the quantity and quality of milk produced by the ewe during lactation (normalized to allow comparisons).
(4) *Stud market*. Market of males obtained by artificial insemination, which follow the racial standard[3] of Manchego sheep and whose mothers are above the 70th percentile in the genealogical ranking. Acquiring males in the stud market constitutes an easy way to improve the genetic merit of a herd for those stock farmers that cannot enter in the technical part of the ESROM program (artificial insemination and milking monitoring).

The key parameter in the SS is the estimation of animals' genetic merit or breeding value (BV), because it is this value, computed by using the data about the monitored lactations, which allows them to be placed in the genealogical ranking and to be recorded in the stud catalogue or market. In addition, the SS encourages stock breeders to select their flock replacements on the basis of the animal genetic merit. The BV of an animal is a numeric value which represents the deviation of the animal with respect to the averaged breeding value of the Manchego ewes born in 1990 (referred to as the base year).

The estimation of the breeding value[4] has been calculated by using the Best Linear Unbiased Predictor (BLUP) methodology (Henderson, 1975), and more precisely the *animal model* (Jurado, 1994), which is considered to be the most sophisticated method of BLUP analysis available. The estimation of the breeding value using BLUP is a very complex process, in that each specific case of SS needs to be carried out every 6 months in a specialized center.

The aim of this research was to work on the prediction and classification of the breeding value inside the ESROM scheme by using techniques from the machine learning (Mitchell, 1997) and data mining fields (Fayyad et al., 1996). These techniques are embraced by a broader field, called artificial intelligence, whose application to agriculture has sparked interest over recent years (Murase, 2000; Farkas, 2003). Different data mining-based approaches related with dairy production can be found. For example, Abbass et al. (1999) and Macrossan et al. (1999) reported on work to predict dairy daughter milk production from a particular mating by using Bayesian neural networks, while Pietersma et al. (2005) investigated the use of data mining to characterize herds with insufficient growth. As in those studies, our intention was not to replace the use of BLUP, but to work in parallel with it in order to study the possibilities of predicting the BV of an animal by using a data-driven approach, which requires less information input than BLUP. We sought a simpler approach, and one that provides output in less time than a full breeding 6-month evaluation. Since a great part of the current study is based on feature extraction, another aim was to identify those subsets of variables strongly related to the BV variable, which can be used as input knowledge even for BLUP or related techniques.

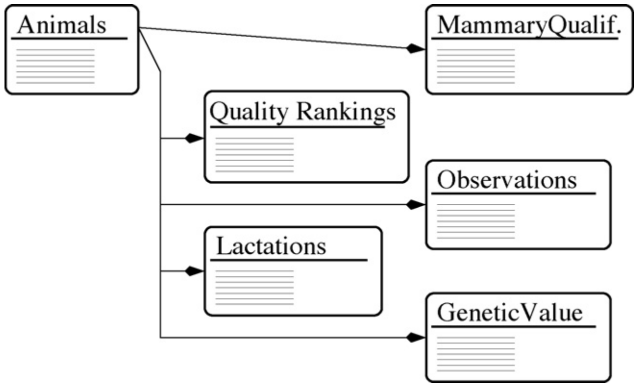

Fig. 1 – Structure of data tables in AGRAMA.

2. Data preparation and selection

Data preparation (Fayyad et al., 1996) is an important process, and in most cases, one of the most time consuming. It comprises a series of stages such as creating a target dataset, as well as cleaning and pre-processing it.

2.1. *Data preparation*

In a data mining project, the source of data is usually a data warehouse, however AGRAMA (National Association of Manchego sheep breeders) does not have a data warehouse in its organization. In AGRAMA, the data[5] are stored in a relational database system as a set of tables which are linked by means of a set of attributes used as keys (the structure is shown in Fig. 1). There are six main tables in the system:

[3] Described in http://www.agrama.org/prototipo.html (in Spanish).

[4] In fact we should use estimated breeding value (EBV), but for the sake of simplicity the notation of breeding value (BV) will be maintained, although it is clear that we are dealing with estimations all the time.

[5] In this work we have used the years from 1989 to 2003 as source data.