Quality Control of National Genetic Evaluation Results Using Data-Mining Techniques; A Progress Report

G. Banos[1], P.A. Mitkas[2], Z. Abas[3], A.L. Symeonidis[2], G. Milis[2] and U. Emanuelson[4]

[1]*Faculty of Veterinary Medicine, Aristotle University of Thessaloniki, Greece*
[2]*Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Greece*
[3]*Department of Agricultural Development, Democretus University of Thrace, Orestiada, Greece*
[4]*Interbull Centre, Uppsala, Sweden*

Background

Data quality constitutes one of the most critical issues in genetic evaluations both at national and international level. International genetic evaluations computed by Interbull are based on the analysis of national genetic evaluation results. Therefore, the validity of international comparisons depends on the quality of the output of the various national genetic evaluation systems. The current method for data quality assurance is mainly determined by the consistency of consecutive evaluation results and is based on thorough statistical examination (Klei *et al.,* 2002). In a separate project, national genetic evaluation programs are being tested on simulated datasets with known properties (Täubert *et al.,* 2002).

Data-mining (DM) provides a different perspective on data quality control. DM is an algorithm-based, data-driven approach in the knowledge discovery process. It is defined as the extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases (Han and Kamber, 2000). In terms of evolutionary steps, DM can be thought of as the new millennium's milestone, following data warehousing and decision support systems (1990s), data management using relational databases (1980s) and traditional data collection (1960s).

An attractive feature of DM, compared to statistical analysis, is that no assumptions on data structure are required to validate consistent and replicable pattern hypotheses. The DM inference process seeks to identify modeling procedures that have a high probability of near-optimality over all possible dimensions of data. The process identifies trends, correlations, discrepancies, irregularities and disruptions, and can make useful predictions and inferences to data continuity and quality. In other words, DM algorithms "learn" from the data and ultimately create "knowledge" for the analyst.

DM techniques have been widely applied in various business areas including telecommunications, market research, financial data analysis and the retail industry. Furthermore, one of the pioneering application domains of DM technology is bioinformatics and genetic analysis. In animal science, Abbass *et al.* (1999) considered DM techniques in deriving predictions of dairy bull daughter performance from specific matings, to be later incorporated into a comprehensive Intelligent Decision Support System (IDSS). In earlier studies, neural networks had been considered to generate knowledge and provide input to IDSS (Wade and Lacroix, 1994). These are undoubtedly useful approaches; DM techniques, however, can also be used for the investigation of all possible associations among different variables and for the extraction of information from very large databases. Thus, DM may provide the basis for a broadly generic, dynamic, flexible and easily used framework for data analysis.

The overriding goal of this project was twofold: I) to investigate the possibility of employing data-mining techniques for the analysis and quality control of national genetic evaluation results that form the basis for international genetic comparisons of dairy bulls and II) to determine whether data-mining application on national genetic evaluation data could lead to useful knowledge discovery in bull evaluations. More specifically, the objectives of this preliminary study were to a) develop a platform for identifying patterns/trends in routine genetic evaluation

data, discovering potential error patterns and isolating possible error causes, b) define a methodology for comparing/predicting consecutive data models and c) compare this new data quality control framework to the existing statistical method in use.

Material and Methods

Data description

National genetic evaluation results (file 010) for production traits (milk, fat, protein) of 17 routine national genetic evaluations computed between February 1999 and February 2003 in 9 countries were used. One of the datasets submitted from a country contained known errors that had been detected by the current Interbull procedure. This country also submitted an official dataset, without known errors, that was included in the analysis. A separate analysis was performed using the dataset with known errors, to test the error detecting capacity of the data-mining algorithm.

Variables considered and data pre-processing

The estimated genetic merit (proof) of every bull, as expressed in each country, was the dependent (response) variable. Preliminary tests revealed interesting correlations for the following four variables, which were subsequently included in the algorithm training set:

1. Birth year of the bull (35 birth years identified).
2. Type of proof of each bull in each country (11=first crop daughters, 12=first and second crop daughters, 21=imports).
3. Population of origin, determined from the breed and country code in the bull's international registration number (21 populations identified).
4. Number of daughters per bull and country of evaluation.

Birth year, type of proof and population of origin were discrete variables whereas bull proof and number of daughters were continuous variables. The last two were categorized, to facilitate data analysis with DM algorithms. Abbass *et al.* (1999) concluded that in DM applications aiming at supporting IDSS, it is easier and more accurate to use category labels (discrete variables) than numeric values (continuous variables). Bull proofs were categorized in two ways: a) in 10 equally sized classes based on the minimum and maximum value (min-max transformation) and b) in 6 classes to cover the entire distribution (±3 standard deviations) using the z-score transformation. Number of daughters was transformed in two ways: a) similar to (b) for bull proofs and b) by computing single-trait daughter-based reliabilities (range: 1-99).

Data-mining system

Data were stored in a relational database created with the MicroSoft (MS) SQL server 2000. The MS Analysis Manager was used for data-mining. It must be noted here that the database and DM tool choices are not restrictive. Any relational database or even a text file could be used for storing and maintaining data. Data-mining algorithms were also run using the Waikato University Environment for Knowledge Analysis (WEKA) tool. Following data collection, pre-processing and transformation, national genetic evaluation results were analyzed mainly with classification algorithms. The induced models were closely studied and evaluated. Results were used to discover systematic or non-systematic error patterns in the data and to develop means for evaluating and comparing the induced models. Figure 1 illustrates an envisaged complete data-mining system. So far, however, only components in the vertical rectangle have been developed.

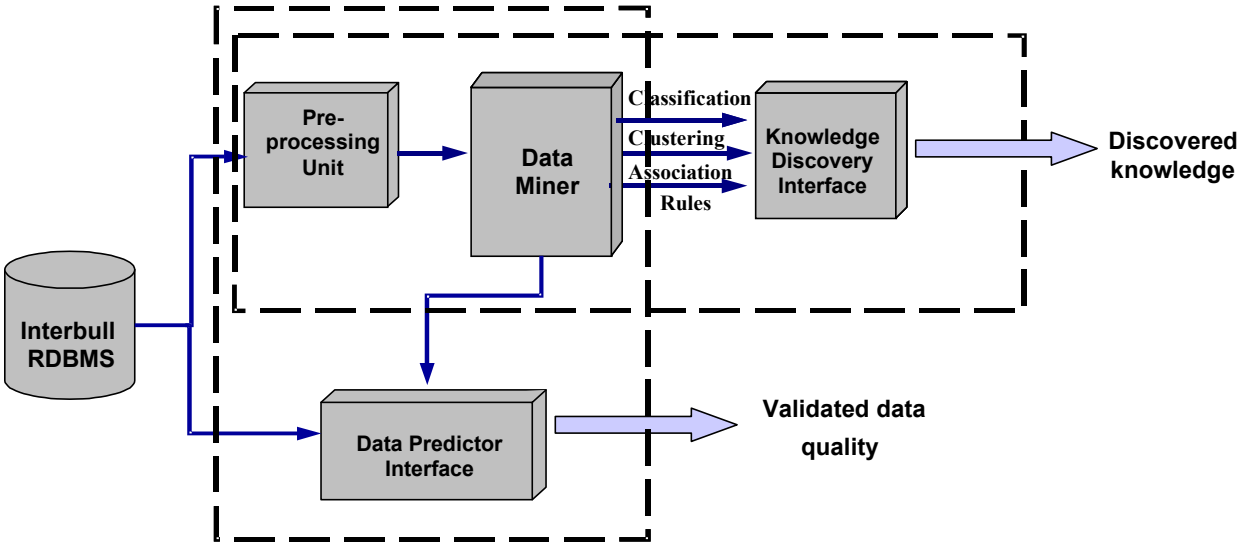

Figure 1. Envisaged data-mining system after complete development (RDBMS = Relational Database Management System).

Data analysis

A classification algorithm based on the C4.5 decision-tree classifier was used. The fact that classification trees are easily applicable and comprehensive and do not have scalability problems made C4.5 our preferred choice. After bull proofs had been categorized, a decision-tree model was induced based on the associations discovered between the response variable (bull proof) and the four input variables described above. The strength of the association was assessed qualitatively and in relative terms for the four variables. The algorithm was applied sequentially to data from each country and each evaluation run, until a model pattern started to emerge for each country. This model pattern was used to assess the consistency of associations across time (i.e. between the different evaluation runs in each country separately) and also to determine a way for deriving predictions for future bull proof models in each country.

Two different datasets were considered. One included only bulls with genetic evaluation available in all runs (number of bulls ranged from 1320 to 26,046 in each of the 9 countries). This was meant to provide evidence of consistency across time; associations between the four variables and proof were expected to be similar in consecutive evaluation runs. The other dataset included all bulls. In this case, association differences might not reflect departures from consistency but changes in breeding philosophy. For example, a new batch of import bulls (type 21) might affect the association of bull proof with type and/or population of origin in a country.

A specific data-mining model was extracted for each country and a Java interface was implemented in order to test and evaluate the algorithm-induced models. This interface provides interoperability between relational databases and text files, on one hand, and the WEKA data-mining tool, one the other, in order to evaluate the C4.5 tree's prediction accuracy and to produce the corresponding confusion matrices. The latter list the actual against the predicted classifications. Correct predictions fall on the diagonals and misclassifications on the off-diagonal of a confusion matrix. In addition, the interface is particularly user-friendly as it may accommodate input data either as text files or relational database tables.

Predictions derived from the selected data-mining model were compared to actual bull proof distribution. In the first instance, this applied to two countries: a) the country with known errors in one of its datasets and b) in the country with the most consistent associations across time. In (a), data-mining models were extracted from evaluation data prior to the "erroneous" run. Predictions were compared to actual bull proofs from the subsequent run

including both the erroneous and the official (without known errors) dataset. The marginal probability of each tree node and the distribution of classified instances in the model confusion matrix were checked, providing both quantitative measures and qualitative assessment of the induced model's predictive capacity. The combination of these two methods (model evaluation and model comparison) has provided us with the necessary power to detect and isolate different error types in the test datasets.

In this first application, all work tasks and routines concerning data pre-processing, prediction queries, model comparison and prediction result extraction, used the Data Transformation Services (DTS) tool provided by MS SQL Server 2000. Nevertheless, they can be easily implemented in Java, if other relational databases or text files are used.

Results and Discussion

Data pre-processing

The four different categorization approaches applied to our continuous variables (bull proof and number of daughters) were tested in two of the nine countries. Bull proofs were categorized using either the min-max transformation or the z-score transformation method. Number of daughters was categorized using either the z-score transformation or the reliability-calculation approach. Analyses revealed only very minor differences between the decision trees in each case. Associations between bull proofs and the four input variables were the same regardless of transformation. Data from all nine countries were then analyzed using min-max transformation for bull proof and z-score normalization for number of daughters.

Data-mining application

In all cases, birth year of the bull had the strongest association with bull proof. This is probably expected because of the genetic progress achieved in each country, following bull selection. Type of proof and population of origin of bull had the second strongest association in three and four cases, respectively, whilst number of daughters had generally the weakest association with bull proof. This may be indicative of different breeding programs. In traditionally importing countries, for example, type of proof and/or population of origin had strong associations with bull proof, especially in early birth years.

The models were derived separately for each country and evaluation run. In the analysis of bulls with genetic evaluation in all 17 runs, we were expecting similar models in every run for each country. Indeed, the induced models and decision trees were generally consistent across evaluation run all countries, meaning that the relative strength of association remained the same for all input variables. The confusion matrix for each model indicated considerable across-time stability for each country.

Figure 2. Data-mining output (decision tree) from the analysis of official and erroneous milk evaluation data from country X; input variables are birth year of bull (Byr), number of daughters z-transformed (Ndau) and type of proof.

When data with known errors from one country (country X) were included, however, a distinct change in the decision tree model and the confusion matrix was observed. Figure 2 shows the decision trees of the evaluation with the official and the erroneous data. Decision trees appear to be distinctly different. These changes were, in fact, due to shifting emphasis to import (type 21) bulls. In all previous official datasets of country X, type of proof had the weakest, if any, association with the predicted class, whereas in the "erroneous" dataset it had the second strongest association with bull proof. In fact, decision trees derived from all previous evaluations of this country were very similar to the first decision tree in figure 2.

A closer inspection of the distribution of categorized bull proofs by type of proof in country X showed that the pattern changed dramatically when data included known errors for type 21 but not for type 11 and 12 proofs (figure 3). Hence, in this example, data quality became an issue with regards to the way import bulls were evaluated in the particular country.

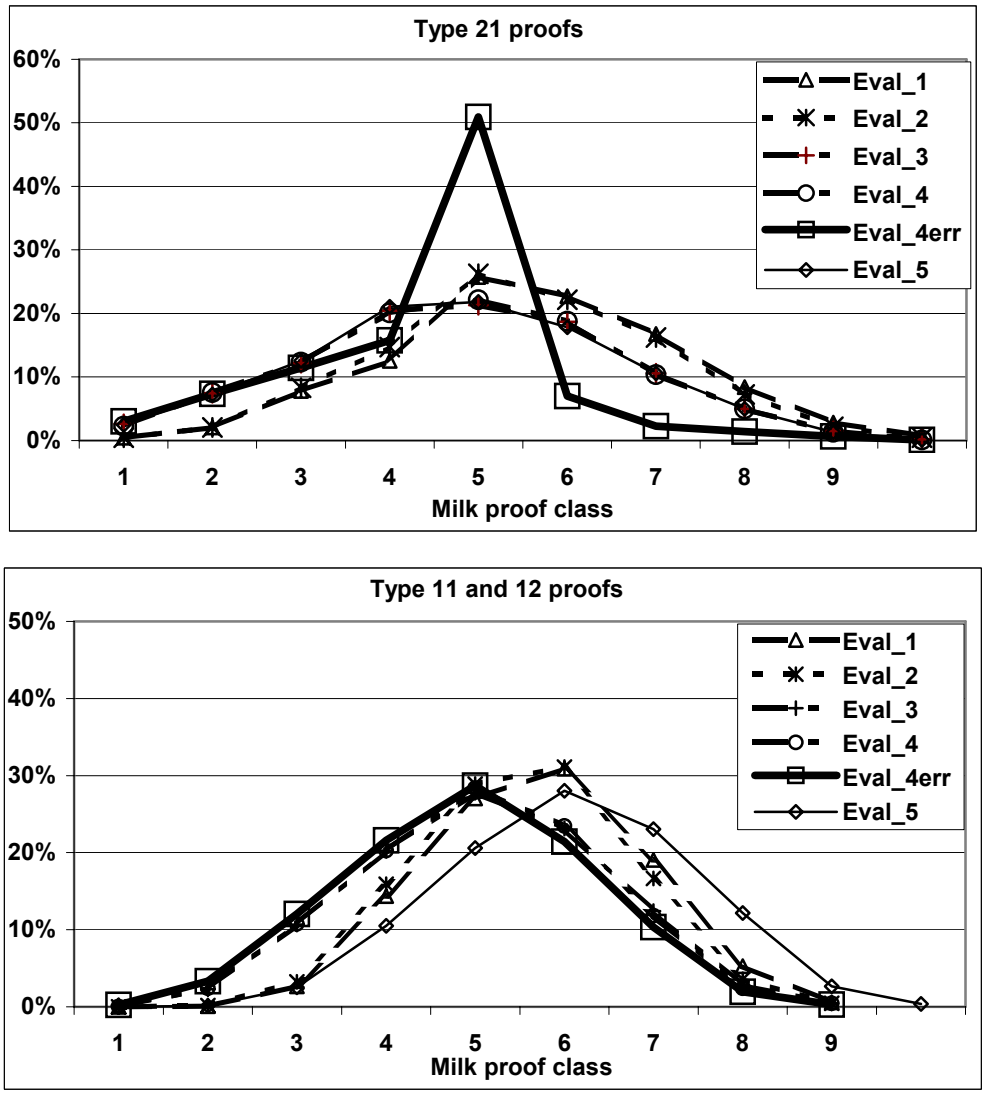

Figure 3. Milk proof distribution by type of proof in five consecutive evaluation runs in country X; Eval_4err includes data with known errors.

In this instance, the data-mining procedure picked an error that had been already identified by the current Interbull procedure. Thus, two distinctly different approaches gave similar results. This is a re-assuring observation in either case.

In order to evaluate and ultimately predict decision-tree models, the implemented architecture compares the node probabilities of each model and decides on the similarity of the results. In these preliminary results, prediction accuracy, calculated from the confusion matrices mostly ranged from 0.40 to 0.60 (note that more than 80% of misclassifications were by 1 class only), suggesting the need for fine-tuning and refining of the procedure. Identifying and including additional input variables that may be associated with bull proofs will most likely improve the accuracy. However, the key utility of this platform lays in its capacity to identify the exact node where disruptions occur leading to erroneous data.

Summary and Conclusions

In this study a new approach to analyzing animal genetic evaluation data was introduced. The method uses algorithms that mine data for links, patterns and predictive clues and, so far, has identified useful associations between bull proofs and a range of attributes, such as type of proof, birth year, population of origin of bull and number of daughters. In this first step the following has been achieved:

1. Data-mining algorithms were applied and models were induced that help us understand the data. These algorithms may become the base of easily usable front-end tools for analysts without previous DM experience.
2. Consistent model patterns (associations) were revealed and identified in most cases.
3. Error patterns in a dataset with known errors were identified. The erroneous dataset was included in the analysis in order to confirm the approach's correctness. This may be potentially useful in confirming and/or complementing the current Interbull procedure.
4. A tool for checking new models has been developed. This model integrates MS SQL Server, MS Analysis Manager and evaluation information in one stand-alone, functional Java application.

Further work on DM application to data quality control may include:

1. Refining the technique for model inspection.
2. Developing new algorithms to study the trends of identified errors.
3. Analyzing and deciding upon criteria that may be used to determine the status of data quality based on comparisons between predictions and actual proofs of the same bulls.
4. Possibly looking at other data-mining techniques (regression analysis, trend analysis)
5. Developing systems for sequential mining that considers all historic information.

The ultimate goal of data-mining is knowledge discovery. In this context, future analysis of genetic evaluation results could be searching for hidden patterns and information. In addition to the four input variable used in this study, additional variables describing the data might be needed. Intelligent Decision Support Systems could use this information to assist selection and the development of breeding strategies.

Acknowledgements

The participation of nine countries that consented to the use of their national evaluation data is acknowledged.

References

Abbass, H.A., Bligh, W., Towsey, M., Flinn, G. & Tierney, M. 1999. Knowledge discovery in a dairy cattle database: Automated knowledge acquisition. *Proc. 1999 Meeting of the International Society for Decision Support Systems,* 13 pages.

Han, J.W. & Kamber, M. 2000. *Data-mining: Concepts and techniques*. Morgan Kaufmann.

Klei, L., Mark, T., Fikse, F. & Lawlor, T. 2002. A method for verifying genetic evaluation results. *Proc. 2002 Interbull Meeting, 29*, 178-182.

Täubert, H., Swalve, H.H. & Simianer, H. 2002. The Interbull audit project Part II: Development of a program for auditing breeding value estimation programs. *Proc. 2002 Interbull Meeting, 29*, 165-167.

Wade, K.M. & Lacroix, R. 1994. The role of artificial neural networks in animal breeding. *Proc. 5th WCGALP, 22,* 31-34.