Available online at www.sciencedirect.com

SCIENCE DIRECT®

Computers and Electronics in Agriculture 52 (2006) 21–35

Computers and electronics in agriculture

www.elsevier.com/locate/compag

A decision-tree-based alarming system for the validation of national genetic evaluations

S. Diplaris[a,*], A.L. Symeonidis[a], P.A. Mitkas[a], G. Banos[b], Z. Abas[c]

[a] Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece
[b] Department of Animal Production, School of Veterinary Medicine, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece
[c] Department of Agricultural Development, Democretus University of Thrace, Orestiada, Greece

Received 28 May 2005; received in revised form 29 November 2005; accepted 9 January 2006

Abstract

The aim of this work was to explore possibilities to build an alarming system based on the results of the application of data mining (DM) techniques in genetic evaluations of dairy cattle, in order to assess and assure data quality. The technique used combined data mining using classification and decision-tree algorithms, Gaussian binned fitting functions, and hypothesis tests. Data were quarterly national genetic evaluations, computed between February 1999 and February 2003 in nine countries. Each evaluation run included 73,000–90,000 bull records complete with their genetic values and evaluation information. Milk production traits were considered. Data mining algorithms were applied separately for each country and evaluation run to search for associations across several dimensions, including bull origin, type of proof, age of bull, and number of daughters. Then, data in each node were fitted to the Gaussian function and the quality of the fit was measured, thus providing a measure of the quality of data. In order to evaluate and ultimately predict decision-tree models, the implemented architecture can compare the node probabilities between two models and decide on their similarity, using hypothesis tests for the standard deviation of their distribution. The key utility of this technique lays in its capacity to identify the exact node where anomalies occur, and to fire a focused alarm pointing to erroneous data.

Keywords: Data mining; Dairy cattle evaluations; Alarming technique; Genetic evaluations; Quality control

1. Introduction

Data quality constitutes one of the most critical issues in genetic evaluations at both national and international levels. International genetic evaluations computed by the International Bull Evaluation Service (Interbull, 2004) 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 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 data sets 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, previ-

[*] Corresponding author. Tel.: +30 2310 996349; fax: +30 2310 996398.
E-mail address: diplaris@danae.ee.auth.gr (S. Diplaris).

0168-1699/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.compag.2006.01.001

ously unknown and potentially useful) information or patterns from data in large databases (Fayyad et al., 1996). 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 makes 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 the field of agriculture, DM applications have emerged in the recent years. Bertis et al. (2001) have applied DM techniques in cornfields, whereas Cunningham and Holmes (2001) showed the efficiency of the application of known DM techniques in a mushroom grading problem. They made use of the model introduced by Garner et al. (1995), by which they induced domain information in a software module by iteratively achieving interaction between the data provider and the DM expert. Other activities in this area involve, among others, the use of DM in textiles (Scherte, 2002), trend discovery in pesticide abuse (Abdullah et al., 2004), and weed recognition in cornfields using artificial neural networks (Yang et al., 2002). Harms et al. (2001) have integrated spatio-temporal knowledge discovery applications into a Geo-Spatial Decision Support System (GDSS), in order to find relationships between specific climatic episodes and other climatic events, and finally predict them. Pietersma et al. (2003) induced decision trees for the classification of services per conception in dairy cattle. Another example that shows the efficiency of knowledge discovery in agriculture is the classic paper by Michalsky and Chilausky (1980), which induced a set of rules for diagnosing soybean diseases that was completely different than the expert suggestions, but was so accurate that it led some experts to reject their own procedures and adopt these rules. 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) (Macrossan et al., 1999; Abbass, 2002). In earlier studies, neural networks had been considered to generate knowledge and provide input to IDSS (Wade and Lacroix, 1994). Both approaches are undoubtedly useful; 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 easy to use framework for data analysis.

In a previous study, a new approach to analyzing animal genetic evaluation data was introduced (Banos et al., 2003). The method used algorithms that mine data for links, patterns and predictive clues, and identified useful associations between bull proofs and a range of attributes, such as type of proof, birth year and population of origin of bull, and number of daughters. As a result, data mining models were induced that helped understanding the data and consistent model patterns (associations) were revealed and identified in most cases. Error patterns in a data set with known errors were identified by subjective inspection of the models, but a generalized method to alert for inconsistent data was missing.

A more advanced step is the evaluation of the DM application results with objective criteria leading, when necessary, to the automatic issuing of warnings or alarm signals. After having successfully employed 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, the technique presented here was able to determine whether data mining application on national genetic evaluation can lead to useful knowledge discovery in bull evaluation. More specifically, the objective of this study was to identify patterns/trends in routine evaluation data, discover potential error patterns and isolate possible error causes. The study also defines a methodology for comparing consecutive data mining models.

2. Material and methods

2.1. Data description and pre-processing

National genetic evaluation results (file 010) for production traits (milk, fat, and protein) of 17 routine national genetic evaluations computed between February 1999 and February 2003 in nine countries were used. One of the data sets contained known errors that had been detected by the current Interbull procedure. The country that submitted the erroneous data set later submitted another error-free data set, which was also included in the analysis. A separate analysis was performed on the data set with known errors, to test the error detecting capacity of the data mining algorithm and the alarming technique.

The estimated genetic merit (proof) of every bull, as expressed in each country, was the dependent (predicted) 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, and 21 = imports).
(3) Population of origin, determined by 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 and produce more comprehensive models. 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. If μ is the mean and σ is the standard deviation of a bull proof distribution, the normalized value $v'$of a bull proof value v after the *z*-score transformation is $v' = v - \mu/\sigma$

The number of daughters variable, N_{dau}, 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 were stored in a relational database created with the MicroSoft (MS) SQL server 2000. The MS Analysis Manager was used for data mining. Following data collection, pre-processing and transformation, national genetic evaluation results were analyzed mainly with classification algorithms. The induced models were used to discover systematic or non-systematic error patterns in the data and to develop the alarming technique for evaluating and comparing the induced models.

2.2. Data mining module

A decision-tree is a flow-chart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions. For example, in Fig. 1, the outcome of the test concerning the bull birth year is first considered. Then the test for bulls born between 1979 and 1990 proceeds by splitting according to the number of daughters of each bull. Each node of the tree contains bull proof value distributions. A classification algorithm-based on the C4.5 decision-tree classifier was used (Quinlan, 1993). C4.5 is a criterion gain algorithm, i.e. an algorithm that decides on the construction of the decision-tree, according to the minimization of a certain criterion. This criterion is the information gain (Quinlan, 1993). The C4.5 decision-tree classifier was selected because classification trees are easily applicable and comprehensive. It also has proved efficient from a computational complexity standpoint, since it is a polynomial time algorithm.

After bull proofs were categorized, a decision-tree model was induced based on the associations discovered between the class variable (bull proof) and the four input variables described earlier. 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

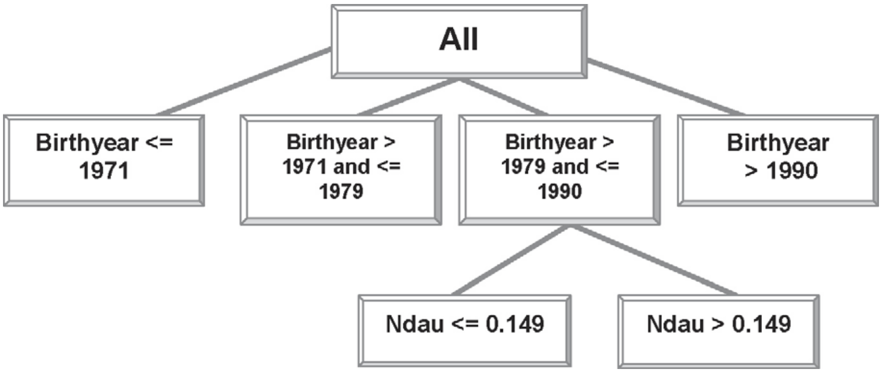

Fig. 1. Decision-tree induced for evaluation run D3.

then used (a) to assess the consistency of associations through time (i.e. between the different evaluation runs in each country separately and (b) to provide input to the alarming system.

The data set included only bulls with genetic evaluation available in all runs (number of bulls ranged from 1320 to 26,046 in each of the nine countries). This grouping 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. A specific data mining model was extracted for each country.

Predictions derived from the selected data mining model were compared to actual bull proof distribution. The alarming technique was applied to the country with known errors in one of its data sets. Data mining models were extracted from evaluation data prior to the “erroneous” run. Using the alarming technique, predictions were compared to actual bull proofs from all the rest of the runs including both the erroneous and the error-free data sets.

2.3. Examining categorized bull proofs

The bull proof scores in the decision trees induced from the data mining procedure were expected to follow normal (Gaussian) distributions, since these scores were normalized during the pre-processing stage. Consequently, if the corresponding dataset was not erroneous, bull proof data in each node of the decision-tree were expected to similarly follow the Gaussian distribution. Therefore, a tool to measure the goodness of fit to the Gaussian distribution would be useful, in order to decide whether or not a decision-tree node includes erroneous data. It should be noted here that in this study, white (Gaussian) noise was not addressed as a type of error and it cannot be identified by this method. Nevertheless, such errors are not considered very significant for the analyst in comparison to any abnormal insertions of bull proofs that had been made into the evaluation system, or the discovery of possible flaws of the existing evaluation model.

Since the data values in each node were categorized (with values 1–10), a binned fitter was applied to fit the data in each decision-tree node to the Gaussian distribution. The fitter exploits the chi^2 technique by fitting histogram values to the Gaussian line. A brief description of the applied technique follows:

Given a linear function *f* and a binned distribution the following variables can be defined:

h_i is the number of bull proofs (height) of the *i*th bin (category),
x_i is the categorized bull proof corresponding to the *i*th bin,
$f(x_i)$ is the function’s value at the *i*th bin,
σ_i the error on the *i*th bin.

One can now define the following fit method where the sum is performed over all the bins:

$$\chi^2 = \sum_i \frac{[f(x_i) - h_i]^2}{\sigma_i^2} \tag{1}$$

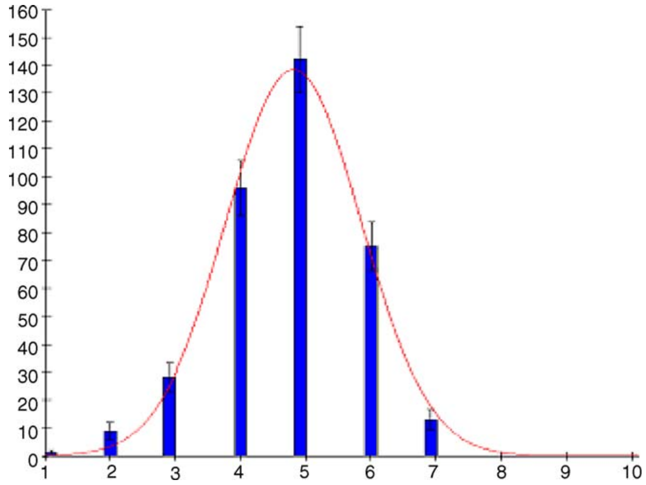

Fig. 2. Gaussian fit for the histogram induced from node distribution 1.2 from evaluation run D3.

By minimizing the quantity in (1), the optimum estimates for mean and standard deviation that define the Gaussian which best fits the data distribution are calculated. Minimizing χ^2 involves the minimization of the numerator in (1), i.e. the standard error of a Gaussian model, so that the distance from the data error (described in the denominator) is minimized. The Gaussian function that is finally computed yields the model that most closely fits the data. Fig. 2 illustrates a binned histogram fit to the Gaussian distribution.

The value of χ^2 was a good measure for the goodness of the performed fit; nevertheless, it was dependent on the number of data entries in each distribution. The more points there were, the harder it would be to get a Gaussian line by chance, unless the data really suggested such a Gaussian distribution. Thus, the value of χ^2 had to be normalized. The key concept here is the number of "degrees of freedom (d.f.)". It is defined as the number of independent data points, N_{data}, minus the number of fitting parameters, n_{param}.

$$\text{d.f.} = N_{\text{data}} - n_{\text{param}} \tag{2}$$

In order to fit a sum of Gaussian distributions into the data, the fitting parameters are two (mean and standard deviation) times the number of Gaussian models. In the case under consideration, it was desirable to fit only one Gaussian model in each node distribution. Therefore, the fitting parameters were always two. Thus, d.f. = $N_{\text{data}}-2$. One could now define a quality of fit measure (criterion) as the normalized value of χ^2 divided by d.f.

$$\text{Quality} = \frac{\chi^2}{\text{d.f.}} \tag{3}$$

This quality of fit criterion was calculated separately for each node of the induced decision trees. By fitting the bull proof values distribution in each node to the Gaussian distribution, a measure of the correctness of data contained in the node could be derived. When results truly followed Gaussian distribution, this value was expected to be no greater than one. In case of significant deviations, a warning was issued, indicating possible erroneous distribution of bull proofs. The binned fit method has been developed as a Java application, which uses the MINUIT fit optimizer (James and Roos, 1989).

2.4. F-tests on node variance in consecutive evaluation runs

Using the chi^2 technique the node distributions in each decision-tree model were individually compared to the corresponding expected Gaussian distribution. A more powerful measure in order to detect anomalies would be to compare corresponding node distributions from different models (evaluation runs) against each other. Since the application of data mining indicated that there was a pattern in the structure of the decision trees induced, it was possible to compare corresponding node distributions in decision-tree models induced from different evaluation runs. It should be noted that there were no differences in national genetic evaluation models across runs and no new bulls were allowed into the system (Banos et al., 2003). Although the mean proof value in corresponding nodes was not expected to be similar, since a bull's proof value may vary from run to run, the variance of the proof values in corresponding nodes was expected to remain stable, at least during subsequent evaluation runs. Consequently, there was a need to compare variances in data distributions in corresponding nodes, in order to discover possible mismatches that could imply the firing of an alarm.

A common statistical technique to examine whether there are any differences between two datasets is the analysis of variance and the associated *F*-test. The *F*-test involves the determination of differences between two variations (versions) of the same dataset. In this study, the *F*-test was applied on two different datasets (subsequent evaluation runs), but it was expected that their variances in corresponding node distributions would remain the same.

The *F*-distribution is formed by the ratio of two independent chi^2 variables divided by their respective degrees of freedom.

$$F = \frac{(\text{d.f.}_1\, s_1^2/\sigma_1^2)/\text{d.f.}_1}{(\text{d.f.}_2\, s_2^2/\sigma_2^2)/\text{d.f.}_2} \tag{4}$$

In Eq. (4), s^2 is the sample variance and σ^2 is the estimate of the true variance of each node in evaluation runs one and two. The *F*-test was designed to test if two population variances are equal (Triola, 2003). It does this by comparing the ratio of two variances. So, if the variances are equal, the ratio of the variances will be one (null hypothesis). All hypothesis testing is done under the assumption the null hypothesis is true. If the null hypothesis is true, then the *F*-test indicates that this ratio of sample variances implies that the two distributions have equal variances. If the null hypothesis is false, then the null hypothesis that the ratio was equal to one is rejected, as well as the initial assumption that the variances were equal. Thus, the difference between the two corresponding nodes is a true difference at the declared level of significance, or *P*-value. A typical threshold for the critical value under which the null hypothesis is true is $P < 0.05$. This *P*-value was also used for the experiments conducted in this study.

Rejecting the null hypothesis for a particular node would imply that the node was "suspicious" in this respect. This criterion was fitted to decision-tree nodes of consecutive evaluation runs. When an exact corresponding node was missing in the following run, then comparisons were made with the closest "parent" node.

2.5. *The alarm firing system*

In order to develop a robust alarming method, the two above-mentioned testing methods (chi^2 for model evaluation and *F*-test for model comparison) were combined. The combination could provide the system with the necessary power to detect and isolate errors in the test datasets. The technique developed threw two levels of alarm; the "yellow" warning and the "red" alarm.

First, each model was evaluated individually. Its node distributions were investigated in order to discover possible deviations from the Gaussian distribution, using the binned fit method. Nodes where the quality measure of the Gaussian fit exceeded the predetermined threshold were labeled as "suspicious" nodes.

Then, the models were evaluated by comparison. Corresponding node distributions from consecutive evaluation runs were compared against each other using the *F*-test. Since older evaluations were considered to be accurate, in cases where the *F*-test indicated that the two distributions had unequal variances, the node that belonged to the most recent evaluation run was marked as a "suspicious" node.

Finally, the technique combined the two evaluating methods by traversing each node in every model. If a node failed one of the two tests (i.e. either the chi^2 or the *F*-test with the corresponding node of the subsequent run) then a yellow warning was issued. If a node failed both tests, then a red alarm was fired for the specific node.

Fig. 3 illustrates the two-level data mining approach developed in this study.

3. Results

For experimental purposes, decision-tree models were first induced from five data sets (D1, D2, D3, D4, and D5) corresponding to five consecutive official genetic eva1uation runs in three countries. Bull proofs for fat yield were considered. In addition, a knowingly erroneous data set (D4_e) had been obtained from one of the countries.

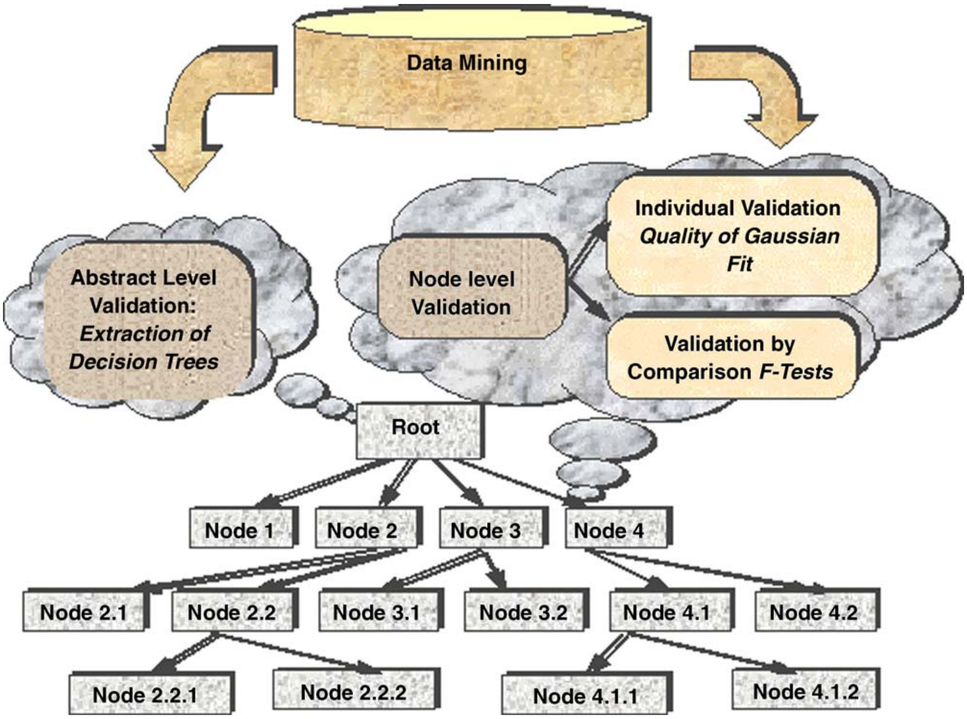

Fig. 3. The two-level approach of data mining application in genetic evaluations. The evaluation models extracted by the data mining procedure are validated both in abstract and in node level. Node level validation involves individual node validation and validation by comparison.

In the analysis of bulls with genetic evaluation in all runs, one would expect similar models in every run. Indeed, by inspection, the non-erroneous models were generally consistent across evaluation runs, meaning that the relative strength of association remained the same for all input variables. For the erroneous evaluation run, however, a distinct change in the decision-tree model was observed. Figs. 1, 4 and 5 illustrate the decision trees for the evaluation runs D1, D4_e, and D5 for Country 1. By close examination of these figures, one can observe that decisions trees are quite similar when official data were used (Figs. 3 and 5), whereas the decision-tree induced from the erroneous data becomes distinctly different (Fig. 4). The alarming technique that was then applied indicated the nature of the erroneous data by measure.

In order to measure the quality of the node distributions in each model, the goodness of fit test (chi^2) was applied. The threshold for the quality measure was set at the $P < 0.05$ significance level. In Figs. 6–10, the fit to the Gaussian

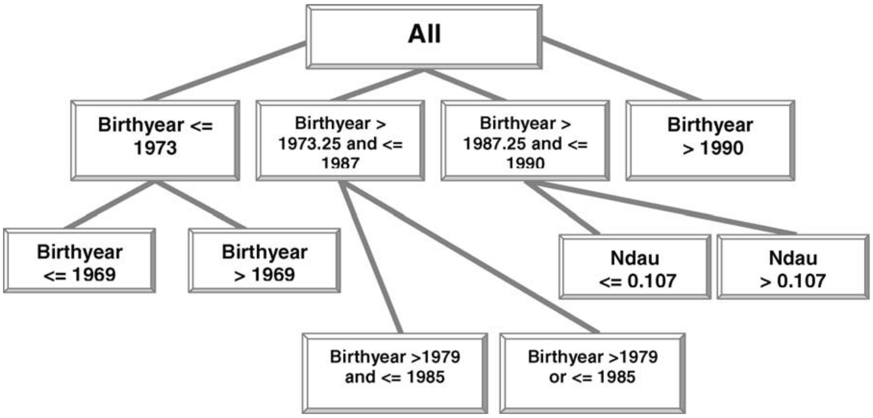
All
Birthyear <= 1973
Birthyear > 1973.25 and <= 1987
Birthyear > 1987.25 and <= 1990
Birthyear > 1990
Birthyear <= 1969
Birthyear > 1969
Ndau <= 0.107
Ndau > 0.107
Birthyear >1979 and <= 1985
Birthyear >1979 or <= 1985

Fig. 4. Decision-tree induced for the erroneous evaluation run D4_e.

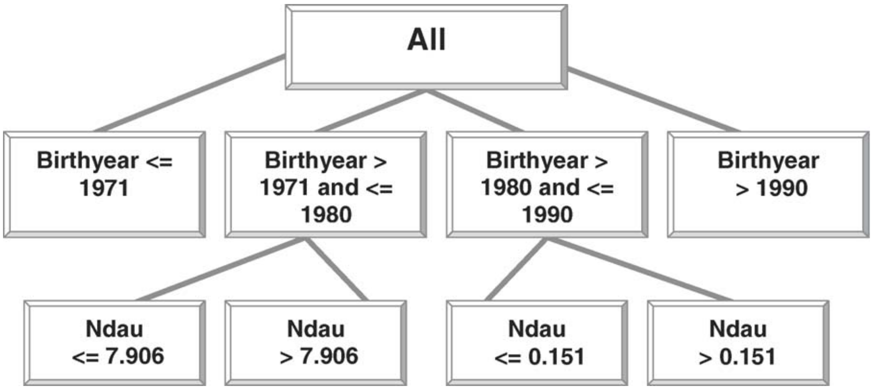

Fig. 5. Decision-tree induced for the evaluation run D5.

distribution for corresponding nodes from two subsequent models for Country 1 is illustrated. The first model is the erroneous one (D4_e), while the second model comes from official data for the next evaluation run. It can be observed that in the erroneous D4_e run the departure from normality is very clear in the nodes 2.1 and 3.2 (Figs. 6–8), while in run D5, data in the corresponding nodes seem to follow the Gaussian distribution (Figs. 9 and 10). Fig. 11 illustrates the measurements of the quality of fit test (chi^2). The dotted horizontal line represents the threshold value, above which there was significant ($P > 0.05$) deviation from the expectation and the node was considered to have failed the test. In D4_e, the values obtained for three nodes (1.2, 1.2.1, and 1.3.2) clearly exceed the others and the threshold. In the case of node 1.2.1, actually, this value extends to infinity. These nodes were associated with bulls born between 1979 and 1985 (node 1.2.1) or bulls born between 1987 and 1990 with $N_{\mathrm{dau}} \geq 0.107$, i.e. having more than 169 daughters (node 1.3.2). In such cases, a yellow warning was issued.

In the other cases, chi^2 values were fairly consistent across all nodes, except for data set D1, node 1.1 (associated with bulls born before 1971), where genetic evaluation results may warrant closer inspection. The same tests were applied to two other countries without known errors in any of their data sets. In all cases, the calculated chi^2 value remained below the threshold (Figs. 12 and 13).

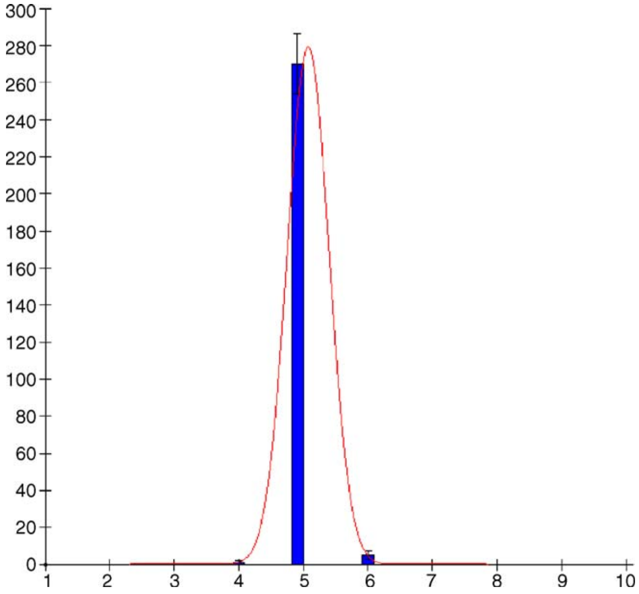

Fig. 6. Gaussian fit for the histogram induced from node distribution 1.2.1 from the erroneous evaluation run D4_e.

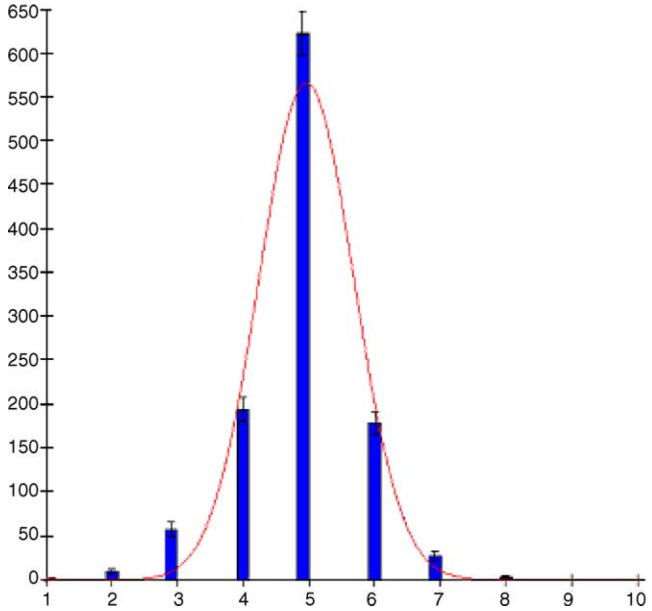

Fig. 7. Gaussian fit for the histogram induced from node distribution 1.2 from the erroneous evaluation run D4_e.

In order to apply the *F*-test method between corresponding nodes of the models, node correspondence should first be established. In most cases this task was trivial, since the structure of the trees was almost the same and, in all the models, birth year of the bull had the strongest association with bull proof. Consider two decision trees T_1 and T_2, which were induced from subsequent evaluation runs, and where the children nodes of a parent node were sorted according to the attribute by which they were split. Without loss of generality, T_1 was considered as the reference tree (i.e. the tree induced at an earlier run and assumed accurate). Two nodes $N_1(i) \in T_1$ and $N_2(i) \in T_2$ were considered corresponding nodes if they lied at the same tree level, their parents were corresponding nodes, and they were similarly categorized, e.g. they were both the *m*th children of their parents. In cases where a corresponding node $N_2(i)$ did not

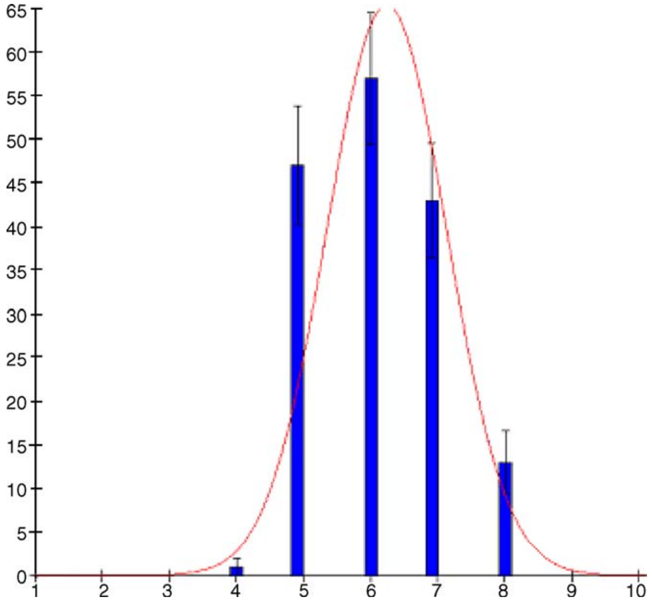

Fig. 8. Gaussian fit for the histogram induced from node distribution 3.2 from the erroneous evaluation run D4_e.

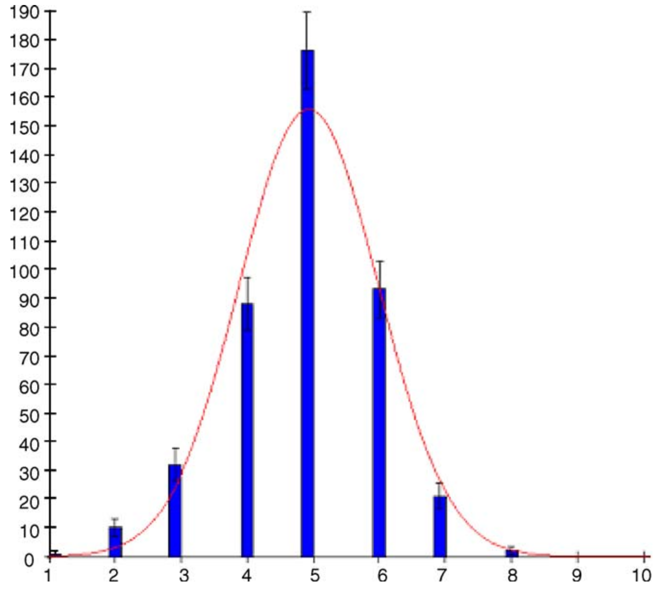

Fig. 9. Gaussian fit for the histogram induced from node distribution 1.2 from evaluation run D5.

exist in the decision-tree T_2, the F-test was administered between N_1(i) and the corresponding node of its parent in T_2. In this way the exact node where an error might occur could be identified, even when a corresponding node in the other model did not exist.

When the F-test was applied to node-by-node comparison of consecutive evaluation runs, four yellow warnings were issued in the case of D4_e. Two of these warnings were associated with nodes which already had yellow warnings issued by the chi^2 test. These two yellow warnings were upgraded to red alarms. Some additional yellow warnings were issued to neighboring nodes that were affected by distribution disruptions in the “red” nodes. Interestingly, a yellow warning was also issued in one case of consecutive data sets without known problems (D4–D5). The node affected

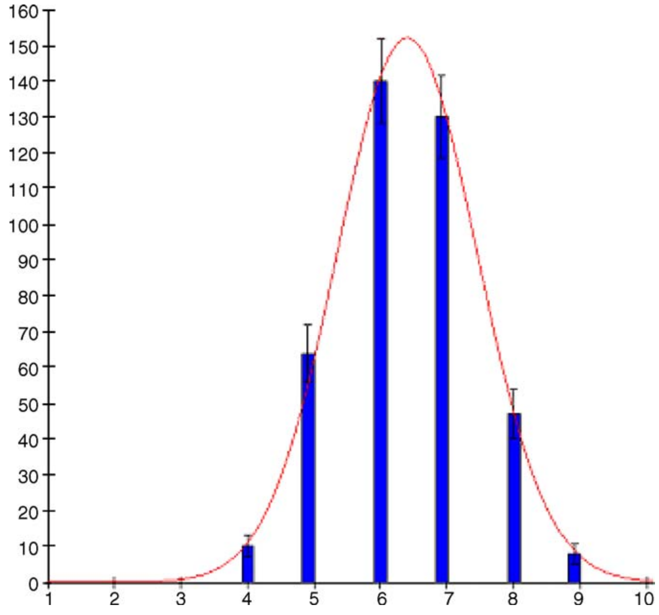

Fig. 10. Gaussian fit for the histogram induced from node distribution 1.3.2 from evaluation run D5.

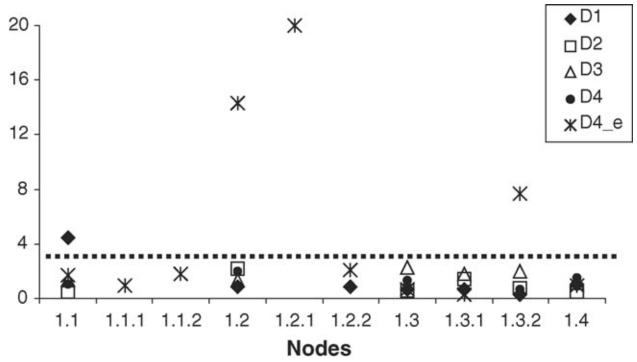

Fig. 11. Quality of fit (chi^2) test for four data sets (runs) without known errors (D1–D4) and one with known errors (D4_e) in Country 1; values exceeding the dotted horizontal line indicate departure from normality.

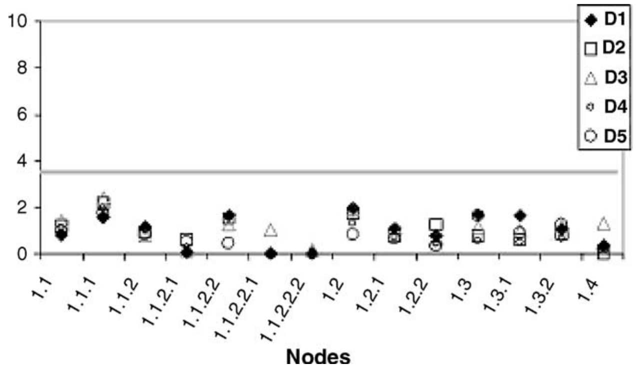

Fig. 12. Quality of fit (chi^2) test for five data sets (runs) without known errors (D1–D5) in Country 2; the horizontal line indicates the threshold above which there is departure from normality.

pertained to bulls born between 1971 and 1980 and with $N_{\text{dau}} \geq 7.906$ (i.e. having more than 131 daughters). The same node had successfully passed the chi^2 test. It should be also noted that node 1.1 in D1, where a yellow warning was issued by the chi^2 test, passed the F-test.

In one of the other two countries, two yellow warnings were issued in two separate comparisons of official national evaluations. One warning was for bulls born after 1986 and the other warning was for bulls in the same period having more than 958 daughters; both nodes had successfully passed the chi^2 test. No warnings were issued for the third country.

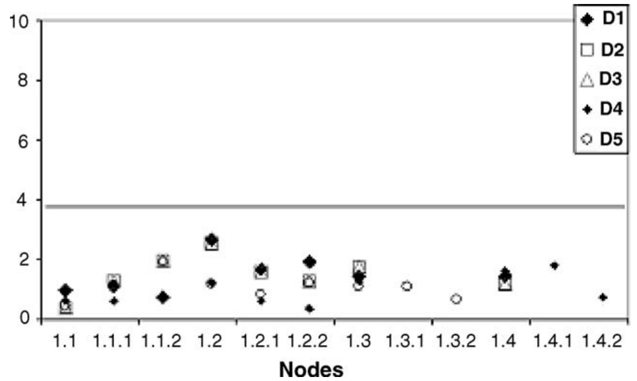

Fig. 13. Quality of fit (chi^2) test for five data sets (runs) without known errors (D1–D5) in Country 3; the horizontal line indicates the threshold above which there is departure from normality.

Table 1
Number of nodes where yellow warnings or red alarms were issued for five data sets (consecutive runs) without known errors (D1–D5), and one with known errors (D4_e) in three countries

Model comparison	Country 1		Country 2		Country 3	
	Yellow	Red	Yellow	Red	Yellow	Red
D1–D2	1	0	0	0	0	0
D2–D3	0	0	0	0	0	0
D3–D4	0	0	1	0	0	0
D3–D4_e	3	2	–	–	–	–
D4–D5	1	0	1	0	0	0

Table 2
Quality values from the quality of fit test for Country 1

Country 1 quality of fit test node	D1	D2	D3	D4	D4_e	D5	*F*-test D3–D4_e
1.1	**4.44**	0.51	1.35	1.09	1.72	1.13	
1.1.1					0.93		
1.1.2					1.81		
1.2	0.89	2.20	1.31	2.01	**14.33**	2.31	Warning
1.2.1					**20.00**	2.90	Warning
1.2.2	0.89				2.07	0.95	
1.3	0.70	0.53	2.28	1.33	0.66	0.53	Warning
1.3.1	0.64	1.47	1.81	0.66	0.25	0.87	Warning
1.3.2	0.31	0.79	1.95	0.70	**7.72**	0.09	
1.4	1.08	0.48	1.15	1.50	0.96	1.21	

Values in bold indicate departure from normality and warning issuing. The rightmost column shows yellow warnings issued by the *F*-tests in the corresponding nodes. A red alarm was issued for nodes 1.2 and 1.2.1 of data set D4_e.

Table 3
Quality values from the quality of fit test for Country 2

Country 2 quality of fit test node	D1	D2	D3	D4	D5	*F*-Test D3-D4	D4–D5
1.1	0.96	0.38	0.38	0.60	0.43		
1.1.1	1.08	1.29	1.29	0.60	1.14		
1.1.2	0.74	1.99	1.99	0.79	1.92		
1.2	2.68	2.54	2.54	1.24	1.17		
1.2.1	1.67	1.59	1.59	0.59	0.84		
1.2.2	1.93	1.25	1.25	0.35	1.24		
1.3	1.45	1.76	1.76	1.25	1.10		
1.3.1					1.10		
1.3.2					0.68		
1.4	1.46	1.17	1.17	1.62			Warning
1.4.1				1.79			
1.4.2				0.77		Warning	

Values in bold indicate departure from normality and warning issuing. The rightmost column shows yellow warnings issued by the *F*-tests in the corresponding nodes. No red alarm was issued.

Table 1 summarizes results from the combination of the chi^2 and *F*-test in five consecutive evaluation run models, in three countries (one of which included the knowingly erroneous data set). The combined results for each decision-tree node and each country are depicted in Tables 2–4, respectively.

4. Discussion

At an abstract level, the application of DM techniques could provide decision trees for the validation of consecutive genetic evaluation runs, based on model inspection. At the decision-tree node level, it was possible to assess data quality

Table 4
Quality values from the quality of fit test for Country 3

Country 3 quality of fit test node	D1	D2	D3	D4	D5
1.1	0.84	1.20	1.40	1.10	1.00
1.1.1	1.58	2.20	2.41	1.95	1.76
1.1.2	1.14	0.92	0.77	1.01	0.79
1.1.2.1	0.04	0.60	0.10	0.28	0.47
1.1.2.2	1.61	1.52	1.25	1.58	0.41
1.1.2.2.1			1.04		
1.1.2.2.2			0.14		
1.2	1.94	1.75	1.88	1.32	0.83
1.2.1	1.08	0.74	1.10	0.69	0.63
1.2.2	0.74	1.23	0.65	0.42	0.30
1,3	1.66	0.77	1.14	0.53	1.61
1.3.1	1.61	0.61	0.99	0.49	0.87
1.3.2	1.04	0.89	1.21	0.66	1.26
1.4	0.31	0.01	1.28	0.18	0.21

No alarms or warnings were issued by either test.

by measure and focus on the validation of tree nodes, using individual node analysis and also comparing corresponding nodes from consecutive evaluation runs. In this study, the individual node analysis was conducted with quality of Gaussian fit tests, while the node comparison technique involved statistical *F*-tests for the node variances.

At the individual node level, bull breeding values were tested for normality (Gaussian distribution). Sometimes, though, this assumption might not hold, especially if animals on a certain node are related to each other. In such cases the splitting criterion of entropy minimization could affect the node normality. A better variable to consider for this test would be the Mendelian sampling, computed as the difference between the breeding value of an animal and the mean breeding value of its parents. Nevertheless, this was viewed as a minor point as far as the design of the system was concerned. The latter constituted the overriding objective of the present study.

When the technique was applied to a data set with known errors, the combination of the two tests fired a red alarm for bull proofs concerning bulls born between 1979 and 1985. 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 yielded similar results. Moreover, the technique could detect by measure the exact node where the errors occured, thus further specifying the error. This was a re-assuring observation in either case. During the experiments the technique also fired “yellow” warnings, mainly when comparisons between non-consecutive evaluation runs were conducted. A “yellow” warning was also thrown for the nodes of the erroneous data adjacent to the “red” flagged node, since their distribution was affected by the erroneous disruption of the sibling node.

It should be noted that all yellow warnings and red alarms issued were for older bulls. This could be either because there were genuinely no problems with the evaluation of younger bulls in the countries studied here or because there was not enough information in the input variables and data available for the system to discover patterns within younger animal groups. More detailed data and additional input variables might be required for this matter.

The main advantage of this technique is that it exploits the ability of mining models to categorize data. The technique is also able to detect the exact node where the erroneous data exist, by producing simple but specific warnings. However, actual verification of the erroneous data can only be accomplished “manually”, by thoroughly examining the sub-dataset indicated by the model. The two-fold nature of the checks performed in the alarming method makes it more robust and illustrative for the quality of data. The technique can be easily included in the standard Interbull dairy evaluation procedure, providing an alternative way to endorse the evaluation analysis results, and possibly discover other types of errors that the standard Interbull procedure might not detect.

5. Conclusion and future outlook

In this study, a new alarm firing system has been presented, which exploits results of DM application to national genetic evaluations of dairy bulls. The system examined the discovered patterns by combining two methods for individual and pair-wise evaluation of decision-tree nodes. Each node distribution in the model was examined with regard to quality of fit to the Gaussian distribution; furthermore, its variance was compared to the corresponding node variance of the subsequent evaluation run. Results so far have shown that this system is capable of capturing errors that have been also confirmed by the standard Interbull procedure. Some additional warnings that deserve closer examination were also issued. Furthermore, the key utility of this platform lies in its capacity to pinpoint the exact node where the alarm is issued, leading to closer inspection of the potentially erroneous data and the genetic evaluation model that generated them.

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, by developing new algorithms and criteria for data quality, in order to study the trends of identified errors. In addition to the four input variables used in this study, additional variables describing the data might be needed. Finally, data mining techniques can be exploited to discover knowledge about the bull proofs directly from the data. In this case, though, there would be the need to utilize much more historic information in order to perform sequential mining or regression techniques, discover trends in bull proofs and finally predict bull proof values.

Acknowledgements

We would like to thank the Interbull Center for providing us with the necessary data to perform this research. Also, the consent of nine countries to the use of their national evaluation data is appreciated.

References

Abbass, H.A., Bligh, W., Towsey, M., Tierney, M., Finn, G., 1999. Knowledge discovery in a dairy cattle database: (automated knowledge acquisition). In: Proceedings of the Fifth International Conference, International Society for Decision Support Systems (ISDSS'99), Melbourne, Australia.

Abbass, H.A., 2002. Computational intelligence techniques for decision making: with applications to the dairy industry. Dissertation. de Verlag Publishing, Germany, ISBN 3-89825-408-9.

Abdullah, A., Brobst, S., Pervaiz, I., Umer, M., Nisar, A., 2004. Learning dynamics of pesticide abuse through data mining. In: Proceedings of the Second Workshop on Australasian Information Security, Data Mining and Web Intelligence, and Software Internationalisation, Dunedin, New Zealand 32, pp. 151–156.

Banos, G., Mitkas, P.A., Abas, Z., Symeonidis, A.L., Milis, G., Emanuelson, U., 2003. Quality control of national genetic evaluation results using data mining techniques; a progress report. In: Proceedings of the 2003 Interbull Annual Meeting, Rome Italy, 31, pp. 8–15.

Bertis, B., Johnston, W.L., Lovell, A., Olson, S., Steed, S., Cross, M., 2001. Data mining in U.S. corn fields. In: Proceedings of the First SIAM International Conference on Data Mining, Chicago, IL, USA.

Cunningham, S.J., Holmes, G., 2001. Developing Innovative Applications in Agriculture Using Data Mining. Department of Computer Science, University of Waikato, Hamilton, New Zealand.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthuruswamy, R. (Eds.), 1996. Advances in Knowledge Discovery and Data Mining. AAAI Press, Menlo Park, CA, USA.

Garner, S.R., Cunningham, S.J., Holmes, G., Nevill-Manning, C.G., Witten, I.H., 1995. Applying a machine learning workbench: experience with agricultural databases. In: Proceedings of the Machine Learning in Practice Workshop, 12th International Machine Learning Conference, Tahoe City, CA, USA.

Harms, S.K., Goddard, S., Reichenbach, S.E., Waltman, W.J., Tadesse, T., 2001. Data mining in a geospatial support system for drought risk management. In: Proceedings of the First National Conference on Digital Government Research, vol. 1, Los Angeles, CA, pp. 9–16.

International Bull Evaluation Service (Interbull), 2004. Web site: www.interbull.org.

James, F., Roos, M., 1989. MINUIT Functional Minimization and Error Analysis, D506-Minuit, CERN.

Klei, L., Mark, T., Fikse, F., Lawlor, T., 2002. A method for verifying genetic evaluation results. In: Proceedings of the 2002 Interbull Meeting, 29, pp. 178–182.

Macrossan, P.E., Abbass, H.A., Mengersen, K., Towsey, M., Finn, G., 1999. Bayesian neural network learning for prediction in the Australian dairy industry. In: Hand, D.J., Joost, N.K., Berthold, M.R. (Eds.), Advances in Intelligent Data Analysis, Proceedings of Third International Symposium, IDA-99. Amsterdam, The Netherlands, pp. 395–406.

Michalsky, R.S., Chilausky, R.L., 1980. Learning by being told and learning from examples: an experimental comparison of the two methods of knowledge acquisition in the context of developing an expert system for soybean disease diagnosis. Int. J. Policy Anal. Inf. Syst. 4 (2), 125–161.

Pietersma, D., Grzesiak, W., Blaszczyk, P., Sablik, P., Lacroix, R., Wade, K.M., 2003. Decision-tree induction for the classification of services per conception in dairy cattle. In: Proceedings of the Joint Conference of ECPA (4th) and ECPLF (1st). Berlin Germany, pp. 775–776.

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA, USA.

Scherte, S.L., 2002. Data mining and its potential use in textiles: a spinning mill. PhD Dissertation. North Carolina State University, Raleigh, NC, USA.

Täubert, H., Swalve, H.H., Simianer, H., 2002. The Interbull audit project. Part II: development of a program for auditing breeding value estimation programs, Procedings of the 2002 Interbull Meeting, Interlaaken Switzerland, 29, pp. 165–167.

Triola, M.F., 2003. Elementary Statistics. Pearson Addison Wesley, Boston, MA, USA.

Wade, K.M., Lacroix, R., 1994. The role of artificial neural networks in animal breeding, Proceedings of the fifth WCGALP 22, pp. 31–34.

Yang, C., Prasher, S.O., Landry, J.A., 2002. Use of artificial neural networks to recognize weeds in a corn field. Trans. ASAE 45 (3), 859–864.