An Alarm Firing System for National Genetic Evaluation Quality Control

S. Diplaris[1], A.L. Symeonidis[1], P.A. Mitkas[1], G. Banos[2] and Z. Abas[3]

1Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Greece
2Department of Animal Production, School of Veterinary Medicine, Aristotle University of Thessaloniki, Greece
3Department of Agricultural Development, Democretus University of Thrace, Orestiada, Greece

1. Introduction

National genetic evaluation results form the basis of Interbull services. 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) offers an alternative way to examine data and extract valuable information (Han and Kamber, 2000), potentially leading to inference on data quality. In a recent progress report, the development of a DM algorithm for the analysis of national genetic evaluation results was presented (Banos *et al.,* 2003). Data quality was assessed by subjective inspection of DM results. The present study introduces a method to evaluate DM application results with objective criteria leading, when necessary, to the automatic issuing of warnings or alarm signals.

2. Material and Methods

2.1 Data description and data-mining module

Data and algorithms used were as described in Banos *et al*. (2003). Briefly, national genetic evaluations for production traits (milk, fat, protein) computed between February 1999 and February 2003 in 9 countries that had not changed their national genetic evaluation model during this period, were obtained from the Interbull Center. A separate data set of a certain country-evaluation run combination that included known errors was also obtained. A classification algorithm was developed, based on the C4.5 decision-tree classifier (Quinlan, 1993). Bull proofs were first categorized (min-max categorization on a scale 1-10) and a decision-tree model was then induced based on the associations discovered between the class variable (bull proof) and four input variables (bull birth year, type of proof, number of daughters and origin of bull). In the previous report (Banos *et al.,* 2003) predictions led to by associations discovered had been qualitatively compared to actual proofs and discrepancies had been confirmed in the data set with the known errors.

2.2 The bin fitter (or examining categorized bull proofs)

Bull proofs in the various decision-tree nodes are expected to follow normal (Gaussian) distribution. This is a result of normalization during data pre-processing. A tool measuring the correctness (or goodness or quality) of fit to the Gaussian distribution in each node would be useful. Since data were binned (i.e., categorized, with values 1-10), a bin fitter was used to fit data in each decision-tree node to the Gaussian distribution. The following chi-square value was estimated for this matter:

$$\text{chi}-\text{square} = \sum_{i} \frac{(f(x_i) - h_i)^2}{\sigma_i^2}$$

where 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 a linear function of x_i that minimizes the chi-square value, and σ_i is the error variance in the i^{th} bin. Minimization of the above quantity leads to the optimum estimates of the mean and standard deviation that define the best fitting Gaussian to our distribution.

The value of chi-square is a good measure for defining the quality (goodness) of fit, nevertheless, it is dependent on the number of data entries in each distribution. The more

points there are, the harder it will be to get a Gaussian distribution by chance, unless the data really follow such distribution. Thus, we had to normalize the value of chi-square. The key concept here is the degrees of freedom (df) that were computed as follows:

$$df = N_{data} - n_{param}$$

where N_{data} is the number of independent data points and n_{param} is the number of fitting parameters. In this case there were always two fitting parameters (mean and standard deviation). Thus, df $=N_{data} - 2$. We can define now as "quality of fit" measure (criterion) the normalized value of chi-square over df.

$$\text{Quality} = \frac{\text{chi} - \text{square}}{\text{df}}$$

This quality of fit criterion was calculated separately for each node of the induced decision-trees. When results truly follow Gaussian distribution, this value is expected to be no greater than 1. In case of significant deviations, a warning is issued, indicating possible erroneous distribution of bull proofs.

2.3 F- tests on node variance in consecutive evaluation runs

With the bin fitter, the nodes of each decision-tree model were individually compared to the expected Gaussian distribution. A more powerful approach would be to compare corresponding node distributions from different models (evaluation runs) against each other. Since the application of data mining indicated that there is a pattern in the structure of the decision-trees, it is possible to compare corresponding node distributions in decision-tree models induced from different evaluation runs. It should be noted there were no differences in national genetic evaluation models across runs and no new bulls were allowed into the system (Banos *et al.,* 2003), therefore, the variance of genetic proofs in corresponding nodes of two consecutive evaluation runs is expected to remain stable. This was tested with the following F-test, based on the ratio of two independent chi-square variables divided by their respective degrees of freedom.

$$F = \frac{\frac{df_1 \cdot s_1^2}{\sigma_1^2} / df_1}{\frac{df_2 \cdot s_2^2}{\sigma_2^2} / df_2}$$

where s^2 is the sample variance and σ^2 the estimate of the true variance of each node in evaluation runs 1 and 2. This F-test is designed to compare two sample variances (s_1 and s_2,) (Triola, 2003). So, if variances are equal, the F-value will be 1 (null hypothesis). Rejecting the null hypothesis for a particular node would imply that the node is "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 with the closest "parent" node.

2.4 The alarm firing system

A combination of the above two criteria was used to provide us with the necessary power to detect and isolate potential disruptions in the data sets. The technique developed issues warning at two levels, the "yellow" and the "red". If a node fails one of the two tests (i.e. either the chi-square or the F-test with the corresponding node of the subsequent run) then a yellow warning is issued. If a node fails both tests, then a red alarm is fired.

3. Results

Figure 1 shows the quality of fit test (chi-square) applied to five data sets, corresponding to consecutive genetic evaluation runs in a country; one of these data sets was knowingly erroneous (D4_e) and the other four had no known errors. 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

tended to infinity. These nodes were associated with bulls born between 1979 and 1985 (node 1.2.1) or bulls born between 1987 and 1990 that had more than 169 daughters (node 1.3.2). In such cases, a yellow warning was issued. Figure 2 shows the Gaussian fit for the histogram induced in these two nodes, demonstrating clear departure from normality. In the other cases, chi-square 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-square value was lower than the threshold.

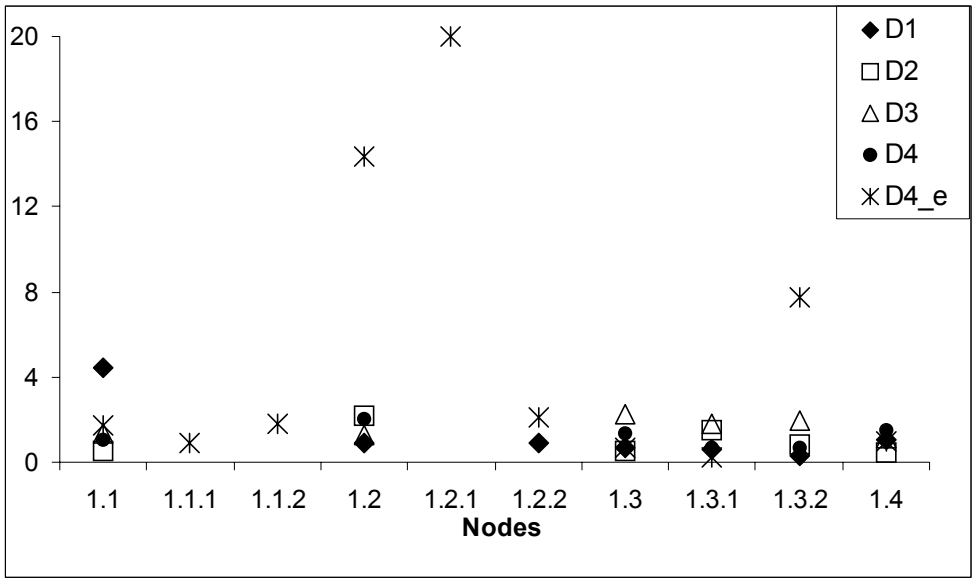

Figure 1. Quality of fit (chi-square) test for four data sets (runs) without known errors (D1-D4) and one with known errors (D4_e) in one country; values exceeding the dotted horizontal line indicate departure from normality.

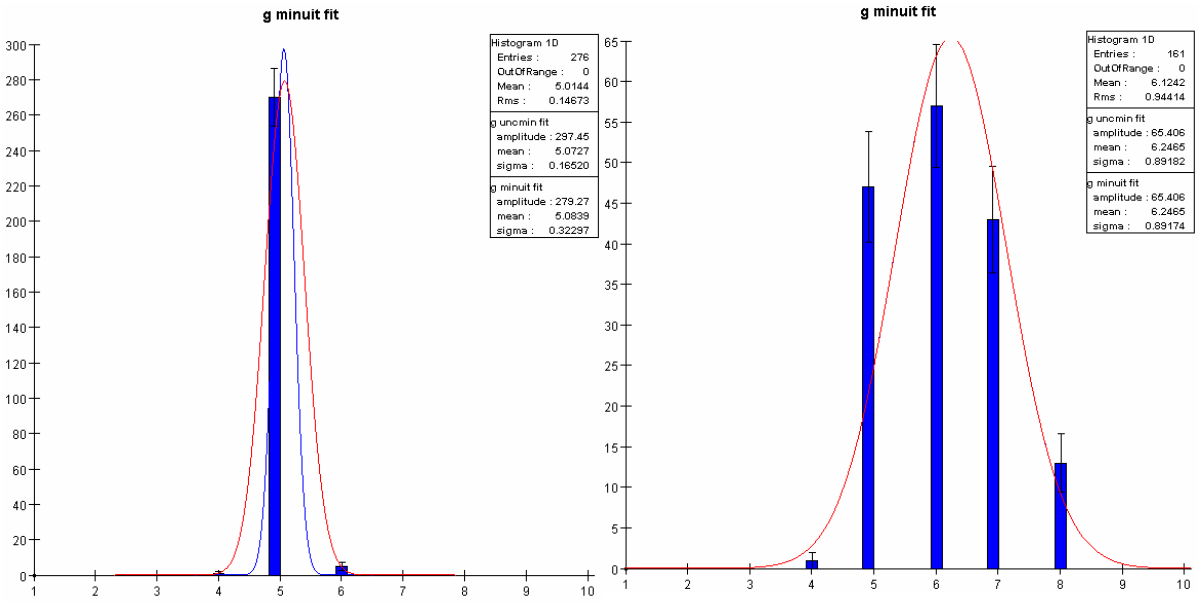

Figure 2. Gaussian fit for the histogram in nodes 1.2.1 (left) and 1.3.2 (right) from D4_e dataset.

When the F-test was applied to node-by-node comparison of consecutive evaluation runs, four (4) yellow warnings were issued in the case of D4_e. Two of these warnings were associated with nodes where yellow warnings had been also issued by the chi-square test. These two 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 pertained to bulls born between 1971 and 1980 and had more than 131 daughters. The same node had successfully passed the chi-square test. It should be also noted that node 1.1 in D1, where a yellow warning was issued by the chi-square test, passed the F-test.

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

Table 1 summarizes results from the combination of the chi-square and F-test in five consecutive evaluation run models, in three countries (one of which included the knowingly erroneous data set).

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.

	Country 1		Country 2		Country 3	
Model comparison	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

It should be noted that all yellow warnings and red alarms issued were for older bulls. This could be either because there are genuinely no problems with the evaluation of younger bulls in the countries studied here or because there is 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.

4. Summary and Conclusions

We presented a new alarm firing system that exploits results of DM application to national genetic evaluations of dairy bulls. The system examines the discovered patterns by combining two methods for individual and pairwise evaluation of decision-tree nodes. Each node distribution in the model is examined with regards to quality of fit to the Gaussian distribution; furthermore, its variance is compared to the corresponding node variance of the subsequent evaluation run. Results so far showed that this system is able to capture errors that are also confirmed by the standard Interbull procedure. Some additional warnings warranting closer examination were also issued. Furthermore, the key utility of this platform lays in its capacity to identify 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. In addition to the four input variables used in this study, additional variables describing the data might be needed.

References

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, Proc. 2003 *Interbull Annual Meeting, 31*, 8-15.

Han, J.W. & Kamber, M. 2000. *Data-miming: 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.

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

Triola, M.F. 2003. *Elementary Statistics,* Pearson Addison Wesley.

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.