AGELI: An Integrated Platform for the Assessment of National Genetic Evaluation Results by Learning and Informing

H. Eleftherohorinou[1], S. Diplaris[1], P.A. Mitkas[1] and G. Banos[2]
[1]Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Greece
[2]Department of Animal Production, School of Veterinary Medicine, Aristotle University of Thessaloniki, Greece

1. Introduction

We present an integrated platform for preprocessing, analysis, alarm issuing and presentation of national genetic evaluation data based on data-mining. Our goal is the integrated qualitative description of national genetic evaluation results, concerning three milk yield traits that constitute a critical issue in the range of services provided by Interbull. Although the standard method for quality assurance appears sufficiently functional (Klei *et al.,* 2002), during the last years there has been a progress concerning an alternative validation method of genetic evaluation results using data-mining (Banos *et al.,* 2003; Diplaris *et al.,* 2004; Han and Kamber, 2000), potentially leading to inference on data quality. A new alarming technique based on multiple criteria was recently established in order to assess and assure data quality (Diplaris *et al.,* 2004). The whole idea was to exploit data-mining techniques, i.e. decision trees, and then apply a goodness of fit test to individual tree nodes and an F-test in corresponding nodes from consecutive evaluation runs, aiming at discovering possible abnormalities in bull proof distributions at various regions. In a previous report (Banos *et al.,* 2003) predictions led to by associations discovered had been qualitatively compared to actual proofs and discrepancies had been confirmed using a data set with known errors.

AGELI (the Greek word for herd) is a software platform that integrates the whole data-mining procedure developed thus far. It can receive data from external remote databases and transform them to a form suitable for input to the Microsoft SQL Analysis Manager data-mining suite. Decision tree models can then be created and data quality can be assessed both by inspection of data-mining results and evaluation with objective criteria.

2. Material and Methods

2.1 AGELI description and functionality

AGELI has been developed in Java, in order to be able to communicate with the SQL Analysis Manager module, while maintaining the option to embed other data-mining algorithms in the future. The preprocessing procedure, as implemented in AGELI, requires the use of database tables, thus allowing the compact representation of data. The standard SQL Server database format was used, to be compatible with the form that data is represented in the main Interbull database. The Microsoft Decision Tree algorithm was used in order to mine bull evaluation data. This is a variation of the C4.5 algorithm (Quinlan, 1993) developed by Microsoft SQL Analysis Manager. The new algorithm was based upon the notion of classification. The algorithm builds a tree that will predict the value of a trait based upon the corresponding attributes in the training set. Therefore, each node in the tree represents a particular case for an attribute. The decision on where to place this node is made by the algorithm and a node at a different depth than its siblings may represent different cases of each attribute.

AGELI's preprocessing procedure creates new data files stored in SQL format in a local database. Furthermore, it exploits Data Transformation Services (DTS) in order to transform the SQL files to matrix format, thus enabling the communication with SQL Analysis Manager.

From a functional point of view, AGELI offers various services that are presented below:

- *Insert new data*: New genetic evaluation data can be inserted to the system. The new data is inserted in the form of SQL tables.
- *Preprocess data:* Bull proof data can be transformed using categorization techniques (min-max categorization on a scale 1-10) and attributes, such as number of daughters, can be normalized using z-score normalization.
- *Build decision tree models:* Decision tree models can be built in a batch form or one by one. The data-mining technique used is the one described in Diplaris *et al*. (2004). The trees induced are based on associations discovered between the class variable (bull proof for milk, fat and protein yield) and four input attributes (bull birth year, type of proof, number of daughters and origin of bull).
- *Chi-square test:* Individual decision tree nodes can be tested for goodness of Gaussian fit using the chi-square test as described in Diplaris *et al.* (2004). The platform allows for the dynamic set of an alarm threshold.
- *Comparative node validation:* F-tests can be applied in corresponding nodes from consecutive evaluation runs in order to compare them and discover interesting patterns, possible irregularities and different cases (Diplaris *et al.,* 2004).
- *Alarm issuing:* The results of the validation tests can be combined and various alarms are issued, in order to detect and isolate potential disruptions in the data sets. Specific warnings and alarms issued are described later.
- *Simple and functional user interface:* The user interface was developed in order to simplify its use and be comprehensible even to users who might not be sufficiently familiar with computer technology. A graphical representation of trees and chi-square distributions was implemented.
- *Integration of multiple tools in one program:* Since the target was the development of an integrated system that can process bull evaluation data and issue warnings and alarms, interaction with the SQL Analysis Manager was incorporated, along with the implementation of the two validation algorithms (individual node validation and validation through node comparison).

2.2 The alarm issuing procedure

The decision trees induced in AGELI are presented in an elegant way, such that all useful information can be easily accessed and viewed by the user. For this purpose, two special modes were designed that allow the user to view results of the validation tests in tabular and graphical form. Figure 1 depicts an AGELI screenshot that can provide the user with detailed information regarding the numerical results of the chi-square and F-test criteria, as well as the combined results, along with the potential issuing of warnings and alarms. These two tests were described by Diplaris *et al.* (2004). Briefly, the chi-square test checks the normality of each node, whereas the F-test compares the distribution variance between corresponding nodes in two consecutive evaluations. Users may select the country, evaluation run and trait they wish to examine and are provided with three analytical tables containing the following information for each node of the decision tree induced: **i)** the values of the chi-square test and a pass/fail qualifier, **ii)** the values of the F-test and a pass/fail qualifier, and **iii)** a table with combined results from the two tests, indicating possible warnings or alarms.

A more powerful and informative representation of the results is the graphical viewing of the decision trees. Using the tree viewing AGELI module, it is possible for the user to further examine the state of the distribution on the decision tree nodes, as various cases can be distinguished concerning the behavior of these nodes. Such cases can be observed with respect to the response of the tree node distributions to the various validation criteria, as shown in Table 1.

Table 1. Appearance of decision tree nodes based on two validation criteria (tests).

Chi-square test	F-test	Node color	Explanation
pass	pass	green	Problem-free node
pass	fail	yellow	Warning
fail	pass	yellow	Warning
fail	fail	red	Alarm
N/A	pass	green-grey	Cautiously problem-freed node
N/A	fail	yellow-grey	Cautious warning
pass	N/A	green-grey	Cautiously problem-freed node
fail	N/A	yellow-grey	Cautious warning
N/A	N/A	grey	Inconclusive node

N/A = non-applicable

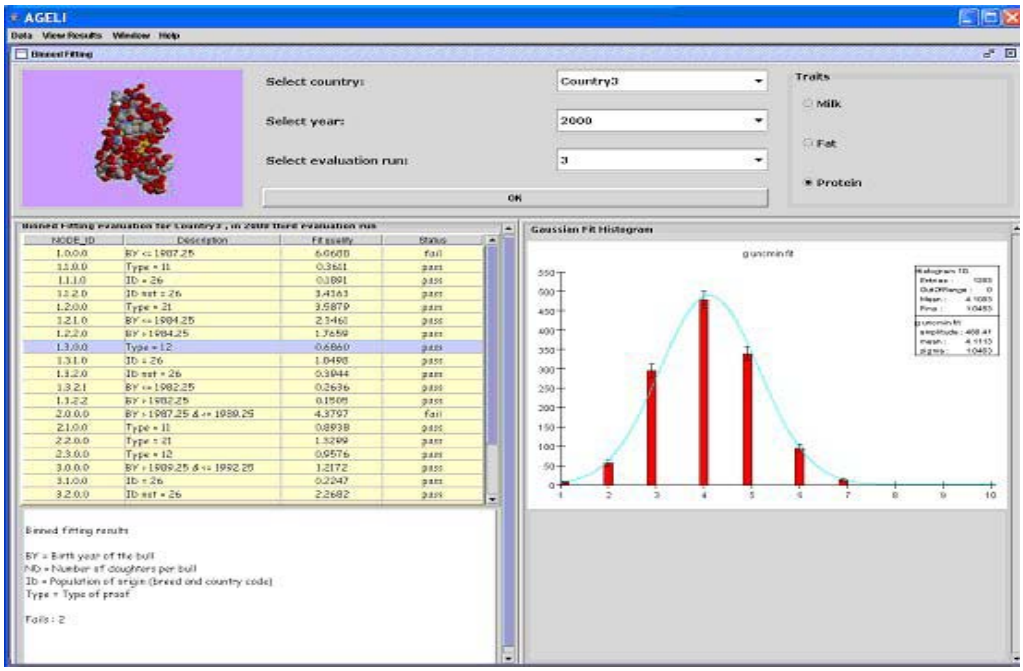

Figure 1. A screenshot of AGELI's graphical module for Gaussian fit node distributions.

Sure warnings are issued when both tests are applicable and one of them fails. A red alarm is fired when both tests fail, whereas the node is considered problem-free when it passes both tests. When either test is not applicable, then the node is considered inconclusive. The chi-square test may not be applicable when a node contains a very small number of bulls; the F-test can not be implemented if there is no direct node correspondence in two consecutive evaluation runs.

In the AGELI platform, all nodes are active and, by clicking on any of them, the user can access a separate window with all the information available about the specific node. A Gaussian fit on a histogram of the bull proof node distribution is formed, along with the numerical values and thresholds of the two validation tests. An informative label guides the user to look for the values and behaviors that introduced problems to the particular node.

Moreover, in a separate module, the user can graphically view the Gaussian fit distribution for any country, year, evaluation run, trait and node. Besides, the user can examine qualitatively the overall behavior of a tree throughout the whole sequence of evaluation runs, thus developing knowledge of the whole model behavior across time. Figure 2 illustrates a snapshot of the tree viewing AGELI module.

3. Experiments and Results

AGELI was used in order to conduct the whole series of experiments described in Diplaris *et al.* (2004). Briefly, national genetic evaluations for three milk production traits (milk, fat and protein yields) computed between February 1999 and February 2003 in 9 countries that had not changed their national genetic evaluation model during that period were obtained from the Interbull Center. Only bulls with a genetic evaluation in all 17 runs in a country were included in the analysis. The platform presented here was applied to all countries, traits and evaluation runs.

The platform was proved very efficient in time considering the data loading and preprocessing, and the decision tree model building procedure. The whole algorithm training procedure lasted about three hours in a Pentium 4 processor at 3.5GHz with 1GB RAM. The validation tests were conducted dynamically on demand and the results were stored locally on the disk in ASCII mode for the numerical values and in GIF format for the Gaussian fit distribution histograms.

Concerning the results of the validation procedure, 90.28% of the nodes were green, 2.5% yellow, 0.02% red and the remaining 7.2% in various combinations with grey. Results were in agreement with those described in the previous report (Diplaris *et al.,* 2004). Some new inconclusive cases were discovered, suggesting that more research is required to investigate their possible cause. Five out of the nine countries were characterized as quite stable. In one country there were known problems in an unofficial run that were picked up by the system. In three countries there were a number of alarms and inconclusive nodes, warranting further probing.

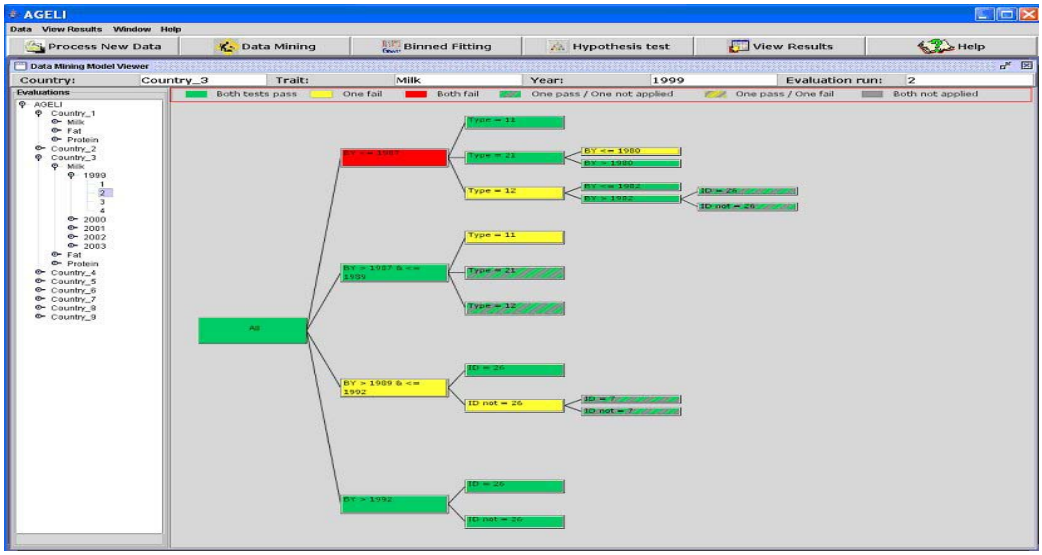

Figure 2. A screenshot of AGELI's tree visualization module.

4. Summary and Conclusions

We presented AGELI, a software platform that integrates two validation tests into a data-mining procedure, in order to validate routine national genetic evaluation results for dairy bulls. The platform induces decision trees and examines them thoroughly by combining two methods for individual and pairwise evaluation of their nodes. Several different cases were defined to better describe the behavior of the bull proof distribution in each decision tree node.

AGELI's user-friendly graphical interface provides the user with all available information in cases of particular interest. Results have demonstrated the platform's efficacy in finding and depicting special cases of interest and localizing potentially erroneous data. Future add-ons to the platform could be the

integration of new, flexible data-mining techniques and the quantification of the degree of matching between consecutive decision trees in order to provide additional measures for model validation and enhance the platform utility.

Acknowledgements

The authors thank the Interbull Center and the nine countries that made data available for this study.

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.

Diplaris, S., Symeonidis, A.L., Mitkas, P.A., Banos, G. & Abas, Z. 2004. An Alarm Firing System for National Genetic Evaluation Quality Control, Proc. 2004. *Interbull Meeting 32,* 146-150.

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.

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