Mendelian Sampling Mining and Cluster Monitoring of National Genetic Evaluation Data with the AGELI Software Platform

P. Tsimpos[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 innovative approach for pre-processing, analysis, alarm issuing and presentation of national genetic evaluation data with AGELI using Mendelian sampling mining and clustering techniques. AGELI (Eleftherohorinou *et al.,* 2005) is a software platform that integrates the whole data mining procedure in order to produce a qualitative description of national genetic evaluation results, concerning three milk yield traits. Quality assurance constitutes a critical issue in the range of services provided by Interbull. Although the standard method appears sufficiently functional (Klei *et al.,* 2002), during the last years there has been progress concerning an alternative validation method of genetic evaluation results using data mining (Banos *et al.,* 2003; Diplaris *et al.,* 2004), potentially leading to inference on data quality. This methodology was incorporated in AGELI in order to assess and assure data quality. The whole idea was to exploit decision trees and apply a goodness of fit test to individual tree nodes and an F-test to corresponding nodes from consecutive evaluation runs, aiming at discovering possible abnormalities in bull proof distributions. In a previous report (Banos *et al.,* 2003) predictions led to associations, which were qualitatively compared to actual proofs, and existing discrepancies were confirmed using a data set with known errors.

In this report we present AGELI's novel methods of performing data mining by using a series of decision tree and clustering algorithms. Different decision tree models can now be created in order to assess data quality by evaluating data with various criteria. To further assess data quality, a novel technique for cluster monitoring is implemented in AGELI. It is possible to form clusters of bulls and perform unsupervised monitoring on them over the entire period of genetic evaluation runs. Finally, analyses were conducted using bull Mendelian sampling over the whole dataset.

2. Material and Methods

2.1 Data description

AGELI uses data-mining algorithms in order to mine bull evaluation data. These algorithms induce decision-tree models based on the associations between four input variables (birth year of the bull, type of proof, population of origin, number of daughters) and the class variable (milk, fat, protein yield proof). National genetic evaluations for production traits computed between February 1999 and February 2003 in 9 countries that had not changed their national genetic evaluation model during that period were considered (Banos *et al.,* 2003). True IDs of bulls were recoded to ensure data confidentiality.

Previous work was based on the analysis of actual bull proofs. An alternative approach is to use the Mendelian sampling as predicted variable. Mendelian sampling is unaffected by selection and in this study it was computed as follows:

$$MS = EBV_{BULL} - \frac{1}{2}\left(EBV_{SIRE} + \frac{1}{2} EBV_{MGS} \right) \quad (1)$$

where MS is the Mendelian sampling, EBV_{BULL} is the genetic proof of a bull, EBV_{SIRE} the genetic proof of the bull's sire and EBV_{MGS} the genetic proof of the bull's maternal grandsire.

For any particular bull in the dataset, Mendelian sampling was computed if the genetic proofs of his male ancestors (sire and maternal grandsire) were recorded over the

whole evaluation period (February 1999 – February 2003, total of 17 runs). Table 1 depicts the number of bulls for which the Mendelian sampling value could be calculated.

Table 1. Number of bulls with proofs and bulls with ancestors' proofs in all 17 runs, in 9 countries.

Country	1	2	3	4	5	6	7	8	9	Total
Bulls evaluated in all runs	4162	4408	7650	12393	5607	1320	3240	3403	26046	68,229
Bulls with ancestors evaluated in all runs	2299	2663	5371	7812	2087	236	2492	1640	21880	46,480
Percentage	55.2%	60.4%	70.2%	63.0%	37.2%	17.9%	76.9%	48.2%	84.0%	68.1%

2.2 Data-mining modules

In the previous reports (Diplaris *et al.*, 2004) the decision-tree classifier developed in AGELI was a Microsoft Decision Tree algorithm aiming at gaining knowledge and discovering patterns in bull evaluation data. This classification algorithm is part of the Microsoft SQL Analysis Manager and cannot be customized by the user. Six additional decision-tree classifiers, drawn from the WEKA data-mining suite (WEKA3, 2007), were included in the new version of AGELI (v2.0):

- *C4.5*: Algorithm for generating a pruned or unpruned C4.5 decision tree (Quinlan, 1993).
- *M5P*: Decision tree algorithm (Wang and Witten, 1997).
- *LMT*: Classifier for building 'logistic model trees', which are classification trees with logistic regression functions at the leaves (Landwehr *et al.*, 2003).
- *REPtree*: Simple fast decision tree learner; builds a decision/regression tree using information gain/variance and prunes it using reduced-error pruning (with back fitting).
- *RandomTree*: Simple classifier for constructing a tree that considers K randomly chosen attributes at each node without performing any pruning.
- *NBTree*: Classifier for generating a decision tree with naive Bayesian classifiers at the leaves (Kohavi, 1995).

The basic feature of these decision-tree classifiers, except for NBTree, is that each one has a set of customizable attributes, enabling the user to control the model building and training procedure.

Moreover, four customizable clustering algorithms were also incorporated in AGELI v2.0:

- *FarthestFirst*: Implements the "Farthest First traversal algorithm" (Hochbaum *et al.*, 1985).
- *SimpleKmeans*: Simple k-means clustering algorithm (Kanungo *et al.*, 2000).
- *EM*: Simple EM (expectation maximization) classifier; EM assigns a probability distribution to each instance which indicates the probability of it belonging to each of the clusters.
- *Cobweb*: Classifier implementing the Cobweb clustering algorithm (Kaldor, 1934).

Since clustering algorithms do not create any graph like decision-trees, the parameter that mainly affects a model shape is the number of clusters (classes) to which the data instances are separated.

2.3 The alarm firing system

Diplaris *et al.* (2004) described in detail two criteria which provide a way to detect and isolate possible irregularities or potential disruption in the datasets. Briefly, the first criterion (chi-square test) checks the quality of fit to the normal (Gaussian) distribution in each node and raises a flag if this quality is over a predefined threshold. The second criterion (F-test) compares corresponding node distributions from two consecutive evaluation runs. Node correspondence is established if the two nodes follow exactly the same path to the root. If the corresponding node distribution variance ratio is statistically different from unity then the test has failed.

AGELI's tree viewing module paints each node of a decision tree with a color that denotes which tests have been successfully passed (Eleftherohorinou *et al.,* 2005).

However, in cluster models, the technique for corresponding cluster identification in consecutive models needs to be different from the one implemented for decision-tree models. To this end, a special technique was implemented in AGELI's software to find cluster matches: cluster X of evaluation dataset T_i is a match with cluster Y of evaluation data set T_{i+1} if and only if cluster Y contains more than 50% of the data elements (i.e., recoded bull IDs) found in cluster X and less than 50% of the elements of any other cluster of evaluation data set T_i (Spiliopoulou *et al.,* 2006). F-test is performed only when a cluster match is identified. In any other cluster transitions (i.e. cluster split, cluster absorption, cluster disappearance or new cluster emergence) the F-test is not applicable.

3. Results and Discussion

3.1 Decision-tree Classification

Decision-tree models built using the above presented classifiers were applied to country-datasets with known errors or those found to be unstable in the previous report (Eleftherohorinou *et al.,* 2005).

When bull proofs were used as predicted variables, decision-trees that were induced by the presented algorithms appeared quite similar to those induced by Microsoft Decision Trees. These similarities were mostly detected in the lowest node levels of the tree, where the main splitting (decision-making) criterion was the bull birth year. The known problems in one country-dataset were discovered by most classifier models, either by failing both tests, or by inducing model patterns that were completely different than the previous evaluation models, thus making it impossible to perform F-tests. In some unstable country-datasets, where in the previous report a number of alarms and inconclusive nodes had occurred, the same behavior was observed by using the new tools, thus warranting further probing.

When the Mendelian sampling was used as a predicted variable, the amount of bull evaluation data was decreased considerably due to absence of certain ancestor proofs. This limited the ability of the algorithms to detect patterns and trends. The dataset with known errors yielded a model that changed completely its pattern in comparison with the previous evaluation, but in the subsequent evaluation run the pattern changed back to normal. All tested algorithms were able to detect the problem.

3.2 Clustering

Clustering is the partitioning of a data set into subsets (clusters). It does not create any kind of graph like decision-tree classification, but it is a powerful technique that enables monitoring of the clusters in consecutive evaluation runs. AGELI's model visualizer module can illustrate a clustering model as shown in Figure 1.

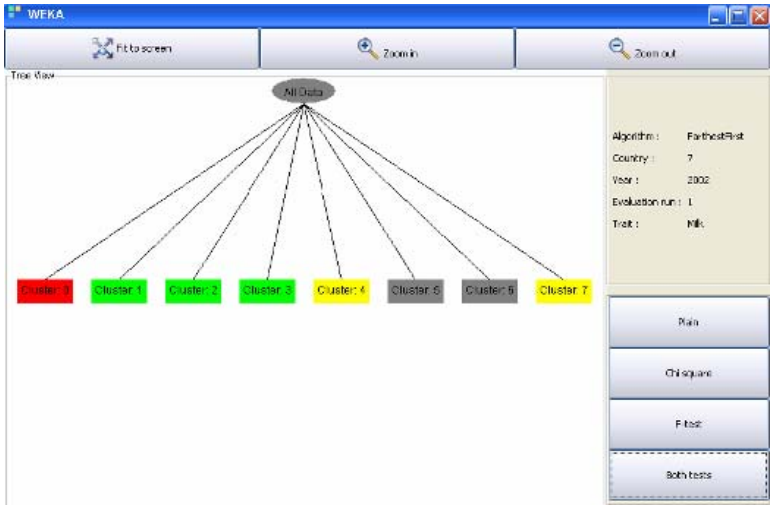

Figure 1. Clustering model.

In the conducted analyses the number of clusters for each model was fixed to a value at which the model would contain as few clusters as possible, while also optimizing inter-cluster data distribution. A minimum of 50 elements per cluster was required in order to avoid small clusters with outliers. The number of clusters ranged between 8 and 12 per country-dataset.

Most (over 90%), of the induced clusters survived through the whole evaluation period since most of them were also found in previous evaluation datasets. That makes the F-test comparison much more applicable to clustering models than to decision-tree classifiers.

The erroneous dataset triggered red alarms (both tests failed) in all models no matter which clusterer was chosen, even when the predicted variable was the Mendelian sampling of bull proofs. Country-datasets with unstable (inconclusive) patterns in decision-tree models induced some yellow clusters (one test failed) warranting more research to investigate their possible cause. Very rarely, cluster models would be induced for unstable country-datasets, where F-test was not applicable.

4. Summary and Conclusions

We presented an enhanced version of AGELI (v2.0), a software platform that integrates data mining tools (such as decision-tree classification and clustering) for the quality assessment of national genetic evaluation data.

AGELI is now a fully integrated software platform with graphical representation capabilities for the data mining models and the identification of potential erroneous data. The second version of the software incorporates a new series of data-mining algorithms, which, in contrast to the previous ones, can be customized by the user. Each algorithm has a set of attributes that can be configured appropriately by data-mining experts, in order to achieve more robust results.

The data mining modules of the platform induce patterns in data, which are examined using individual and pairwise validation tests in order to check data integrity. The fact that most algorithms tested here detected similar patterns of potential data disruptions suggest that these tools can be considered as indicators of data quality. More research is needed to refine the presented techniques and investigate their applicability to individual cow records. The latter will be the ultimate way to validate the developed tools.

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. *Interbull Bulletin 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. *Interbull Bulletin 32*, 146-150.

Eleftherohorinou, H., Diplaris, S., Mitkas, P.A. & Banos, G. 2005. AGELI: An Integrated Platform for the Assessment of National Genetic Evaluation Results by Learning and Informing, Proc. 2005 Interbull Meeting. *Interbull Bulletin 33*, 183-187.

Hochbaum, D.S. & Shmoys, D. 1985. A Best Possible Heuristic for the K-Center Problem, *Mathematics of Operations Research, 10(2)*, 180-184.

Kaldor, N. 1934. A Classificatory Note on the Determination of Equilibrium. *Review of Economic Studies, 1*, 122-136.

Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C., Silverman, R. & Wu, A.Y. 2000. The Analysis of a simple k-means Clustering Algorithm, *Proc. 16th ACM Symposium on Computational Geometry*, pp. 101-109.

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

Kohavi, R. 1996. Scaling up the accuracy of naive-Bayes classifiers: a decision tree hybrid, *Proc. of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 202-207.

Landwehr, N., Hall, M. & Frank, E. 2003. Logistic Model Trees, *Proceedings of the 16th European Conference on Machine Learning*, pp. 241-252.

Spiliopoulou, M., Ntoutsi, I., Theodoridis, Y. & Schult, R. 2006. MONIC – Modeling and Monitoring Cluster Transitions, *Proc. 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 706-711 .

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

Wang, Y. & Witten, I.H. 1997. Induction of model trees for predicting continuous classes. *Proceedings of the poster papers of the European Conference on Machine Learning*, pp. 128-137.

WEKA3, 2007. http://www.cs.waikato.ac.nz/ml/weka/