International Journal of Biological, Life and Agricultural Sciences
ISSN: 2415-6612
Vol:8, No:5, 2014

Methods for Distinction of Cattle Using Supervised Learning

Radoslav Židek, Veronika Šidlová, Radovan Kasarda, Birgit Fuerst-Waltl

***Abstract*—**Machine learning represents a set of topics dealing with the creation and evaluation of algorithms that facilitate pattern recognition, classification, and prediction, based on models derived from existing data. The data can present identification patterns which are used to classify into groups. The result of the analysis is the pattern which can be used for identification of data set without the need to obtain input data used for creation of this pattern. An important requirement in this process is careful data preparation validation of model used and its suitable interpretation. For breeders, it is important to know the origin of animals from the point of the genetic diversity. In case of missing pedigree information, other methods can be used for traceability of animal´s origin. Genetic diversity written in genetic data is holding relatively useful information to identify animals originated from individual countries. We can conclude that the application of data mining for molecular genetic data using supervised learning is an appropriate tool for hypothesis testing and identifying an individual.

***Keywords*—**Genetic data, Pinzgau cattle, supervised learning.

I. Introduction

CONSIDERING that wild cattle no longer exist and that all surviving genetic diversity is now present in domestic animals, a better understanding of cattle genetics could help us to reduce some of these undesirable effects [1]. One of the important steps in development of efficient breed protection programs is characterization of population genetic variability and assessment of genetic structure [2].

Pinzgau cattle is an originally alpine breed adapted in mountain areas. Nowadays, this breed belongs to the endangered populations [3] due to drastic decreasing of the animal counts. Taking in the account the situation of alternatively breeding programs were optimized [4]. Currently loss of genetic resources concerns not only the extinction of traditional breeds, but also the loss of genetic diversity within breeds. Therefore, through information on diversity and population structure in cattle is urgently needed to serve as a rational basis for the conservation and possible use of indigenous cattle breeds as genetic resources to meet potential future demands [5].

This project work was co-funded by European Community under project no 26220220180: Building Research Centre "AgroBioTech".

R. Židekis with the Department of Food Hygiene and Safety, Slovak University of Agriculture in Nitra, 94976 Slovakia (e-mail: radoslav.zidek@uniag.sk).

V. Šidlová and R. Kasarda are with the Department of Animal Genetics and Breeding Biology, Slovak University of Agriculture in Nitra, 94976 Slovakia.

B. Fuerst-Waltl is with the Department of Sustainable Agriculture Systems, University of Natural Resources and Life Sciences Vienna, Gregor-Mendel-Strasse 33, A-1180 Vienna, Austria.

Markers are used by population genetics to investigate the origin, genetic diversity and population structure of alleles, by evolutionists to describe genetic relationship among species or populations and by geneticists to study linkage disequilibrium within or between genes [6]. Molecular markers based on DNA have a very high polymorphism level, and they have been successfully used for evaluation of genetic diversity and variation in breeding programmes with an impact on the level of genetic conservation schemes [7]. Microsatellite markers are considered as a marker of choice to characterize breeds for diversity assessment [8]. Their short length makes them amenable to amplification by polymerase chain reaction. So far, autosomal microsatellites have been the most popular markers for characterizing the genetic constitution of breeds, establishing breed relationships, describing the history of livestock, the uniqueness at the breed level [9], [10] for the selection of breeding animals from divergent groups in order to maximize the genetic variation and consequently the fitness [11].

Recently machine learning techniques have gained popularity in the field for their ability to successfully classify unknown samples [12]. This process of automatically learning from data and in turn using that acquired knowledge to inform future decisions is extremely powerful. At the core of machine learning lies a set of complicated algorithms which have been developed over the course of the past few decades by academics in a diverse set of disciplines [13], [14]. Two facets of mechanization should be acknowledged when considering machine learning in broad terms. Firstly, it is intended that the classification and prediction tasks can be accomplished by a suitably programmed computing machine. That is, the product of machine learning is a classifier that can be feasibly used on available hardware. Secondly, it is intended that the creation of the classifier should itself be highly mechanized, and should not involve too much human input which could affect the selection and performance of the algorithm. Both the creation of the algorithm and its operation to classify objects or predict events are based on concrete, observable data [15].

Supervised machine learning is the search for algorithms that reason from externally supplied instances produce general hypotheses, which then make predictions about future instances. On the other hand, the goal of supervised learning is to build a concise model of the distribution of class labels in terms of predictor features. The resulting classifier is then used to assign class labels to the testing instances where the values of the predictor features are known, but the value of the class label is unknown [16]. Supervised learning provides algorithms to automatically build predictive models only from

observations of a system. During the last twenty years, supervised learning has been a tool of choice to analyse the always increasing and complexifying data generated in the context of molecular biology, with successful applications in genome annotation, function prediction, or biomarker discovery [17].

The aim of this study was to develop a model for verifying animal identity and to classify observed individuals into Slovak and Austrian group using supervised learning models.

II. Material and Methods

Selected cows of Pinzgau cattle originated from Austria and Slovakia were analysed. DNA of 412 animals was isolated from hair roots and amplified in one multiplex PCR with 8 microsatellites (TGLA227, SPS115, ETH3, BM1824, CSRM60, CSSM66, TGLA122, INRA23). To determine the polymorphism of microsatellite DNA sequences fluorescent fragmentation analyses by capillary electrophoresis (ABI PRISM 310 Genetic Analyser)were used and the allele sizes were evaluated using software Gene Mapper 4.0.

All observed animals were divided into 2 logical groups according to countries of origin. A dataset consisted of 346 Slovak animals and 66 Austrian animals. The data were used to develop classification models for identity verification of animals. Statistical analysis was conducted using Tanagra 1.4 software [18]. Supervised learning accounts for a lot of research activity in machine learning and many supervised learning techniques have found application in the processing of multimedia content. The typicalcharacteristic of supervised learning is the availability of annotated training data. The name invokes the idea of a “supervisor” that instructs the learning system on the labels to associate with training examples. Typically these labels are class labels in classification problems. Supervised learning algorithms induce models from these training data and these models can be used to classify other unlabelled data.

Data mining statistical approaches using supervised classification were used in the learning phase. Supervised learning methods, which can analyse continuous data, were used to ensure high quality and relevant outputs. In the learning phase were analysed 20 methods of supervised machine learning and their ability to classify examined data (n=412). The basic output of supervised learning methods is "confusion matrix" or error rate matrix. This table represents the number of classified individuals using statistical method to some logical group (e.g. Slovak animals) expressed by index. Matrix itself operates with two outputs. One of the values is the "recall" value, which refers number of animals correctly excluded from evaluated group of individuals. Number of animals incorrectly included to evaluated group of individuals represents "1-precission" value. Values in rows represent the Slovak (SK) and Austrian (AT) population estimated using individual models.

Sometimes method for the classification of a known data set can classify correctly all analysed data to the logical groups. Providing further exploiting of this method as a suitable for classifying without testing, may be seen as a phenomenon called "rote" or "memorization". Statistical method classifies to the logical groups without errors, but when data are changing or adding, shows the considerable errors due to memorization. Logical pattern to classifying of unknown data for statistical method is not available, but it proves precisely describe data from the learning phase. The verifying of the algorithm reliability based on processes "bootstrapping" and "cross validation" are used to detect this phenomenon. Both these methods are designed for estimating the generalization error based on resampling [19]. Generally, validation can be executed using the same set of samples (i.e. Leave-out cross validation) or using a new set of samples (external validation). In the cross validation, some (n) samples are removed from the model, the model is recalculated and with the new model, predictions are obtained for each of the removed samples. The samples are placed back into the data set and other samples are again removed. This procedure is repeated as many times as necessary to obtain predictions for all samples. In the external validation, a new set of samples of known class are analysed and predicted with the model, and these predictions are compared with the real identity of the samples.

In the using phase was submitted algorithm to the test. For construction of the algorithm 75% of the data were used and remaining 25% were presented to algorithm as unknown classification.

III. Results

A model to verifying animal identity was developed by the useof microsatellite panel and multivariate statistics. Application of all available models of supervised learning for data set preparation, we observed the reliability of individual methods in order to choose the (approximate) best one. From 20 tested methods three of them have been selected with highest value of reliability.

TABLE I
Reliability of Learning Process and Validation Reliability

Method		Recall	1-Precision	Algorithm error	Bootstrap .632+	CV error
C4.5	AT	0.879	0.0794	0.0316	**0.074**	0.081
	SK	0.986	0.023			
CS-MC4	AT	0.575	0.036	0.068	0.081	**0.078**
	SK	0.996	0.071			
Rnd Tree	AT	1.000	0.000	**0.000**	0.09	0.12
	SK	1.000	0.000			

As shown in Table I, method with the lowest algorithm error in direct classification was the Rnd Tree method, by applying the classification techniques using decision trees. Machine learning based on decision trees is currently used for data mining analysis in many sectors of biology and food [20].

Methods C4.5 and CS-MC4 were appeared as preferred due to phenomenon called memorization of Rnd Tree method. Although these methods recorded the higher value of the algorithm error in the phase of direct learning, after verifying the reliability using bootstrapping (Bootstrap .632+) and cross validation (CV error), was recorded lower error rate. Even 99.6% of animals can be correctly assigned to Slovak

subpopulation and excluded from Austrian subpopulation only with 3.6% error rate using CS-MC4 method.

Simple decision trees are easy to understand how the classification has been built and which attributes were used. New instances are classified by following the tree along the relevant branches, depending on the attributes of sample. Methods, such as C4.5, start with an empty tree and iteratively split the data, creating branches of the tree, until they decide to assign all examples of a branch to a specific class, creating a leaf of the tree, based on a certain criteria (e.g., all examples in the node belong to the same class) the error in the branch of the tree is small enough [21]. Upgrade of C4.5 method is using of one variable, with efficient dividing the data set into groups. The criterion for data dividing is normalized information gain of variable. Providing the variable shows a high information gain, is then used as a condition for the decision. The superstructure of C4.5 is CS-MC4 method, described in [22]. This method is optimized for huge data set with a large number of variables.

TABLE II
RELIABILITY OF USING PROCESS

Model		Recall	1-Precision	Generalization error
C4.5	AT	0.842	0.111	**0.048**
	SK	0.976	0.035	
CS-MC4	AT	0.79	0.063	**0.048**
	SK	0.988	0.046	
Rnd Tree	AT	0.737	0.300	0.107
	SK	0.929	0.060	

The test set (25% of animals) is used for assessment of the generalization error of the final chosen model (Table II). The methods C4.5 and CS-MC4 have been confirmed again as the most reliable for classification of animals to the group by country origin, reaching the equal errors of observed models (4.8%). We can conclude the correct classification rate obtained with the reliability validation of the model were sufficient for identifying of animals.

IV. CONCLUSION

Genetic structure of two Pinzgau cattle populations has been analyzed and used as a model for supervised learning of different statistical methods. A result of provided study shows the possibility to classify unknown samples according to genetic data. Genetic diversity written in genetic data presents useful information to identifying country of origin for individual animals. Model is also useful for classification on many logical levels as breed type, breeding system, herd and many others. Supervised learning is the suitable tool for hypothesis testing using genetic data. Using supervised learning allowed us to clearly distinguish between animals of Austrian and Slovak origin. This is in opposite with generally accepted idea of closed genetic relationship between population due to commonly used sires. On the level of DNA it is still possible to classify animals to separate populations.

REFERENCES

[1] F. C. Canavez, D. D. Luche, P. Stothard, K. R. M. Leite, J. M. Sousa-Canavez, G. Plastow, J. Meidanis, M. A. Souza, P. Feijao,S. S. Moore,L. H. Camara-Lopes, "Genome Sequence and Assembly of Bosindicus," *J. Hered.*, vol. 103, no. 3, pp. 342-348, Feb. 2012.

[2] M. Simčič, M. Čepon, S. Horvat, S. Jovnovac, V. Gantner, P. Dovč, D. Kompan, "Genetic characterization of autochtonous cattle breeds, Cika and Busha, using microsatellites," *ActaAgr.Slovenica*, supl. 2, pp. 71-77. Sept. 2008.

[3] O. Kadlečík, H. H. Swalve, J. A. Lederer, H. Grosu, "*Development of dual –purpose Pinzgau cattle*," SPU, Nitra, Slovak Republic, pp. 128, ISBN 80-8069-439-7. 2004.

[4] O. Kadlečík, R. Kasarda, L. Hetényi, „Geneticgain, increase in inbreeding rate and generation interval in alternativesofPinzgaubreeding program," *Czech J. Anim. Sci.*, vol. 49, no. 12, pp. 524-531. Dec. 2004.

[5] P. Taberlet, A. Valentini, H. R. Rezaei, S. Naderi, F. Pompanon, R. Negrini, P. Ajmone-Marsan, "Are cattle, sheep, and goats endangered species?"*Mol. Ecol.*,vol. 17, no. 1, pp. 275-284.Jan. 2008.

[6] K. Liu, S. V. Muse, "Integrated analysis environment for genetic marker data,"*Bioinformatics*, vol. 21, no. 9, pp. 2128-2129.Feb. 2005.

[7] R. Židek, R. Kasarda, "Distribution of genetic distance within groups with different relationship coefficient," *ActaFytotech.Zootech*.,vol. 13, pp. 73-76.Oct. 2010.

[8] FAO. *The state of the world's animal genetic resources for food and agriculture*. FAO, Rome. 2007.

[9] W. B. Sun, H. Chen, C. Z. Lei, X. Q. Lei, Y. H. Zhang,"Study on population genetic characteristics of Qinchuan cows using microsatellite markers," *J. Genet. Genomics*, vol. 34, pp. 17-25.Jan. 2007.

[10] J. A. Lenstra,L. F. Groeneveld, H. Eding, J. Kantanen, J. L. Williams, P. Taberlet, E. L.Nicolazzi, J. Sölkner, H. Simianer, E. Ciani, J. F. Garcia, M. W. Bruford, P. Ajmone-Marshan, S. Weigend, "Molecular tools and analytical approaches for the characterization of farm animal genetic diversity,"*Anim. Genet*., vol. 43, no. 5, pp. 483-502.Oct. 2012.

[11] J. Goudet,l. Keller,"The correlation between inbreeding and fitness: does allele size matter?"*Trends Ecol. Evol*., vol. 17, no. 5, pp. 201-202.May 2002.

[12] D. L. Samson, T. J. Parker, Z. Upton, C. P. Hurst,"A Comparison of Methods for Classifying Clinical Samples Based on Proteomics Data: A Case Study for Statistical and Machine Learning Approaches,"*PloS one*, vol. 6, no. 9, e24973.Sept. 2011.

[13] H. Brink, J. W. Richards, "*Real-Word Machine Learning*," Manning Publications Co. Pp. 400, ISBN 9781617291920. 2013.

[14] A. L. Swan, A. Mobasheri, D. Allaway, S. Liddell, J. Bacardit, "Application of Machine Learning to Proteomics Data: Classification and Biomarker Identification in Postgenomics Biology," *OMICS*, vol. 17, no. 12, pp. 595-610. Dec 2013.

[15] A. L.Tarca, V. J. Carey, X. Chen, R. Romero, S. Drăghici,"Machine Learning and Its Applications to Biology," *PlosComput. Biol*., vol. 3, no. 6, pp. 116.Jun 2007.

[16] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," Informatica, vol. 31, pp. 249-268.2007.

[17] P. Guerts, A. Irrthum,L. Wehenkel,"Supervised learning with decision tree-based methodsin computational and systems biology,"*Mol. BioSyst*., vol. 5, no. 12, pp. 1593-1605. Dec.2009.

[18] R. Rakotomalala, "TANAGRA: a free software for research and academic purposes", *Proc. of EGC'2005*, RNTI-E-3, vol. 2, pp.697-702. 2005. (in French)

[19] R. Kohavi, "A study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," *Proc. of the 14th IJCAI*, vol. 2, pp. 1137-1143, ISBN 1-55860-363-8, 1995.

[20] R. C. Barros, M. P. Basgalupp, A. C. P. L. F. de Carvalho, A. A. Freitas, "A Survey of Evolutionary Algorithms for Decision Tree Induction," *IEEE Transactions on Systems, Man, and Cybernetics – Part C:Appl. Rev*., vol. 42, no. 3, pp. 291-312, May 2012.

[21] R. Quinlan, "*C4.5: programs for machine learning*," San Francisco, CA: Morgan Kaufmann Publishers Inc., ISBN 1-55860-238-0, 1993.

[22] J-H. Chauchat, R. Rakotomalala, M. Carloz, C. Pelletier, "Targeting customer groups using gain and cost matrix; a marketing application," *Mining for Marketing*, 2001.