PRACTICAL ASPECTS FOR CLASSIFICATION OF PERFORMANCE PARAMETERS OF PIGS BY DECISION TREES

Schlegel H., Fischer K., Wähner M.

Anhalt University of Applied Sciences, Germany

Abstract

The potential of performance of breeding sows and other livestock gets described by a variety of data. Those data get analysed constantly. Aims of investigations are defined and approved. Therefore different methods are used, well known ones and relatively new ones. With the help of those different methods relationships should be represented and interpreted. One – in comparison – relatively new way of Data Mining is the classification with the help of decision trees. There is one big advantage, because decision trees do not have any conditions according to the scale and distribution of parameters. Many factors can get analysed at the same time. Different kinds of decision trees can be used for known and unknown interactions between parameters. That is also an advantage. For representation of the usage of decision trees in livestock production one practical example has been chosen. The main aim of this investigation is the interpretation of the results from a biological and economical point of view. In addition to that the way of using a decision tree should get demonstrated. In this case, 1.780 sows of one pig breeding farm are used. They got classified according to the parameter of number of total born piglets, live born piglets, weaned piglets and number of litter, sire and technician for insemination. It is very important to derive the correct sequence steps from the statistical results. That's necessary for all interpretations.

Key Words: Pig, performance parameters, decision trees.

Sows of modern breed do have a high genetic potential. The parameter of longevity and fertility are very important for a good lifetime performance of sows. A good longevity and lifetime performance are the basis for an efficient pig production. Another important aspect you need to consider is animal welfare.

By using a variety of different statistical methods different relationships between selected parameters can get detected in complex. A statistical analysis of single or only a few numbers of parameters happens continuously. By using decision trees you are able to analyse more parameters in complex. There is no condition on known or unknown relationships between those parameters.

Material and Methods

In order to demonstrate the application of decision trees 1.780 sows of one pig breeding company have been used. These sows are between the 1st and 12th litter. They got grouped with the help of the program SPSS 21.0. For a more complex interpretation of the characteristics of a sow in each group six impact factors have been chosen. If you need an interpretation in the way of economy and animal welfare you need a higher number of parameters. The used ones in this example are three metric parameters (number of total born piglets, live born piglets and weaned piglets) and three nominal parameters (number of litter, sire and technician for insemination).

Methods of classification are cluster analysis, discriminant analysis and artificial neural nets and decision trees. The application of decision trees has been announced first by the publication of Breiman, Friedman, Olshen and Stone (BREIMAN et al., 1984). This method is one more easy and also successful way of machine learning (RUSSEL et al., 2012). Those tree structures have been developed first for categories, now one can use it for numeric parameters also (ESTER et al., 2000).

On this basis the reasons for those optimal characteristics need to get determined. This can be genetic parameters as well as management tools. To find them in mathematical way knowledge of decision trees and biological topics is necessary. If you use decision trees all cases are divided into classes and rules are created. Those rules can get visualised as a tree or as a classical rule. Starting point of a visual description is a root node, which contains all data. Ending points are leaves, which represent single groups and the number of rules. The structure is a hierarchical one. In the following figure one simple decision tree is demonstrated.

The creating of a decision tree is divided into four phases (PETERSON, 2005). In the 1st phase one attribute gets selected to split the root node. In the 2nd phase the splitting of the root node is performed into two or more subsets. In the 3rd phase includes the stop of splitting and the conformation of leafs. This happens if additional information leads to a stop of the process of splitting. The last and 4th phase is called pruning. This avoids the effect of overlearning and simplifies an interpretation. This classification can be performed with one dependent and as many independent variables as you need. The complexity of the decision tree is not limited. In some programs it is also possible to use only independent variables. With SPSS you can just analyse questions with one dependent and as many independent variables as you want. The advantages of decision trees are their explainability and traceability of the record based classification rules (SCHLITTGEN, 2009). There will remain a certain amount of mistakes, outliers and missings. Additional the meaning of variables can get derived. The quality of decision trees is depending on the quality of the data. The more representative the data, the better the decision tree. This way to solve a problem is also used, if there are minor changes in the data base (e.g. missing variables) which should not lead to a changing in the result (SCHLITTGEN, 2009).

There are some problems. Heuristic methods of classification are only able to detect a local not a global optimum. Other risks are the so called overfitting and strong correlations between the examined variables (SCHLITTGEN, 2009). Nether the less classification trees are used very often in other disciplines (BÜHL, 2012). In agriculture decision trees are not a favoured way to solve mathematical or statistical problems.

With the help of the table 1 an example of a structure of a tree should get demonstrated. Therefore mushrooms are used.

The characteristics of the parameter color, size, points and class lead to several possibilities to divide mushrooms according to the first parameter. This is shown in the figure 2. In brackets you find the number of mushrooms in each leaf of the tree.

The classification on the left side in figure 2 is very useful and clear. Despite that, the classification on the right side is incorrect. In one group (leaf) there are toxic and eatable mushrooms classified together. The quality of a classification tree depends on the kind of classification, which I choose. Each parameter needs to get proved before to find the best variable for splitting (SCHLTTGEN, 2009).

Figure 1: Example of a classification tree

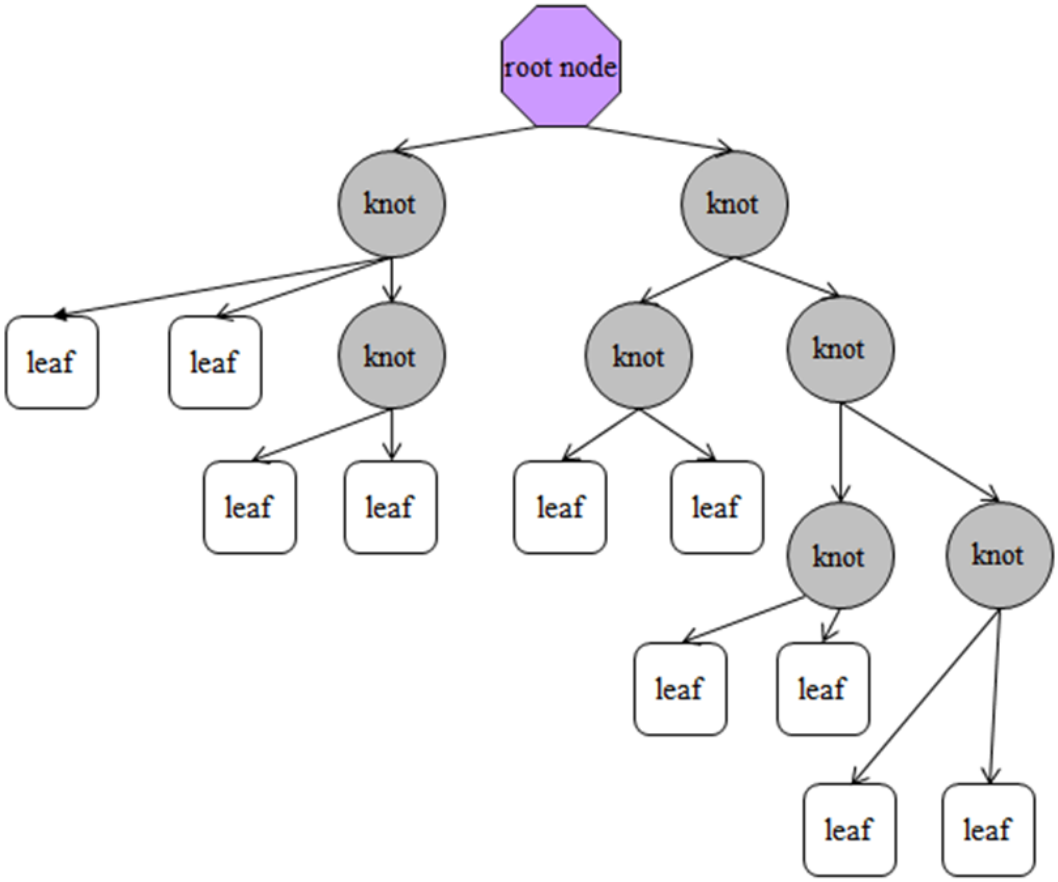
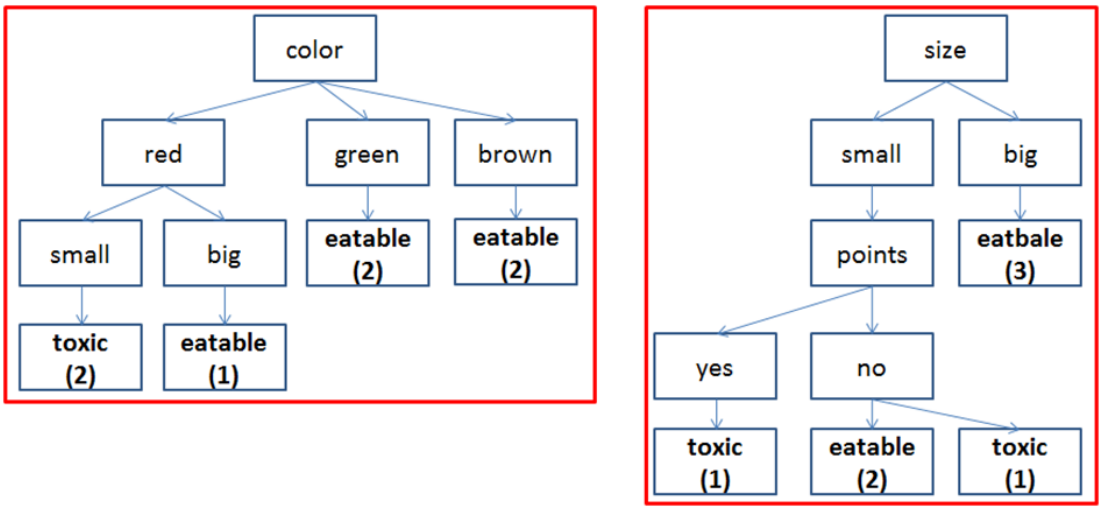


Table 1. Example of a classification by decision trees

color	size	points	class
red	small	yes	toxic
green	big	no	eatable
brown	small	no	eatable
brown	big	yes	eatable
red	small	no	toxic
green	small	no	eatable
red	big	no	eatable

Figure 2. Examples for classification by decision trees



For classification by decision trees you can operate with several methods in SPSS. These are the CHAID Algorithm (Chi-Squared Automatic Interaction Detection), Exhaustive CHAID Algorithm, the CRT Algorithm (Classification and Regression Trees) and the QUEST Algorithm (Quick, Unbaised, Efficient Statistical Tree). Those algorithms differ most in two facts, the parameter for splitting and the sensitivity to stop the splitting.

The CHAID algorithm uses the Chi Square Test to examine the meaning of parameters. In each step of creating a decision tree the independent variable gets determined which has the strongest relation to the dependent one. This relation may not be significant. In the next step the different categories of the independent variable which differ not significantly get combined. They get put together in one node or one leaf at the end of the process. Despite that significant different categories of one parameter get split. Metric variables need to get grouped into a defined number of classes before starting the classification tree. SPSS uses 10 categories as a standard. The user is able to choose up to 200 categories.

The Exhaustive CHAID Algorithm analyses all possible splitting possibilities and chooses the best result at the end. That's why this method is much more precise and in theory the best method (BÜHL, 2012). The CRT Algorithm divides the data into homogenous groups. One possibility to split those groups are e.g. a priori probabilities. The dependent variable can be of each scale (ECKSTEIN, 2008). The QUEST Algorithm can only get performed with nominal variables. That's the reason why it is not used very often in praxis.

Within the different ways of classification there are several possibilities to split. To calculate the split attribute you may use Entropie, Information Gain or Information Gain Ratio. Entropie is a factor which reflects the gain of additional information which you can get by splitting (SCHLITTGEN, 2009). There is more gain in information if the insecurity of possible mistakes decreases. One possible formula to calculate Entropie is the following:

$$H(S) = -p_+ \log_2 p_+ - p_- \log_2 p_-$$

Definition: S amount of data
p+ amount of positive examples
p- amount of negative examples
$\log_2$ logarithm on basis 2

The gain of information gets calculated subsequently by subtraction of the Entropie of the latest variants from the Entropie before. The highest number in each case represents the split attribute which fits most. If you use Information Gain p^2 will be used instead of $\log_2$ in the formula (STEINLEIN, 2004). The method of Information Gain Ratio considers also the uniformity of the splitting. This means that the level of purity gets weight by all amounts of the data in the following nodes. The next splitting follows at the end of this selection process and in connection with the algorithm of classification.

One main problem of classification by trees is the perfect fitting on the data set. This leads to a limited possibility to generalize. Decision trees are a good method for classification if you need exact „if than" rules and if the data set are final ones. If the data are not final, you need to choose a method to generalize. Deviations of biological data are a normal case. There are two approaches to generalize. The first possibility of validation is the using of one data set for training and one or more data sets for testing. The structure of a tree gets determined by the help of a training data set. One or more data sets of validation are used at the same time to prove the possibility to generalize. The data set for testing are defined by cross validation and split-sample validation. The second approach starts at the end the classification process. After that leaves or nodes which are not necessary any more get erased by different methods (STEUERER, 1997). Another point is to define the structure of a tree easier and to simplify the possibility to generalize. Therefore you need to define a minimum number of cases per leaf. You need to consider the sum of all data. Pruning can get started by calculating the mistakes of classification. According to the rules this factor represents the number of wrong assigned data. In the next step the leaves and parts of a tree get cut off which have the highest influence on the factor of mistakes of classification (SCHLTTGEN, 2009). In all cases you need to consider the aim to classify all data. If there are too many cases with problems you need to create a node with those data. Another way of pruning is using the test of significance (RUSSEL et al., 2012). The method of Cost – Complexity - Prunig calculates the relevance of all leaves and nodes and erases the nodes and leaves of the lowest relevance (PETERSON, 2005). In that case to erase data means that data of those sows are included in the leaf on the next leaf. The formula of Cost – Complexity – Pruning is as follows:

$$g(n_k) = \frac{R(n_k) - R(UB_{blatt})}{h_{abs}(UB_{blatt} - 1)}$$

with:

$$R(n_k) = \left(1 - \max \frac{h_{abs}(n_k, k_r)}{h_{abs}(n_k)}\right) * \left(\frac{h_{abs}(n_k)}{h_{abs}(T)}\right)$$

and

$$R(UB_{blatt}) = \sum_{k=1}^{K} R(n_k)$$

Definition:

$g(n_k)$: relevance of the subsequent node or leaf

$R(n_k)$: statistical significance of the following node

$R(UB_{blatt})$: statistical significance of the examined leaf

n_k : following node

k_r : following category

h_{abs} : observed frequency

T: data for trainig

Pruning can get executed subsequently and also during the classification algorithm (PETERSON, 2005). Nodes can get changed into leaves by special criteria for stopping the classification process. This may happen if there is only a few number of cases in one leaf or if a following splitting does not lead to a new knowledge.

Results and Discussion

Six methods have been used to classify 1.780 sows. The number of total born piglets is used as a dependent variable. All other variables (number of live born piglets and weaned piglets, number of litter, sire and technician for insemination) were used an independent variables.

For classification the following methods have been chosen CHAID, Exhaustive CHAID and CRT. Afterwards a cross validation and a split sample validation is used in combination with the Exhaustive CHAID method. The 6th attempt was an Exhaustive CHAID method in combination with a cross validation and 25 categories instead of 10.

Those six methods of classification created from 9 (CRT) up to 18 (last attempt) leaves. The sire was not used in any case for classification. An additional test by a contingency analysis confirmed that there are no significant correlations ($0.064 \leq p \leq 0.535$) between leaves and sire in all six methods. All other parameters have been used for classification. The 6th attempt of classification fits most for interpretation. This is the method which worked with the Exhaustive CHAID method combined with a cross validation. As result 18 leaves have been created by using 25 intervals. The figure 3 shows one part of the classification tree.

In the figure 4 and table 2 the results from the worst (one and two) and the two best leaves (three and four) are demonstrated combined with the cross table for the parameter technician for insemination.

Figure 3. One detail of a classification tree of 1.780 sows

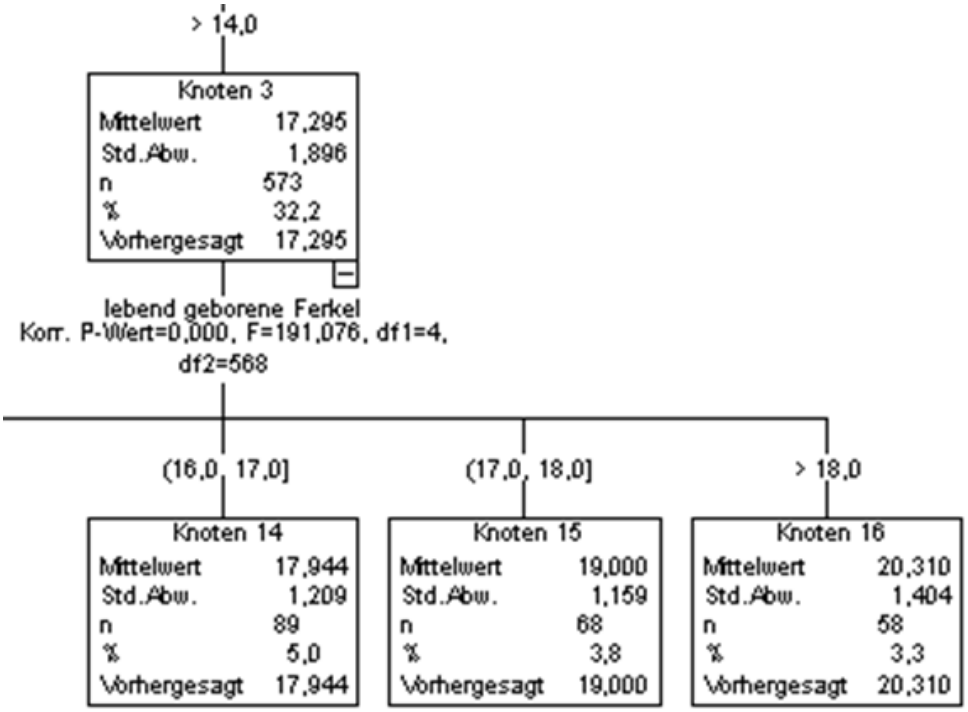


Figure 4. Mean values of sows within the leaves of a classification tree

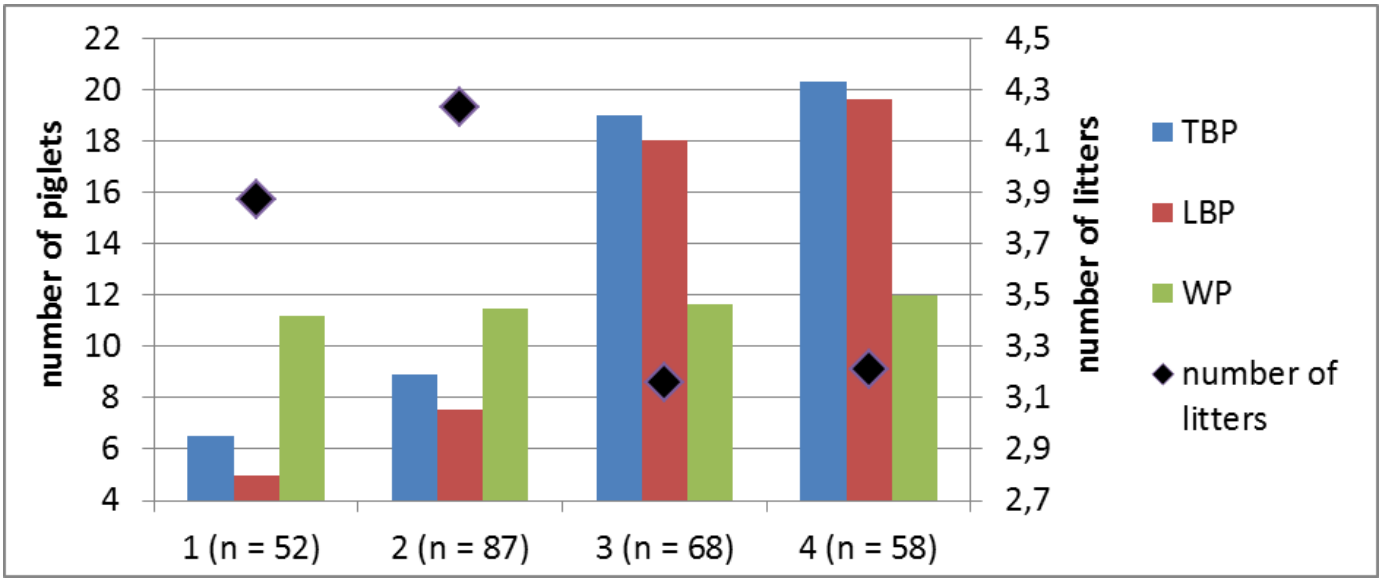


Table 2. Descriptive statistics of the classification of the sows

leaves	TBP	LBP	WP	number of litters
1 (n = 52)	6.52 ± 2.96	4.96 ± 1.20	11.21 ± 1.64	3.87 ± 2.37
2 (n = 87)	8.89 ± 2.07	7.52 ± 0.50	11.46 ± 1.61	4.23 ± 2.76[a]
3 (n = 68)	19.00 ± 1.16	18.00 ± 0.00	11.65 ± 1.29	3.16 ± 1.75[b]
4 (n = 58)	20.31 ± 1.40	19.62 ± 0.99	12.00 ± 1.62	3.21 ± 1.45[b]

[a:b;] $p<0.05$

Sows of the 3rd and 4th leaf achieve best results with 20.3 and 19.0 total born piglets and 19.6 and 18.0 live born piglets in average. Those numbers are above the mean results of whole farm. In table 2 one can see also the low standard deviation in this parameter. The number of piglets is not the only essential parameter. The birth weight of the piglets needs also to get considered. The results from sows in the 1st and 2nd leaf are too low.

The numbers of weaned piglets per litter are not informative because of the system of cross fostering. In general the number of weaned piglets is at an adequate level. The number of litter ranges between 3.2 and 4.2. There was only a significant difference between sows from the 2nd leaf to sows from the 3rd and 4th leaf ($p=0.016$; $p=0.033$).

In table 3 data of the technician for insemination are shown. The influence of the technician is significant on creating nodes and leaves of a tree. But it is very difficult to interpret those results. This demonstrates that not all data, which are of mathematical relevance, do also have an influence in praxis.

In general the classification by a tree was successful. All sows of the 3rd and 4th leaf should get examined further to find the reason for their good performance. Influences of genetics and management should be in focus.

Table 3. Frequency of the use of the technician for insemination for the sows of different leaves

technician for insemination	1st leaf	2nd leaf	3rd leaf	4th leaf
1	3.40%	7.55%	6.79%	5.28%
2	6.42%	9.06%	4.15%	4.91%
3	4.15%	4.15%	4.15%	4.53%
4	0.38%	0.38%	0.75%	0.00%
5	1.89%	2.64%	1.13%	1.51%
6	0.38%	1.89%	1.51%	1.89%
7	0.00%	0.00%	0.38%	0.00%
8	3.02%	6.79%	6.42%	3.40%
9	0.00%	0.38%	0.38%	0.38%

Conclusion

With the help of a theoretical reflection and a practical example it is proven that classification trees are a useful tool to group sows by different parameters. In a second investigation possible impact factors for a good performance need to get examined. Classification trees can be used not only for questions in pig breeding and hog feeding but also for questions in other fields of activity.

References

BREIMAN, L.; FRIEDMAN, J.H.; OLSHEN, R.A.; STONE, C.J.: Classification and regression trees. Wadsworth International Group, Belmont CA, 1984

BÜHL, A.: SPSS 20 - Einführung in die moderne Datenanalyse. 13. aktualisierte Auflage, München, 2012

ECKSTEIN, P.P.: Angewandte Statistik mit SPSS. Praktische Einführung für Wirtschaftswissenschaftler, 6. überarbeitete Auflage, Wiesbaden, 2008

ESTER, M.; SANDER, J.: Knowledge Discovery in Databases – Techniken und Anwendungen. Berlin, 2000

PETERSON, H.: Data Mining: Verfahren, Prozesse, Anwendungsarchitektur. München, 2005

RUSSEL, S.; NORVIG, P.: Künstliche Intelligenz - ein moderner Ansatz. 3. aktualisierte Auflage, München, 2012

SCHLITTGEN, R.: Multivariate Statistik. München, 2009

STEINLEIN, U.: Data Mining als Instrument der Responseoptimierung im Direktmarketing: Methoden zur Bewältigung niedriger Responsequoten. Göttingen, 2004

STEURER, E.: Ökonometrische Methoden und maschinelle Lernverfahren zur Wechselkursprognose. Heidelberg, 1997

Corresponding Address:
Hannes Schlegel
Anhalt University of Applied Sciences,
Strenzfelder Allee 28, 06406 Bernburg, Germany
E-mail: Hannes.Schlegel@gmx.de