Cluster analysis application in research on pork quality determinants

W. Przybylski[1)], P. Wąsiewicz[5)], P. Zieliński[5,6)], J. Gromadzka-Ostrowska[2)], E. Olczak[1)], D. Jaworska[1)], S. Niemyjski[3)] and V. Santé-Lhoutellier[4)]

Warsaw University of Life Sciences,
[1)]Faculty of Human Nutrition and Consumer Science,
Department of Engineering and Catering Technology
[2)]Faculty of Human Nutrition and Consumer Science, Department of Dietetics
Nowoursynowska 159C, 02-787 Warszawa, Poland
[3)]PenArLan Spółdzielcza 2a, 64-100 Leszno, Poland
[4)]Quality of Animal Products INRA 63122 Saint Genés Champanelle, France
[5)]Institute of Electronic Systems, Warsaw University of Technology
[6)]Interdisciplinary Centre for Mathematical and Computational Modeling, University of Warsaw

ABSTRACT

In this paper data mining methods were applied to investigate features determining high quality pork meat. The aim of the study was analysis of conditionality of the pork meat quality defined in coherence with HDL and LDL cholesterol concentration, plasma leptin, triglycerides, plasma glucose and serum. The research was carried out on 54 pigs. originated from crossbreeding of Naima sows with P76-PenArLan boars hybrids line. Meat quality parameters were evaluated in samples derived from the *Longissimus* (LD) muscle taken behind the last rib on the basis: the pH value, meat colour, drip loss, the RTN, intramuscular fat and glycolytic potential. The results of this study were elaborated by using R environment and show that cluster and regression analysis can be a useful tool for in-depth analysis of the determinants of the quality of pig meat in homogeneous populations of pigs. However, the question of determinants of the level of glycogen and fat in meat requires further research.

Keywords: pigs, meat quality, cluster analysis, regression tree, decision tree, lda

1. INTRODUCTION

The increase of pig meatiness in swine industry caused by selective breeding of high muscularity foreign pigs may lead to decrease porcine quality (greater incidence of defects such as PSE, ASE or DFD). Lack of desirable characteristics of meat depends on genetic and environmental factors or may be result of expression of unknown defects. Recent studies showed in high muscularity pigs increased drip loss [Bertram, 2000] and high decrease of level of intramuscular fat. Their effect on eating pork quality such as tenderness, juiciness, flavour and colour was also investigated. Full explanation of the causes of the decreasing meat quality is aim of many studies and researches. Despite knowledge of many genes such as RYR1, RN$^-$, CAST, GLUT4, H-FABP affecting meat quality, considerable differences are observed. Recent studies on energy balance regulation in muscle tissue and interaction between intramuscular fat and glycolytic potential in muscle showed that it is possible to observe groups with differences in metabolism in population of high muscularity pigs[P]. Result of these studies carried out that improvement of meatiness in breeding programs, which is the main purpose in European countries, may be associated with decreasing sensory pork quality due to interactions between genes with roles in energy balance, lipid and insulin metabolism. At the same time reducing the fat usually leads to a lower level of intramuscular fat in muscle tissue and consequently to increase the amount of glycogen in the muscles which may have important implications for their development of technological and sensory quality of meat. In consequence this process increases the quantity of meat with a low pH_{24}.

The aim of the study was analysis of conditionality of the pork meat quality defined in coherence with HDL and LDL cholesterol concentration, plasma leptin, triglycerides, plasma glucose and serum. Cluster analysis was described in this paper.

[1] Corresponding autor: e-mail: wieslaw_przybylski@sggw.pl

Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments 2010, edited by Ryszard S. Romaniuk, Krzysztof S. Kulpa, Proc. of SPIE Vol. 7745, 77450Y
 · doi: 10.1117/12.872255

2. MATERIAL AND METHODS

The research was carried out on 54 pigs (36 barrows and 18 gilts) originated from crossbreeding of F1 sows polish landrace with polish large white breed with P76-PenArLan boras hybrids line. One day before slaughter blood was taken from the anterior caval vein into heparinized tubes and centrifuging at 3000 rpm for 10 min. Obtained blood plasma was transferred into vials and frozen at -28ºC.

Plasma level of triglycerides (TG), total cholesterol (CHOL), HDL cholesterol and glucose was determined enzymatically using PTH Hydrex kits (Warsaw, Poland). LDL cholesterol was calculated according formula: LDL=CHOL - (TG/5+HDL). Plasma leptin concentration was measured using the DSL (Webster, TX, USA) porcine leptin IRMA (DSL-82100) kit. Assays were performed as per the manufacturer's instructions with sensitivity 0.05 ng/ml.

On the next day the animals were slaughtered in the slaughterhouse situated 50 km from the farm, according to the following conditions: the rest time- two hours of pre-slaughter, automatic electric stunning and exsanguinations in the horizontal position, carcass was chilled in fast cooling system. The backfat thickness and loin thickness was determined by using CGM apparatus and on this basis meat percent in carcass was calculated. Meat quality parameters were evaluated in samples derived from the *Longissimus* (LD) muscle taken behind the last rib. The pH value was measured in 1, 3 and 48 h after slaughtering directly in the muscle tissue. Meat colour was measured in CIE L*a*b* system by Minolta CR310 chromameter in 48 h *post mortem*. The natural drip loss was determined [Prange ,1977] methods. The RTN (Rendement of Technological Yield of Napole) as yield of meat in curing and cooking processing was determined [Naveau ,1985]. The fat in muscle was determined in according to Soxhlet method (PN-73/A-85111). In LD glycogen, glucose and glucose-6-phosphate after glycogen hydrolysis with amyloglucosidase [Dalrymple and Hamm, 1973] and lactate [Bergmeyer, 1974] were determined. On the basis of them the glycolytic potential (GP) was calculated [Monin and Sellier, 1985].

3. CLUSTER ANALYSIS AND REGRESSION TREES

A decision tree is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a classification or decision. A decision tree takes as input an object or situation described by a set of properties, and outputs a yes/no decision. Functions with a larger range of outputs can also be represented [Russel,2009].

Linear discriminant analysis (LDA) method finds a linear combination of features, which characterize or separate two or more classes of objects or events. LDA is closely related to principal component analysis (PCA), analysis of variance (anova) and regression analysis, especially logistic and probit regression.

Data clustering is the unsupervised clustering of patterns e.g. observations, data items, feature vectors into groups, so the data in each subset share common trait like proximity according to some defined distance measure [Kaufman et al.1990]. Data clustering algorithms can be hierarchical or partitional. Hierarchical algorithms find successive clusters using previously established clusters, whereas partitional algorithms determine all clusters at once. Hierarchical algorithms can be agglomerative ("bottom-up") or divisive ("top-down"). Agglomerative algorithms begin with each element as a separate cluster and merge them into successively larger clusters. Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters.

A key step in a hierarchical clustering is to select a distance measure. A simple measure is Manhattan measure equal to the sum of absolute differences for each variable. A more common measure is Euclidean distance, computed by finding the square of the distance between each variable, summing the squares, and finding the square root of that sum. A property of the Euclidean space is that distances are symmetric (the distance from object A to B is the same as the distance from B to A).

Given a distance measure, elements can be combined. Hierarchical clustering builds (agglomerative nesting hierarchical clustering), or breaks up (divisive analysis clustering), a hierarchy of clusters. The traditional representation of this hierarchy is a tree data structure (called a dendogram), with individual elements at one end and a single cluster with every element at the other. Agglomerative algorithms begin at the top of the tree, whereas divisive algorithms begin at the bottom. Cutting the tree at a given height will give a clustering at a selected precision.

In partitional clustering the k-means algorithm assigns each point to the cluster whose center (also called centroid) is nearest. The center is the average of all the points in the cluster — that is, its coordinates are the arithmetic mean for each dimension separately over all the points in the cluster.

In fuzzy clustering (fuzzy analysis clustering), each point has a degree of belonging to clusters, as in fuzzy logic, rather than belonging completely to just one cluster. Thus, points on the edge of a cluster, may be in the cluster to a lesser degree than points in the center of cluster. For each point its coefficients of being in the clusters are computed. As in k-means algorithm the minimum is a local minimum and the results depend on the initial choice of weights.

4. EXPERIMENT RESULTS

For clustering several procedures from R environment were used: diana (DIvisive ANAlysis Clustering), fanny (Fuzzy Analysis Clustering with squared dissimilarity input matrix), hclust (Hierarchical Clustering with squared dissimilarity input matrix), agnes (Agglomerative Nesting - Hierarchical Clustering with dissimilarity input matrix), kmeans (K-Means Clustering), cutree (dendogram cutting into groups function used for diana, hclust, agnes). A cluster number was set to two, three or four clusters as is seen in Table 1. After creating clusters on each attribute column and cluster label column analysis of variance (ANOVA) was calculated. Numbers of attributes with p value equal or less than 0.05 were put in Table 1. As is seen fanny procedure found the best four groups with 20 columns accepted by anova analysis.

In the further investigation some attributes were removed like glycogen, lactate, marbling, which were connected with other remaining ones e.g. glycolytic potential of LD muscle. In every cluster means of parameters were computed and for all clusters their statistical significance (p_anova – p value) was determined with a use of anova tests for each parameter. Finally, standard deviation (sd) for every attribute was calculated as is provided in Table 2.

In this paper we research a relationship between parameters measured in vivo (last 6 rows in Table 2) and after slaughter. In the mentioned table the C2 group has a very good quality meat due to the highest level of pH48, hot carcass weight, RTN, the lowest level of glycolytic potential of LD muscle, colour attributes, driploss, protein, drymass, water, cholesterol HDL, total cholesterol sum, glucose in serum and mean level of leptin, triglicerides, cholesterol HDL. The worst quality meat was in the C3 group where the highest levels of total cholesterol sum, intramuscular fat level, glycolytic potential, drymass, leptin, colour attributes and the lowest levels of pH, water, hot carcass weight, back fat and loin thickness. Thus, the meat quality is determined by pH48, hot carcass weigth, RTN, glycolytic potential of LD muscle, colour attributes, total cholesterol sum.

The two groups C1 and C4 are stranger ones. In C4 the highest levels of glucose in serum and triglicerides, driploss, cholesterol HDL are joint with the lowest level of intramuscular fat level, LDL and sufficient level of pH48 and in consequence a sufficient meat quality (especially cholesterol levels are good for special healthy diet). In C1 worse quality meat is connected with the highest levels of back fat, loin thickness, meat percent, cholesterol LDL, the lowest levels of RTN, leptin, triglicerides, glucose in serum.

Finally, for good quality meat the following parameter levels measured in vivo has great influence: low values of cholesterol sum, cholesterol LDL, mean values of leptin, rather high values of cholesterol HDL, triglicerides. Rather high means values are high or near mean.

In Table 2 only glucose in serum levels has no relationship with meat parameters. In order to find it regression trees which usually have quite large classification error, but represent main tendencies in the given data, were created ten times on randomly chosen each time 70 percent of our dataset, which was earlier discretized into 3 levels. The best tree was selected with the smallest error of classification of the remaining 30 percent of the given dataset (the testing set). The whole procedure was repeated three times.

A tree depicted in Fig. 1 is one of two calculated trees for glycogen and has classification error on the whole dataset about 40 percent and appropriate attribute means at its leaves are put in Table 3. The another one was similar, but has no intermediate node cholesterol_sum and his classification error was larger. Thus, this tree represents the main scheme of hormones' relations in the given pig population and for its leaves means of IMF, glycolytic potential of LD muscle, leptin, total cholesterol sum, glucose in serum were computed. At leaf 1 the more leptin the less glycogen. Leptin also is in a relationship with IMF. For low leptin values and small total cholesterol or for rather low leptin levels and high total cholesterol glycogen has mean values. For mean leptin level, rather high cholesterol level and low glucose (in serum)

level glycogen has the highest values. The more glycogen, the less intramuscular fat. The less fat, the better meat quality and production efficiency.

The mentioned results confirmed cluster significance, not only statistical, but there was no tool to guess a meat quality before slaughter, so classification trees with nodes chosen from blood serum (which was obtained in vivo) attributes were computed as is provided in Fig. 2.

Table 1. For a given method of clustering and number of clusters a number of our dataset columns (num_anova_rows) accepted by anova – analysis variance with p-value of F-statistic equal or less than 0.05

method	num_clusters	num_anova_rows
diana	2	12
fanny	2	16
hclust	2	4
agnes	2	11
kmeans	2	16
diana	3	15
fanny	3	15
hclust	3	5
agnes	3	14
kmeans	3	17
diana	4	18
fanny	4	20
hclust	4	8
agnes	4	16
kmeans	4	18

Table 2. Parameters' means in clusters, where C1,C2,C3,C4 denotes a cluster name, p_anova – p value of clustering statistical significance, sd – standard deviation

Castrated males	9	11	9	7		
Gilts	4	6	1	7		
Clusters	C1	C2	C3	C4	p_anova	sd
Hot carcass weight, kg	93,415	95,741	88,31	93,6	0,106	7,626
Back fat thickness, mm	15,385	15,059	13,9	15,286	0,689	3,159
Loin thickness, mm	62,308	60,824	55,9	59,643	0,096	6,327
% of meat in carcass, %	57,338	57,135	56,52	56,671	0,845	2,537
pH1	6,527	6,524	6,271	6,356	1,18E-002	0,241
pH3	6,278	6,231	6,141	6,21	0,642	0,251
pH48	5,488	5,626	5,452	5,503	1,79E-005	0,11
Driploss, %	4,878	3,087	5,292	5,603	2,99E-006	1,605
Intramuscular fat level, %	1,112	0,978	1,229	0,746	0,335	0,682
Glycolytic potential of LD muscle, µmol/g	143,651	124,86	161,999	147,512	6,07E-009	17,886
Colour_L	55,597	52,969	56,135	55,491	4,37E-004	2,374
Colour_a	15,568	15,512	15,671	15,624	0,983	1,052
Colour_b	6,299	5,065	6,187	5,996	6,92E-003	1,129
RTN, %	91,303	97,455	92,375	92,891	4,35E-002	6,57
Protein, %	22,778	22,689	23,088	23,311	2,60E-002	0,633
Drymass, %	26,253	25,905	26,903	26,25	4,14E-002	0,887
Water, %	73,747	74,095	73,097	73,75	4,14E-002	0,887
Leptin (ng/ml)	33,285	39,486	41,412	35,903	0,367	12,296
Triglycerides, (mg/100ml)	62,869	86,551	87,406	124,904	3,21E-006	34,374
Cholesterol HDL (mg/100ml)	35,987	33,117	34,163	36,349	0,571	7,098
Cholesterol LDL (mg/100ml)	48,356	45,962	46,251	35,41	3,11E-002	12,781
Total cholesterol (mg/100ml)	96,92	96,388	97,894	96,741	0,991	11,493
Glucose in serum (mg/100ml)	92,404	89,737	106,841	160,879	8,00E-010	38,742

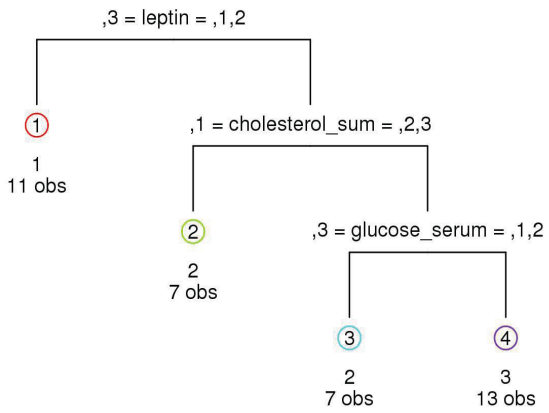

Fig. 1. Rpart tree classificator found for glycogen. Intermediate nodes were chosen from 6 attributes: cholesterol_sum, HDL, LDL, glucose_serum, leptin, triglycerides. All attributes were discretized into 3 levels: 1 – low, 2 – mean, 3 – high. Leaves are denoted by numbers in circles. Glycogen level values with numbers of related rows (observations) in a training set (70 percent of our dataset) are below circles

Table 3. Means for glycogen rpart tree nodes (Fig. 1) calculated on our full dataset. The classification error on full dataset is equal 0.4, but still majority of pigs fits in the tree schema

leaf	nr	IMF	GlyPo	Leptin	Total cholesterol	Glusose in serum
1	18	1,11	144,63	51,29	91,23	113,44
2	10	1,05	136,81	28,1	86,46	107,54
3	17	0,86	138,99	29,93	106,5	117,31
4	9	0,97	149,01	34,2	101,62	103,99

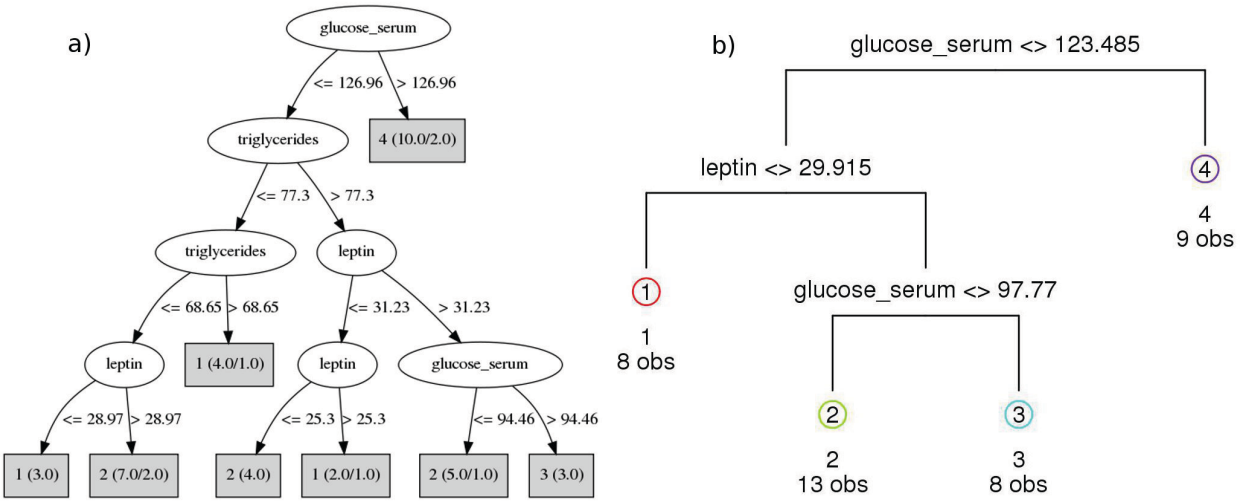

Fig.2 Cluster classification trees calculated with a help of R procedures a) J48 with a prunning parameter C equal to 1.5 b) rpart with default settings

In Fig. 2a from the blood serum parameter set of total cholesterol sum, cholesterol HDL, cholesterol LDL, glucose in serum, leptin, triglycerides, the attributes: glucose serum, triglicerides, leptin were chosen to find an appropriate cluster for a pig in the decision process through intermediate nodes up to leaves with final cluster numbers. A classification error on the entire data set was equal to about 20%. The same error was obtained by the best linear discriminant analysis classifier which involved all blood serum attributes after testing lda classifiers with all possible input variable combinations. In Fig. 2b only the glucose serum and leptin were in node tests, but a classification error on the entire data set was greater (about 30%). These errors could be smaller with greater trees or random forest consisting of many such trees voting for an answer. The pruned trees were chosen to show main inner dependencies between groups with different meat quality and blood serum measurements results.

As was shown in many research papers [Monin, 1985] the final meat quality depends from changes occurred in muscle after slaughter. The principal changes occurs in the glycogenolysis, that makes the basis for other changes. This depends from genetic and other factors. As was mentioned in introduction the determinism of meat quality is not fully understood but many factors are good determinants. During selection the metabolism of animals changes them. It is interesting (in case of breed when the genes bad influencing on meat quality are removed) what factors determine the quality of the meat. In this study we wanted to check whether two major components, which determine the technological and sensory quality of pork, glycogen and fat can be predicted on the basis of estimated level of lipids and glucose in blood serum. The results presented in Table 2 show that the group which was associated with lower quality of meat and a higher level of glycogen was characterized by higher levels of glucose in the blood serum and certain lipids. However, these dependencies are not clear. Since the cluster analysis are focused on many of element characteristics we decided to carry out an analysis based solely on metabolites determined in serum taken in vivo (Figs. 1 and 2). However, results show that the emerging four groups did not achieve enough significant differences in the level of glycogen and fat in meat. Thus, it is a need for further research.

5. CONCLUSIONS

Cluster analysis and regression decision trees proved that relationships between blood serum parameters taken in vivo and pork quality parameters measured after slaughter exist and are strong enough to enable good regression tree and cluster classifier creation. Obtained results showed significant associations and interactions between eating pork quality clusters and biochemical content of blood serum. In further research we try to describe more carefully in detail and with a help of new data and new function parameter tunning the mentioned pork attributes relationships.

REFERENCES

1. Barb C.R., Hausman G.J., Houseknecht K.L., 2001. Biology of leptin in the pig. Domest. Anim. Endocrinol. 21, 297-317
2. Bertram H.C., Petersen J.S., Andersen H.J., 2000. Relationship between RN genotype and drip loss meat from Danish pigs. Meat Sci., 2000, 56, 49-55
3. Bergmeyer H.U., 1974. Methods of Enzymatic Analysis. Academic Press, New York, pp. 1127, 1196, 1238, 1464
4. Dalrymple R.H., Hamm R., 1973. A method for the extraction of glycogen and metabolites from a single muscle sample. J. Food Technol. 8, 439-444
5. De Vries A.G., Faucitano L., Sosnicki A., Plastow G.S., 2000. The use of gene technology for optimal development of pork meat quality. Food Chem. 69, 397-405
6. Kaufman L, Rousseeuw PJ, 1990. Finding Groups in Data: An Introduction to Cluster Analysis. J. Wiley and Sons, Inc., New York.
7. Monin G. and Sellier P., 1985. Pork of low technological quality with a normal rate of muscle pH fall in the immediate post-mortem period: The case of the Hampshire breed. Meat Sci. 13, 49-63
8. Naveau J., Pommeret P., Lechaux P., 1985. Proposition d'une methode de mesure du rendement technologique: la „methode Napole" - Techni Porc 8, 7-13.
9. Prange H., Juggrt L., Scharner E., 1977. Untersuchungen zur Muskel Fleischqualität beim Schwein. Archiv. Experim. Vet. Med. 30, 2, 235-248
10. Przybylski W., Sieczko L., Jaworska D., Czarniecka-Skubina E., Niemyjski S., 2007. Estimation of conditionality of pork sensory quality by using multivariate analysis. Arch. Tierz. 50, 125-135
11. Przybylski W., Gromadzka-Ostrowska J., Olczak E., Jaworska D., Niemyjski S., Santé-LhoutellierV., 2009, Analysis of variability of plasma leptin and lipids concentration in relations to glycolytic potential, intramuscular fat and meat quality in P76 pigs. Journal of Animal and Feed Sciences, 18, 296-304
12. Przybylski W., Jaworska D., Czarniecka-Skubina E., Kajak-Siemaszko K., Wachowicz I., 2007, Using multidimensional analysis in the evaluation of technological value and sensory quality of pork meat. Pol. J. Food Nutr. Sci., Vol. 57, No. 4(B), pp. 449-455
13. Russell S., Norvig P., 2009. Artificial Intelligence: A Modern Approach (First Edition). Prentice Hall