Review article

Economics of Agriculture 4/2015
UDC: 311.2:631.1

QUALITY OF RESEARCH RESULTS IN AGRO-ECONOMY BY DATA MINING

Gordana Vukelić[1], Slobodan Stanojević[2], Zorica Anđelić[3]

Abstract

Data Mining (DM) through data in agroeconomy is a scientific method that enables researchers not to go through set research scenarioes that are predetermined assumptions and hypotheses on the basis of insignificant atributes. On the contrary, by data mining detection of these atributes is made possible, in general, those hiden facts that enable setting a hypothesis. The DM method does this by an iterative way, including key atributes and factors and their influence on the quality of agro-resources. The research was conducted on a random sample, by analyzing the quality of eggs. The research subject is the posibility of classifying and predicting significant variables-atributes that determine the level of egg quality. The research starts from the use of Data Mining, as an area of machine studies, which significantly helps researchers in optimizing research. The applied methodology during research includes analytical-sintetic procedures and methods of Data Mining, with a special focus on using Supervised linear discrimination analysis and the Decision Tree. The results indicate significant posibilities of using DM as an additional analytical procedure in performing agro-research and it can be concluded that it contributes to an improvement in effectiveness and validity of process in performing these researches.

Key words: *machine studies, data mining, prediction, classification, supervised discrimination analysis, decision tree, effectiveness of agro-research.*

JEL: *Q14*

Introduction

Application of Data Mining in the last ten years has brought to focus of scientific

1 Gordana Vukelić, Ph.D., Full Professor, Beogradska bankarska akademija, Fakultet za bankarstvo, osiguranje i finansije, Zmaj Jovina street no. 12, Belgrade, Republic of Serbia, Phone no: +381 11 263 58 23, E-mail: gordana.vukelic@bba.edu.rs

2 Slobodan Stanojević, Ph.D., Assistant Professor, Univerzitet privredna akademija u Novom Sadu, Fakultet za mandžment, ekonomiju i finansije, Belgrade, Nemanjina street no. 4, Belgrade, Republic of Serbia, Phone no: +381 62 644 358, E-mail: slobe_leo@yahoo.com

3 Zorica Anđelić, Master, VCC Akademija, E-mail: zorica.andjelic@vccsrbija.rs

researches in agro-economy a significant methodological turn. Classical methods that subsum the normative-descriptive methods supported by classical multivariate statistical methods have become a basis for establishing machine studies, as a more productive and exact scientific method in agro-research (Bohanec et al., 2003). The method of machine studies gives a clear positive answer on the question whether a computer can execute functions that we conside thought (Alan Turing, 1912-1954). Machine learning is a set of processes which include: collecting new declarative knowledge, development and specialization of motor and significant capabilities through practice, structuring existing knowledge and discovering new facts and theories by observing and active experimenting (Breiman, 2001). This learning process includes knowledge acquisition, ie. learning new information of symbolic character and training, which means improvement of acquired knowledge, which is a equal to the way people learn. Machine studies includes two models, *learning based on examples* and *learning by observation and own discovery.* They produce two types of scientific knowledge: explicite knowledge that is represented by mathematical logic, production rules and systems alike, which are the subject of authors application, and those are data mining, using different techniques of the decision tree, production rules of discrimination analysis, neuron network models etc (Cherkassky, Mulier, 2007).

Research in the domain of data mining relates to the area of inductive learning and performing general laws based on insight in specific occurences – cases. The procedure of applying these methods implies the grade of validity of learned knowledge, so it is necessary to methodologically divide the set on a *learning set*, which is used for learning and a *test set*, which is used for testing learned knowledge (Stanojević, 2013).

Usefull inductive knowledge must have *predictive accuracy,* the percentage of success of classifying new, examples of using learned rules which were not considered, done by grading accuracy in classifying the method of cross-validation and the bootstrap method (Kohavi, 1995). The method of machine learning, especially the area of finding hidden knowledge, or data mining includes an iterative process of discovering patterns, whether automatically or manually, in a surrounding where there are no predetermined assumptions or hypotheses, and that is the research goal (Stanojević, 2013).

In general, machine learning is a special area of *artificial intelligence*[4] relates to the development of algorithms and techniques which enable computers to "learn" (Platt, 1998). The method of inductive machine learning creates computer programs by extracting rules and patterns of behavior from sets of data. Data mining is also known as a process of *knowledge-discovery in databases (KDD) or knowledge-discovery and*

4 **Artificial intelligence (AI)** is the area in computer sciences that studies intelligent behavior, learning and machine adaptation. Research in the area of Artificial intelligence is related to machine production with the goal of task automation from required intelligent behavior.

data mining[5] (Chang, Lin, 2001). Data mining is defined as a process of recruiting one or more computer techniques with the goal of automatic analysis and extraction of knowledge from data which are found in a certain data base (Witten et al., 2011). The purpose of data mining is to find and identify certain patterns and trends in data. All methods of data mining are used for induction based learning (Kantardžić, 2002). That is the process of defining general conceptual definitions by observing specific examples from which learning is being done (Stanojević, 2013).

Within the research analysis of available data in the area of agro-research was conducted, by applying statistical method of Artificial Intelligence and DM of classified methods, and the goal is identification of rules of classification of data in the area of agriculture, poultry raising and analyzing egg quality (Birch et al., 2003). The research used two techniques of DM, and those are supervised in learning by using discrimination lineal analysis and the Decision Tree (Maindonald, Braun, 2007).

1. Research methodology

The research procedure starts from defined and well grouped data, determining variables and using chosen methods with analyzing and interpreting received results (Mihajlović, 2014).

2. Data

The authors' research relates to research results of identifying key factors on egg quality. Data was grouped in two samples, and those are: sample A – first quality category, and sample B – second quality category. The sample contains 25 participants, while sample B contains 33 participants.

2.1. Variable[6]

Four characteristic of egg qulity on both samples were measured, and those are: X_1= yolk shadow[7], X_2 = yolk color[8],

5 **KDD (Knowledge Discovery in Databases) -** Technically KDD is the application of using the scientific method of data mining. With the goal of performing data mining, a typical model of the KDD model includes the methodology of extracting and preparing data as well as making decisions on actions that need to be taken with the goal of analyzing and data mining.

6 The data are taken from a study made by A. Johnston, Poiltry Division, Centra Experiment Farm, Ottawa Cyril H. Gouldeb, „Methods of Statistical Analysis", New York, John Wiley and Sons, Inc1952.

7 Out of interior properties, the freshness of an egg is valued the most, determined by measuring the height of the air chamber and density of the egg white. An egg no older than three days has an immobile air, smaller than 4mm (which is determined by illuminating-yolk shadow).

8 Intensity of color of yolk determines quality. Intensive yellow color indicates on an egg with more quality. It is measured by the Roš fan. Grade 1 signifies the palest and 15 a yolk with most color.

X_3= height of egg white[9] , i X_4 = egg white index[10]. The mentioned variables are determined as predictive atributes. The research goal is to identify the key variable which influences the classification of eggs in A and B cathegory (cathegory or class which are goal variables in our research – class of atributes) (Breiman et al., 1984).

2.2. Methods

Identification of key variables from X_1,X_2, i X_3, i X_4 can be identified as a typical problem of classifying, and it occurs in two procedural phases. The first phase, machine model of learning is trained and used as a training sample. The samples is organized in rows and collumns. One of the atribute collumns ie. the class atribute dominantly influences the quality of eggs. This phase is called *Supervised learning* (Sohl, Venkatachalam, 1995). The second step of the model tries to classify objects which don't belong to the training sample.

The authors used Supervised linear discrimination function in the paper, with using validation methods of accuracy of classification as follows: *cross validation*[11] and the *bootstrap* method.[12]

2.2.1. Linear discrimination function (LDF)

The goal of aplying this statistical method is to determine useful variable for the purpose of classifying.

In the first step, the method of supervised learning through linear discrimination of the function with continuous variables, while the *predictory variable of cathegories is A or B.*

The results show that the performance of classification is with a mistake of 1.7%, while the variance mistake in relation to total Wilk"s Lambda (within the MANOVA method) is of a small value with sa p=0,0 [13].

9 Quality of egg white is graded by breaking and egg on a flat surface and measuring the height of dense egg white, which is expressed by Hog units (HU).

10 An egg of good eating quality, has a flat yolk of bigger radius and egg white which is watery and covers a big area.

11 The cross validation method or rotational estimation divides the set of examples D on k mutually excludable subsets D_1,D_2,...,D_k of approximately same size by accidental method.

12 The bootstrap method represents a family of methods for estimating precision of prediction For the assigned set *n* the application of bootstrap sample is formed by accidental choice *n* examples from the set of examples, by switching.

13 MANOVA is the statistical method in analyzing general linear model when there are multiple independent variables with the goal of seeing their interaction and identifying differences between the groups.

Table 1. Performance of classification

Status	Value	p-value
Wilks' Lambda	**0.2593**	**0,0000**
Bartlett -- C(4)	72.8853	0.0000
Rao -- F(4, 53)	37.8470	0.0000

Source: Goulden, 1952

Sum of the result of LDF can't be reliable because the essential question is not being asked, which is the relevant variable for the research. According to the following LD function, the discrimination equation would be as follows.

$$Z = 12{,}21\ X_1 + 6{,}584\ X_2 + 4{,}923\ X_3 - 67{,}83 + 8{,}423\ X_4 - 254{,}467$$

Table 2. Linear discrimination

	Classification function		Statistical valuing			
Attributes	**A**	**B**	**Wilks L.**	**Partial L.**	**F(1,53)**	**p-value**
Yolk shadow	**12.214336**	15.720853	0.586076	0.442453	66.78682	0.000000
Yolk color	**6.584722**	6.917973	0.264190	0.981530	0.99734	0.322492
Egg white index	**4.922898**	4.976528	0.259857	0.997897	0.11169	0.739542
Egg white level	**8.423514**	8.565462	0.261181	0.992839	0.38227	0.539041
Constant	**-251.467754**	-291.374507	-			

Source: Goulden, 1952

Results of the estimation of accuracy of classification with the bootstrap method.

From the table of results given above, it was determined that the actual percentage of error in predicting the quality of eggs was 3.5%.

Table 3. Error percentage in predicting egg quality

Error percentage	
.632+ bootstrap	**0.0348**

Source: Goulden, 1952

The next step is introducing STEPDISK[14] component, whose purpose is to assign the necessary number of variables for classifying affiliation to class A or B.

Table 4. Subset of selected attributes

N	Selected attribute
1	Yolk shadow

Source: Goulden, 1952

Table 5. Detailed results

N	Degree of freedom	The best	Sol.1	Sol.2	Sol.3	Sol.4	Sol.5
1	(1, 56)	**Yolk shadow** **L : 0.2663** **F : 154.26** **p : 0.0000**	**Yolk shadow** L : 0.2663 F : 154.26 p : 0.0000	**Egg white level** L : 0.6730 F : 27.20 p : 0.0000	**Egg white index** L : 0.7206 F : 21.71 p : 0.0000	**Yolk color** L : 0.8472 F : 10.10 p : 0.0024	-

Source: Goulden, 1952

According to the STEPDISK analysis results the only relevant attribute is the yolk shadow. The next step is control of effectiveness of the set model (Demšar et al., 2003). In that sense it is necessary to do the analysis of the supervised linear discrimination function of bootstrap method, whose error is 1.7%, which means it is less in relation to 3.5%, however only one variable occurs in the discrimination function, which is of crucial significance for determining the key factor in classifying within classes A and B

Table 6. Sum view of discrimination linear function

	Classification function		Statistical valuing			
Attributes	A	B	Wilks L.	Partial L.	F(1,56)	p-value
Yolk shadow	**7.565802**	10.951009	1.000000	0.266336	154.26036	0.000000
Constant	**-27.927139**	-57.310071	-			

Source: Goulden, 1952

Where the classification function is $\mathbf{Z= 7.65\ X_1\ -28}$

With a decrease in bootstrap error to 1.2%, as shown in the following table.

14 Stepwise discriminant analysis (STEPDISC) is discrimination analysis which determines relevant variables for the purpose of classifying by using WILKS' LAMBDA method. Wilks' lambda is a statistical test used in multivariation analysis of variance (MANOVA) for the purpose of testing the difference between mediums of identified groups of subjects of combinations of dependent variables.

Table 7. Error rate

Error rate	
.632+ bootstrap	0.1253

Source: Goulden, 1952

2.2.2. Application of the decision tree

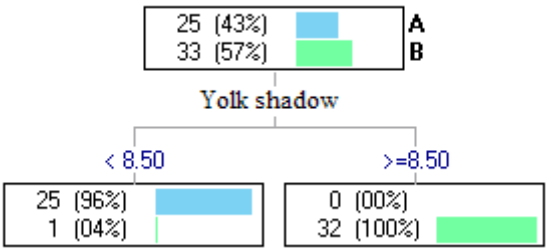
Next to discimination analysis, the research was done by the *Decision Tree*, by using C4.5 algorithm, which is based on the structure of the tree, where every leaf represents an atribute test and every branch a test result (Quinlan, 1996). The goodness of split is based on selection of atributes which are best separated in the sample.

Results with the classification accuracy of 9.6% give the following decision tree:

- **Yolk shadow < 8.5000 then Category (class) = A (96.15 % out of 26 examples)**
- **Yolk Shadow >= 8.5000 then Category (class) = B (100.00 % out of 32 examples)**

Or graphically:

Graph 1. Yolk Shadow



Source: Goulden, 1952

Estimating the accuracy of classification with the *Cross Validation* method we get that the error percentage is 0%.

Further steps in examining the value of classification include the application of bagging method (Sadok et al., 2009).[15] Integrally with this method we use the random tree bagging algorithm within the targets of the supervised learning, which gives 0% error or 100% accuracy (Đinović, 2013).

15 Bagging generates multiple versions classifications which are used as a whole, through the mechanism of voting. Multiple classifications are generated by using the bootstrap method. Every training set is an independent example sample, ie. some examples are excluded, while other are repeated. As well as other ensemble methods, the procedure is suitable for aggregation of work results of "unstable" algorithms, the relation of algorithms in which small changes in the whole set cause big changes in the learned set of rules.

Conclusion

Agro-economy faces great challenges, especially in the domain of research of not only the quality of ground, but also other food resources as well as sources of ecological food. The method of finding hidden knowledge has the assumption in relation to classical methods because they are more precise at classifying, as well as having greater predicting capacities.

The aim of this research was to examine the usefulness and exactness of these methods on the example of examining the presence of egg quality (category A and B) based on examining samples. The *Supervised Linear Discrimination Analysis* was used with the purpose of identifying the specific influence of variables on the quality of eggs with the variation method of accuracy in classifying the influence of variables and identifying the key variables, in this case it is enlightenment –shadow of eggs. Other than this method, the *Decision Tree* was used, which gave results which are more precise in relation of determining the level to which is the influence of certain variables. Given results are at the level of 99% precise, in relation to classical multivariate researches, this is the research where, by using supervised discrimination analysis, the influence of four variables on the presence of egg quality was revised, out of which three variables weren't the key for qualification. All that was needed for the research to come in the foreground was achieved, and that is great degree of accuracy of research (level of 99%).

Usage of this methodological apparatus was of significant help to researchers in the area of agriculture, especially due to the possibility that the research is done on scarce training sets which have a big number of attributes (the entity of the research subject, for example land, quality of agricultural products, fruit, vegetables, eggs, meat and many other) and a very small number of examples (so called scarce sets). The problem of scarceness is related to the evaluation of task difficulty, which in the domain of data mining is solved by reducing the number of attributes-variables. These methodological approaches enable revelation of, until now hidden knowledge in agro-economy and agronomy, and primarily on the causes that determine key-deciding variables and attributes and factors for solving research problems and the correct setting of a hypothesis, in the area of agro-economy, as well as in other areas of research.

Literature

1. Birch, A.N.E., Krogh, P.H., Cortet, J., Tabone, E., Griffiths, B.S., Džeroski, S., Wesseler, J., Gomot de Vaufleury, A., Badot, P-M., Andersen, M.N., Messéan, A. (2003): *Soil ecological and economic evaluation of genetically modified crops,* Poster at Biodiversity Implications of Genetically, ECOGEN, Vol. 51, pp. 171-173.
2. Bohanec, M., Džeroski, S., Žnidaršić, M. (2003): *Multi-attribute modelling of economic and ecological impacts of cropping systems,* September 7-12, 2003 Monte Verità, Ascona.
3. Breiman, L. (2001): *"Random Forests",* Machine Learning, Vol. 45, No. 1, pp. 5–32.

4. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984): *Classification and Regresssion Trees*, Wadsworth, Belmont.
5. Chang, C.C., Lin, C.J. (2001): *ACM Transactions on Intelligent Systems and Technology, (TIST* LIBSVM) A library for support vector machines, Vol. 2, No. 3, p 27.
6. Cherkassky, V., Mulier, F. M. (2007): *Learning from Data: Concepts, Theory, and Methods*, 2nd edition, John Wiley - IEEE Press, USA.
7. Demšar, D., Džeroski, S., Krogh, P.H., and Larsen, T. (2003): *Modeling microathropods and identifying the most important agricultural factors for the soil community of microathropods*, Proceedings of the International Electrotechnical and Computer Science Conference. Ljubljana, Slovenia.
8. Đinović, V. (2013): *Uticaj postpupka revalorizacije po finansijski položaj preduzeća,* Oditor, Belgrade, Serbia, No. 4, pp. 14-19.
9. Kantardžić, M. (2002): *Data mining: concepts, models, methods, and algorithms'*, John Wiley & Sons, Wiley—IEEE Press.
10. Kohavi, R. (1995): *"A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection"*, in Proc. of International Joint Conference on Artificial Intelligence, Vol. 14, No. 2, pp. 1137-1145
11. Maindonald, J., Braun, J. (2007): *Data Analysis and Graphics Using R,* 2nd Edition, Cambridge University Press, Cambridge, ISBN: 9780521762939
12. Mihajlović, M. (2014): *Menadžment znanja kao factor povećanja efikasnosti organizacije,* Oditor, Belgrade, Serbia, No. 9, pp. 33-36.
13. Platt, J. C. (1998): *Sequential minimal optimization: A fast algorithm for training support vector machines*, Technical Report MSR-TR-98-8, Microsoft Research.
14. Quinlan, J. R. (1996): *"Bagging, Boosting and C4.5"*, in Proc. of AAAI-96 Fourteenth national Conference on Artificial Intelligence, Portland, OR, AAAI Press, Menlo Park, CA, Vol. 1, pp. 725-730.
15. Sadok, W., Angevin, F., Bergez, J. É., Bockstaller, C., Colomb, B., Guichard, L., Reau, R., Doré, T. (2009). *Ex ante Assessment of the Sustainability of Alternative Cropping Systems: Implications for Using Multi-criteria Decision-Aid Methods-A Review*. In Sustainable Agriculture, pp. 753-767, Springer Netherlands.
16. Sohl, J. E., Venkatachalam, A. R. (1995): *A neural network approach to forecasting model selection,* Information and Management, Vol. 29, No. 6, pp. 297-303.
17. Stanojević, S. (2013): *Multivarijaciona analiza finansijskih izveštaja*, doktorska disertacija, Beogradska bankarska kademija, Fakultet za bankarstvo, finansije i osiguranje, Univerzitet UNION, Beograd.
18. Witten, I. H., Frank, E., Hall, M. A. (2011): *Data Mining: Practical Machine Learning Tools and Techniques,* Morgan Kaufmann Publishers.

KVALITET REZULTATA ISTRAŽIVANJA U AGROEKONOMIJI PRONALAŽENJEM IMPLICITNIH ZNANJA

Gordana Vukelić[16], Slobodan Stanojević[17], Zorica Anđelić[18]

Sažetak

Pronalaženje implicitnih znanja (Data Mining-DM) u podacima iz oblasti agroekonomije je naučna metoda, koja omogućuje istraživačima da ne polaze od postavljenih scenarija istraživanja gde su unapred odredjene predpostvake i hipoteze na bazi nesignifikantnih atributa. Suprotno tome, pronalaženje implitnih znanja ili utopljenih znanja (DM) omogućuje detektovanje onih atributa, generalnije onih skrivenih znanja koja omogućuju postavljanje prave hipoteze. DM metode to radi iterativnim putem utvrđivanja ključne attribute i faktora injihov uticaj na kvalitet agroresursa. Istraživanje je izvedeno na slučajnom uzorku analize kvaliteta jaja. Predmet istraživanja su mogućnosti klasifikacije i predikcije signifikantnih varijabli-atributa koje odredjuju nivoa kvaliteta jaja. Istraživanja u prilogu polazi od primene Data Mining, kao oblasti mašinskog učenja, koja značajno pomaže istraživačima u optimizaciji istraživanja. Primenjena metodologija u toku istraživanja uključuje analitičko-sintetičke procedure i metode Data Mininga, sa posebnim fokusom na primenu Nadgledane linearne diksriminacione analizu i Stabla Odlučivanja (Decision Tree). Rezultati indiciraju značajne mogućnosti primene DM kao dodatne analitičke procedure u obavljanja agroistraživanja i može se zaključiti da u istraživačkom postupku doprinosi poboljšanju efektivnosti i validnosti procesa obavljanja tih istraživanja.

Ključne reči: *Mašinsko učenje, data mining, predviđanje, klasifikacija, nadgledana diskriminaciona analiza, stablo odlučivanja, efikasnost agroistraživanja.*

16 Redovni profesor, dr Gordana Vukelić, Beogradska bankarska akademija, Fakultet za bankarstvo, osiguranje i finansije, Zmaj Jovina ulica br. 12, Beograd, Republika Srbija, Telefon: +381 11 263 58 23, E-mail: gordana.vukelic@bba.edu.rs

17 Docent, dr Slobodan Stanojević, Univerzitet privredna akademija u Novom Sadu, Fakultet za mandžment, ekonomiju i finansije, Beograd, Nemanjina ulica br. 4, Beograd, Republika Srbija, Telefon: +381 62 644 358, E-mail: slobe_leo@yahoo.com

18 Master, Zorica Anđelić, VCC Akademija, E-mail: zorica.andjelic@vccsrbija.rs