UDK 004.048

PREDICTION TECHNIQUES AND ECONOMIC BREEDING INDEX FOR ANALYZING MULTIDIMENSIONAL FEATURE VECTORS

Megel Y. E., Mikhnova O. D., Kovalenko S. M.

Kharkiv Petro Vasylenko National Technical University of Agriculture

Dairy farming is an important part of agriculture where farm owners have a necessity of predicting their annual profit. Economic breeding index is a combination of traits and sub-indices that aids in computing farmers' income. In the current paper, data mining and prediction techniques are briefly observed, which can be used for the aforementioned purposes. Regression model is built for predicting annual profit of a farm located in Kharkiv rural region. The proposed model can also be used for any site where many input variables are to be converted and presented in a simple form of linear regression equation with multidimensional feature vector.

Formulation of the problem. The importance of agricultural sector development is apparent for providing people with food. The agricultural sector mainly depends on farming. Dairy farming is an integral part of agriculture. Large population of cows and heifers are bought and sold every year to increase total profit by dairy herd replacements. Cows with high milk performance are only productive for 3 years, then they are sorted out and utilized for beef [1].

Thus, farmers need a breeding tool that will allow computing their profitability. One of such tools is EBI (Economic Breeding Index) that aids identifying the most effective cattle for dairy herd replacements. The aim of EBI is production of highly fertile cows with large amount of milk during the longest possible period of time. The following traits are considered:

– Production (milk, fat, protein);

– Fertility (calving interval, survival);

– Calving (direct calving difficulty, maternal calving difficulty, gestation length, calf mortality);

– Beef (cull cow weight, carcass weight, carcass conformation, carcass fat);

– Maintenance (cull cow weight);

– Management (milking time, milking temperature);

– Health (lameness, SCC, mastitis) [2].

While creating a new or altering existing herd, corrective mating should be performed. Selecting and breeding right pairs is essential for maximizing profit. Along with the aforementioned characteristics, grassland data analysis is made [1].

In this paper, the authors make an attempt of analyzing artificial intelligence techniques designed for data mining and knowledge discovery in application to agricultural data. The information retrieved after data processing is used for prediction of farmers' profit based on EBI. Experimental data were gathered from a dairy farm located in rural Kharkivska oblast region near Zmiiv city. The next two sections observe existing prediction methods and experimental results carried out on real agricultural data.

Prediction Models and Problem Statement. This section briefly describes mathematical models that can be used for agricultural data mining and prediction. In static models, the dependence of future values from the previous ones is specified in a form of some equation. For these purposes regression models (linear, multiple, nonlinear or generalized linear regression) can be used primarily along with auto-regression (ARIMAX, GARCH, ARDLM). Time series analysis, exponential smoothing, clustering and other structural models can also be utilized for solving the above mentioned problems [3, 4].

In the latter type of models (structural models), the dependence of future values from the previous ones is specified in a form of some structure and rules. Neural networks, Markov chains, classification and clustering methods are the examples of structural models. Aside from traditionally used prediction models, SVM (support vector machine), GA (genetic algorithm) and many others have recently gained large popularity. Figure 1 shows an example of zoomed-out graph constructed in accordance with the simplified regression model which was utilized for prediction of a farmer's profit based on input agricultural data.

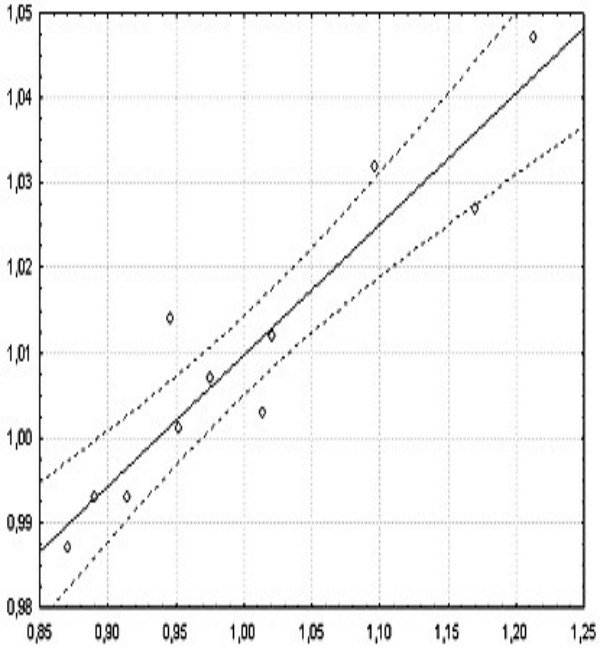


Figure 1 – Regression model for prediction of farmer's profit

Continuous values are predicted by statistical techniques of regression. Many nonlinear regression problems can be converted into linear form by transformation of variables that have an impact on the

model. The following equation of nonlinear regression can be converted into linear form using the least squares technique:

$$Y = \alpha X^{\beta} \quad (1)$$

For the linear regression model, data are presented as a straight line. Linear regression is the simplest form of regression. Bivariate linear regression allows creating a model for a random variable Y as a linear function of another random variable X:

$$Y = \alpha + \beta X \quad (2)$$

where $\alpha = \bar{y} - \beta \bar{x}$, $\beta = \frac{\sum_{i=1}^{s} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{s} \left(x_i - \bar{x}\right)^2}$.

he dispersion of Y, here, is assumed to be constant, and α and β are the regression coefficients that define the Y-intercept and slope of the graph line respectively. α and β are found via the least squares technique which minimizes the error between actual data and estimation of the graph line.

Multiple regression is an extension of linear regression with two and more predictor variables. Multiple regression can also be shrunk into a linear function of a multidimensional feature vector. Here, the least squares technique can be also applied. The following equation shows an example of a multiple regression with two predictor variables, X_1 and X_2:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 . \quad (3)$$

When modeling data without a linear dependence, nonlinear regression is considered by introducing polynomials to the linear model. Transforming the variables, nonlinear model can be converted into a linear one which may be further solved by the least squares technique.

$$Y = \alpha + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 \quad (4)$$

To convert this equation to linear form, new variables are assigned: $X_1 = X; X_2 = X^2; X_3 = X^3$. Now, equation (4) can be transformed into the linear form, resulting in the next formula:

$$Y = \alpha + \beta_1 X + \beta_2 X_2 + \beta_3 X_3 . \quad (5)$$

Though, some nonlinear models cannot be converted into linear ones. In such cases, least square estimates may be obtained using more complex formulae.

Unlike linear regression, in generalized linear models with constant dispersion of Y, the response variable dispersion is presented as a function of the mean value of Y. Generalized linear models are of two types: logistic linear regression and Poisson linear regression. Along with prediction, the log-linear model can be used to compress and smooth data. Data with a Poisson distribution can be modeled with Poisson regression [5, 6].

Experimental Results and Conclusion. The farm under analysis has three lactations with an average calving period that is equal to one year. Its average annual production of milk is 21 826 kg, 770.3 fat kg, 3.5 % fat, 829.2 protein kg, 4 % protein. In order to predict economic values (and thus, the farmer's profit) in future for this particular farm, a simulation model of a herd has been created, where the expenses for animals' food, veterinary, stuff salaries, etc., have been considered.

To perform the initial data analysis, all the input parameter values have been normalized. To shrink the number of variables, correlation analysis has been carried out. Figure 2 illustrates a fragment of pair-wise correlation between the variables under analysis. Figure 3 gives an example of predicted values.

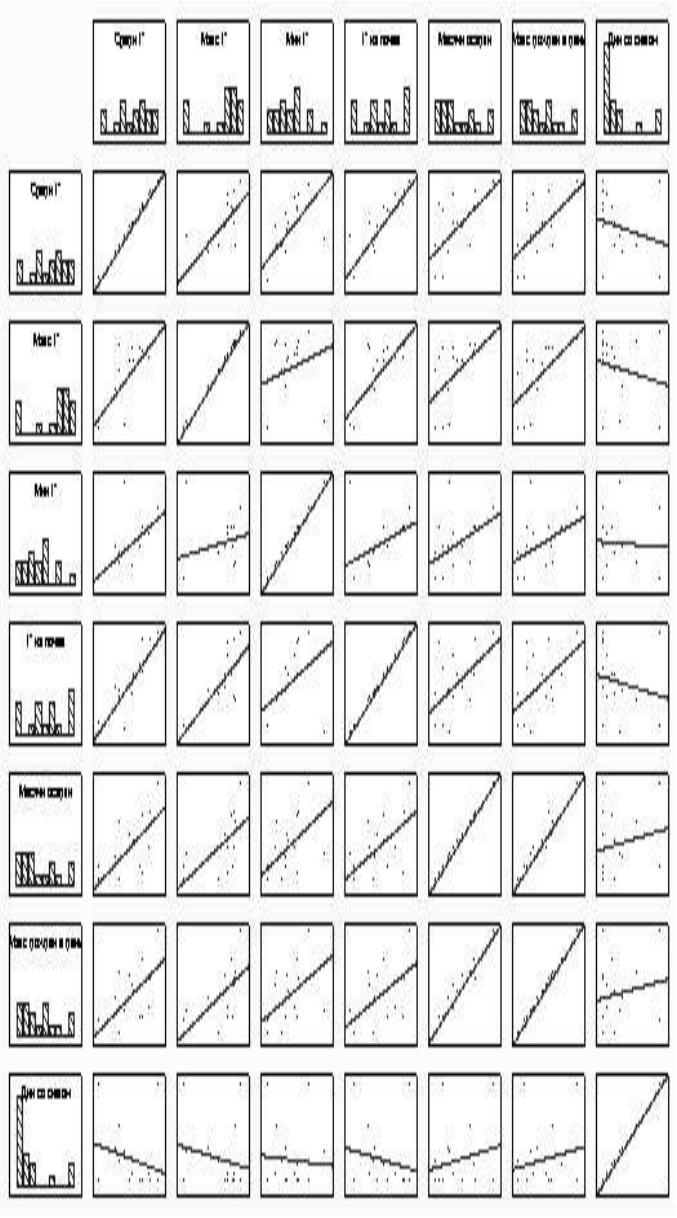
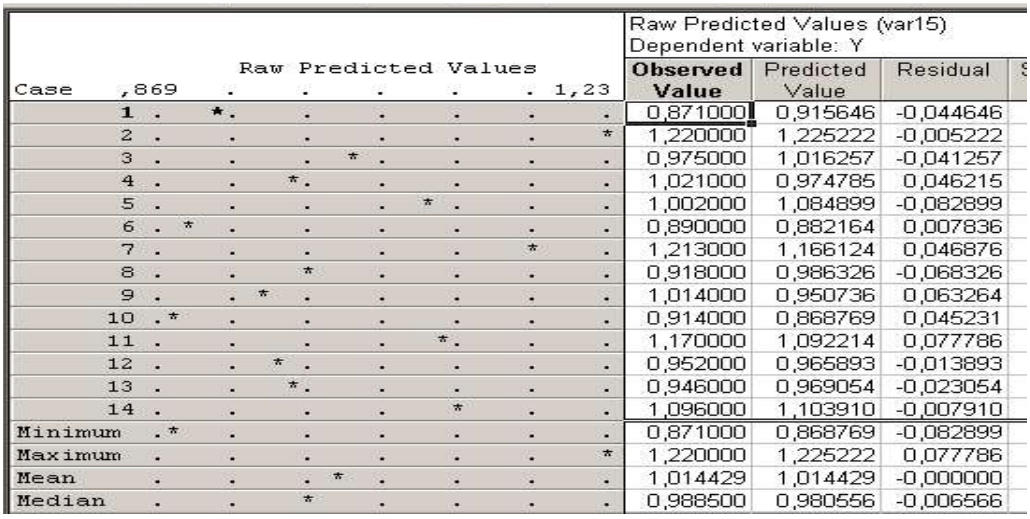


Figure 2 – Correlation of the attributes being under study



Case	Raw Predicted Values ,869 1,23	Raw Predicted Values (var15) Dependent variable: Y Observed Value	Predicted Value	Residual
1	. *.	0,871000	0,915646	-0,044646
2	. . . . . . *	1,220000	1,225222	-0,005222
3	. . . *	0,975000	1,016257	-0,041257
4	. . *.	1,021000	0,974785	0,046215
5	. . . . * . . .	1,002000	1,084899	-0,082899
6	. *	0,890000	0,882164	0,007836
7	. . . . . * .	1,213000	1,166124	0,046876
8	. . *	0,918000	0,986326	-0,068326
9	. . *	1,014000	0,950736	0,063264
10	.*	0,914000	0,868769	0,045231
11	. . . . *. . .	1,170000	1,092214	0,077786
12	. . *	0,952000	0,965893	-0,013893
13	. . *.	0,946000	0,969054	-0,023054
14	. . . . * . .	1,096000	1,103910	-0,007910
Minimum	.*	0,871000	0,868769	-0,082899
Maximum	. . . . . . *	1,220000	1,225222	0,077786
Mean	. . . *	1,014429	1,014429	-0,000000
Median	. . *	0,988500	0,980556	-0,006566

Figure 3 – Example of predicted values

The model under construction has nearly 15 input parameters. Correlation analysis has showed that only 9 of them have significant input to the model. With these data representing economical and biological relationships, the proposed mathematical model along with breeding index can be used for prediction of annual profit of a farm. Despite regression techniques emerged long time ago, they continue to be widely and effectively utilized for analyzing multidimensional feature vectors such as agricultural data in this case study. Unlike complex computational approaches described in the first section, the efficiency here is mainly achieved due to shrinking the initial number of input variables into the linear model.

References

1. Nevens W. B., Kuhlman A. F. Selecting Dairy Cattle. – Urbana : University of Illinois College of Agriculture and Agricultural Experiment Station. 1935. 48 p.
2. Peng L., Fang L., Yu P. Design of evaluation index system of ecological livestock breeding industry based on sustainable development. E-*Product E-Service and E-Entertainment* : proc. of 2010 intern. conf., Henan, 7-9 Nov. 2010. Henan : IEEE, 2010. P. 1-4.
3. Megel Y. E. et al. Matematichni modeli funktsionuvannya yekonomiko-virobnichikh i tekhnichnikh sistem ta metodi ikh doslidzhennya. Kharkiv : Miskdruk. 2013. 389 p.
4. Megel Y. E. et al. Operations research. Kharkiv : Miskdruk. 2015. 386 p.
5. Han J., Kamber M., Pei J. Data Mining: Concepts and Techniques (third edition). San Francisco : Morgan Kaufmann. 2011. 744 p.
6. Hill T., Lewicki P. Statistics: Methods and Applications (first edition). Tulsa : Dell. 2005. 800 p.

Аннотация

МЕТОДЫ ПРОГНОЗИРОВАНИЯ И ПОКАЗАТЕЛЬ ЭКОНОМИЧЕСКОГО ВОСПРОИЗВОДСТВА ДЛЯ АНАЛИЗА МНОГОМЕРНЫХ ВЕКТОРОВ ПРИЗНАКОВ

Мегель Ю. Е., Михнова Е. Д., Коваленко С. Н.

Молочное фермерство является важной составляющей сельского хозяйства. Владельцам ферм необходим интсрумент прогнозирования годовой прибыли. Экономический индекс воспроизводства представляет собой комбинацию признаков и вспомогательных индексов, которые помогают выполнить расчет прибыли фермера. Настоящая статья кратко описывает существующие методы интеллектуального анализа данных и прогнозирования, которые могут использоваться в этих целях. Для прогнозирования годового дохода фермы, расположенной в селе Харьковской области, построена регрессионная модель. Предложенная модель также может быть использована для объекта, где есть потребность преобразования нескольких входных переменных и их представления в более простой форме уравнения линейной регрессии с многомерным вектором признаков.

Аннотация

МЕТОДИ ПРОГНОЗУВАННЯ ТА ПОКАЗНИК ЕКОНОМІЧНОГО ВІДТВОРЕННЯ ДЛЯ АНАЛІЗУ БАГАТОМІРНИХ ВЕКТОРІВ ОЗНАК

Мегель Ю. Є., Міхнова О. Д., Коваленко С. М.

Молочне скотарство є важливою складовою сільського господарства. Власникам ферм необхідне інструмент прогнозування щорічного прибутку. Економічний індекс відтворення представляє собою комбінацію ознак та допоміжних індексів, які допомагають розрахувати прибуток фермера. Ця стаття коротко описує існуючі методи штучного аналізу даних і прогнозування, що можуть бути застосовані для цих потреб. Для прогнозування щорічного прибутку ферми, розташованої у селі Харківської області, побудована регресійна модель. Запропонована модель також може бути використана на об'єкті, де є необхідність перетворення кількох вхідних змінних та їх подання у більш простій формі рівняння лінійної регресії з багатомірним вектором ознак.