UDC 004.942+519.816

[1)]O. Kozhukhivska, [2)]P. Bidyuk, [3)]A. Kozhukhivskyi

[1)] Cherkassy state technological university,
Chair of informatics and information security
[2)] Institute of applied system analysis
of National technical university of Ukraine "KPI"
[3)] Cherkassy state technological university,
Chair of informatics and information security

MODEL BASED DECISION SUPPORT SYSTEM FOR FORECASTING FINANCIAL PROCESSES

Підвищення якості розв'язку задачі прогнозування фінансово-економічних процесів – актуальна задача, яка виникає на виробничих підприємствах, в інвестиційних компаніях, у банківській системі, транспортних компаніях та багатьох інших напрямах діяльності. Високоякісні оцінки прогнозів дають можливість підвищити якість рішень, що приймаються на їх основі. Незважаючи на те, що у цьому напрямі виконано велику кількість досліджень, існує необхідність розв'язання множини задач, спрямованих на прискорення та підвищення якості розв'язання задач такого класу. Зокрема, існує необхідність створення комп'ютерних систем підтримки прийняття рішень (СППР), орієнтованих на побудову високо адекватних математичних моделей та обчислення прийнятних за якістю оцінок коротко- та середньострокових прогнозів.

Ключові слова: модель, прогнозування фінансових та економічних процесів, принципи системного аналізу.

A computer based decision support system is proposed the basic tasks of which are adaptive model constructing and forecasting of financial and economic processes. The system is developed with the use of system analysis principles, i.e. the possibility for taking into consideration of some stochastic and information uncertainties, forming alternatives for models and forecasts, and tracking of the computing procedures correctness during all stages of data processing. A modular architecture is implemented that provides a possibility for the further enhancement and modification of the system functional possibilities with new forecasting and parameter estimation techniques. A high quality of final result is achieved thanks to appropriate tracking of the computing procedures at all stages of data processing: preliminary data processing, model constructing, and forecasts estimation. The tracking is performed with appropriate set of statistical quality parameters. Examples are given for modeling and forecasting of nonlinear and nonstationary financial and economic processes. The examples show that the system developed has good perspectives for the practical use. It is supposed that the system will find its applications as an extra tool for decision making when developing the strategies for enterprises of various types.

Key words: model, forecasting, financial and economic processes, system analysis principles.

Introduction

The forecasting problems are to be solved practically in all areas of human activities. However, the problems of mathematical modeling, estimation and forecasting process dynamics are particularly urgent for micro- and macroeconomics, banking sphere, insurance, investment companies, industrial enterprises that are functioning in conditions of tough competition, and many others kind of activities. There are many ideologically different approaches to mathematical description of processes dynamics and their volatility that are based on known statistical and recently developed techniques of intellectual data analysis.

Volatility forecasts are used widely as a measure of various kinds of financial risks in the frames of Value-at-Risk (VaR) and other methodologies. The market and some other types of risks are estimated with different modifications of VaR methodology that provides a possibility to reach practically acceptable quality of risk estimates [1, 2]. For forecasting financial processes and enterprises bankruptcy the nonlinear classification type models have found wide application, for example, logistic regression as well as support vector machine (SVM), fuzzy logic, neural networks and neuro-fuzzy techniques, Bayesian networks, decision trees, and combinations of the approaches mentioned [3 – 5].

Selection and application of a specific technique for process description and forecasts estimation depends on application area, availability of statistical data, qualification of personnel, who work on the financial analysis problems, and availability of appropriate applied software. Better results for estimation of financial processes forecasts is usually reached with application of ideologically different techniques combined in the frames of one computer based system. Such approach to solving the problems of quality forecasts estimation can be implemented in the frames of modern decision support systems (DSS). DSS is a powerful instrument for supporting decision making as far as it combines a set of appropriately selected data processing procedures aiming to reach final result of high quality – objective high quality alternatives for a decision maker. Development of a DSS is based on modern theories of system analysis, information processing systems, estimation theory, mathematical and statistical modeling and forecasting, decision making theory as well as many other results of theory and practice of processing data and expert estimates [6, 7].

The paper considers the problem of DSS constructing for solving the problems of modeling and forecasting processes evolution in time with the possibility for application of alternative data processing techniques, modeling and estimation of parameters and states for the processes under study.

Problem formulation. The purpose of the study is as follows: 1) analysis and development of requirements to the modern decision support systems; 2) development of the system architecture; 4) selection of mathematical modeling and forecasting techniques for financial and economic processes; 3) illustration of the system application to solving the problem of financial and economic processes forecasting with statistical data.

General requirements to modern DSS

Modern DSS are rather complex multifunctional computing systems with architecture of hierarchical type. Define DSS formally as follows:

$$DSS = \{DKB, PDP, DT, SE, PE, FG, DQ, MQ, FQ, AQ\},$$

where DKB – data and knowledge base; PDP – a set of procedures for preliminary data processing; DT – a set of statistical tests for determining possible effects contained in data; SE – a set of procedures for estimation of mathematical model structure; PE – a set of procedures for estimation of mathematical model parameters; FG – forecasts generating procedures; DQ, MQ, FQ, AQ – the sets of statistical quality criteria for estimating quality of data, models, forecasts, and alternatives, accordingly.

Such systems should satisfy the following general requirements: 1) – contain highly developed bases of data, mathematical models, quality criteria and rules, as well as necessary computational procedures; 2) – their interface should be user friendly, convenient and simple for use, as well as adaptive for the users of various levels (e.g., engineering and managerial staff); 3) – the hierarchy of a system functioning should correspond to the hierarchic process of human decision making; 4) – the system should possess an ability for learning in the process of its functioning, i.e. accumulate appropriate knowledge regarding possibilities of solving the problems of definite (selected) class; 5) – the organization and techniques for computing procedures should provide for appropriate rate of computing that corresponds to the human requirements with regard to the rate of alternatives generation and reaching the final result; 6) – computing (precision) quality should satisfy preliminary established requirements; 7) – intermediate and final results of computations should be controlled with appropriate sets of analytic quality criteria, what will enhance significantly quality and reliability of the final result; 8) – DSS should generate all necessary for a user forms and types of intermediate and final results representations with taking into consideration

the users of various levels; 9) – the system should contain the means for exchange with data and knowledge with other information processing systems via local or global computer nets; 10) – DSS should be easily expandable with new functions.

Satisfaction of all the requirements mentioned above provides a possibility for effective practical application of a system developed and enhancing general behavioristic effect of the DSS as a whole for a specific company or an enterprise.

General and special purpose mathematical tools for DSS

All mathematical methods that are hired for development and implementation of DSS could be divided into two following groups: – general purpose methods that provide for implementation of system functions, and special purpose methods that are necessary for solving specific problems regarding data processing, model constructing, alternatives generating, selecting the best alternative for implementation and forecasting of the implementation consequences.

The group of the general purpose methods includes the following methods: – data and knowledge collecting and editing procedures; – preliminary data processing techniques such as digital filtering, normalization, imputation of missing values, detecting special effects (regime switching, seasonal effects, trends etc); – the methods for accumulating information regarding previous applications of DSS to problem solving for the retrospective use; – computer graphics techniques; – techniques for syntactic analysis to be used in command interpreter; – methods for organizing communications with other information processing systems via local and global nets; – logical rules to control the system functioning. The set of the methods mentioned could be modified or expanded depending on specific application.

Selection of the application defined mathematical methods for a DSS depends on the specific system application area, possible problem statements regarding data processing, model building, processes forecasting, and alternatives generation. However, it is possible to state that in most cases of DSS development it is necessary to use the following mathematical methods: – methods and methodologies for mathematical (statistical and probabilistic) modeling using statistical/experimental data; – forecasting techniques on the basis of the models constructed with possibilities for combining the forecasts computed with different techniques; – operations research optimization techniques and dynamic optimization (optimal control) methods; – the methods for forecasting/foresight of decision implementation consequences; – the sets of special analytic criteria to control the processes of computations performed at each stage of data processing and alternative generation aiming to reach high quality of final results.

All the methods and methodologies mentioned are described well in special modern literature. For example, time series modeling and forecasting are presented in many references, more particularly in [8, 9]. The task for a DSS developer is in appropriate selection of model classes, modeling and optimization techniques, quality criteria as well as relevant methodologies for organizing computational processes.

Generation and implementation of alternatives with DSS

Decision making process includes rather sophisticated procedures that could be partially or completely iterative, i.e. executed repeatedly when the alternative found is not satisfactory for a decision making person (DMP). DSS can return automatically (or on DMP initiative) to the previous stages of data and knowledge analysis.

The whole process of making and implementing decision could be considered as consisting of the stages given below.

1 – A thorough analysis of the decision problem using all available sources of information, collection of data and knowledge relevant to the problem. At this stage it is also important to consider and use former solutions to the problem if such are available. The information regarding former solutions of similar problem can be helpful for correcting problem statement, to select appropriate techniques for data analysis, to speed up alternative generation, and to decline the alternatives that turned out to be ineffective in the past.

2 – Selection of a class (classes) of mathematical models for the problem description, and analysis of a possibility for the use of available (previously developed) models. The models could belong to different classes as far as they can be formulated in continuous or discrete time, be linear or nonlinear, they can be

developed according to the structural or functional approach etc. In some cases it is necessary to construct complex simulative model that would include a set of simpler models of various classes.

3 – Development of new models for the problem (process, object, system) under study what includes structure and parameter estimation for candidate models using available data (and possibly expert estimates) and knowledge of various types. The alternative structures of candidate models provide a possibility for selecting the best one of them for generating alternative decisions (forecasts, control actions, risk estimates etc) on their bases.

4 – Analysis of the candidate models constructed and selecting of the best one of them with application of a set of statistical quality criteria and expert opinion (estimation). At this stage again more than one model can be selected for the further use as far as the best model (for a particular application) can be found only after application of the candidates for solving particular problem, i.e. after alternatives generating and estimating possible consequences of their implementation.

5 – Application of the model (models) selected to solving forecasting and/or control problem (when necessary). If the forecasts or controls computed are not satisfactory we should return back to stage one or stage three, and repeat the process of model constructing. At this stage another set of statistical quality criteria should be applied to the analysis of forecasts or controls determined.

6 – Generating of a set of alternatives with the use of the model (models) constructed and various admissible initial conditions and constraints on variables. In a case of controls generating the alternatives could be built with different optimality criteria, utility functions or other criteria.

7 – Analysis of the alternatives generated with the experts of an enterprise or a company, and final selection of the best one for practical implementation. In a case when no alternative is acceptable we should return back to the model constructing or alternative generating stages. New knowledge or data can be required for the next iteration of computing the decision support.

8 – Planning of actions and estimation of financial, material and human resources that are necessary for implementation of the alternative selected. Determining of a time horizon (horizon of control) necessary for implementing the decision made.

9 – Implementation of the decision made: current monitoring of availability and spending the necessary resources, estimation of necessary time frames, registering and quality estimation of intermediate and final results.

10 – Application of possible analytic and expert quality criteria to estimation of final results.

11 – Analysis of the final results achieved by the company experts, and final elucidation of advantages and disadvantages of the alternative implemented; analysis of the decision making and implementing process, and forming forecasts (foresights) for the future.

12 – Writing the final report on the tasks performed.

Architecture of DSS for forecasting of financial and economic processes

DSS architecture is a generalized large-scale representation of system elements with links between them. Architecture gives a notion for the general purpose of system constructing and its basic functions (Fig. 1). DSS functionality is controlled by user commands, correctness of which is monitored by the command interpreter which constitutes a part of user interface. The user commands are implemented by the main operation module that coordinates functioning of all system elements. The specific commands and actions can be the following: expanding and modification of bases available in the system; initiation and starting of data and knowledge processing procedures; model constructing, forecasts estimation and alternative generating; viewing intermediate and final results of computing; retrospective analysis of previous results of decision making; comparing of current results with previous.

The system interface is considered as its basic element from the point of view of its presentation to the user. This is justified by the fact that interface construction influences substantially convenience as well as rate and effectiveness of user interaction with the system [10]. The principles of interface constructing and its implementation create a separate special task that is not considered here.

It is clear that architecture, given in Fig. 1, is highly generalized. It means that practically the same type of architecture could be used to construct DSS for solving rather wide class of problems that require statistical/experimental data processing, mathematical modeling, optimal state and parameter estimation for dynamic systems, forecasting the future process evolution and making decisions on this basis.

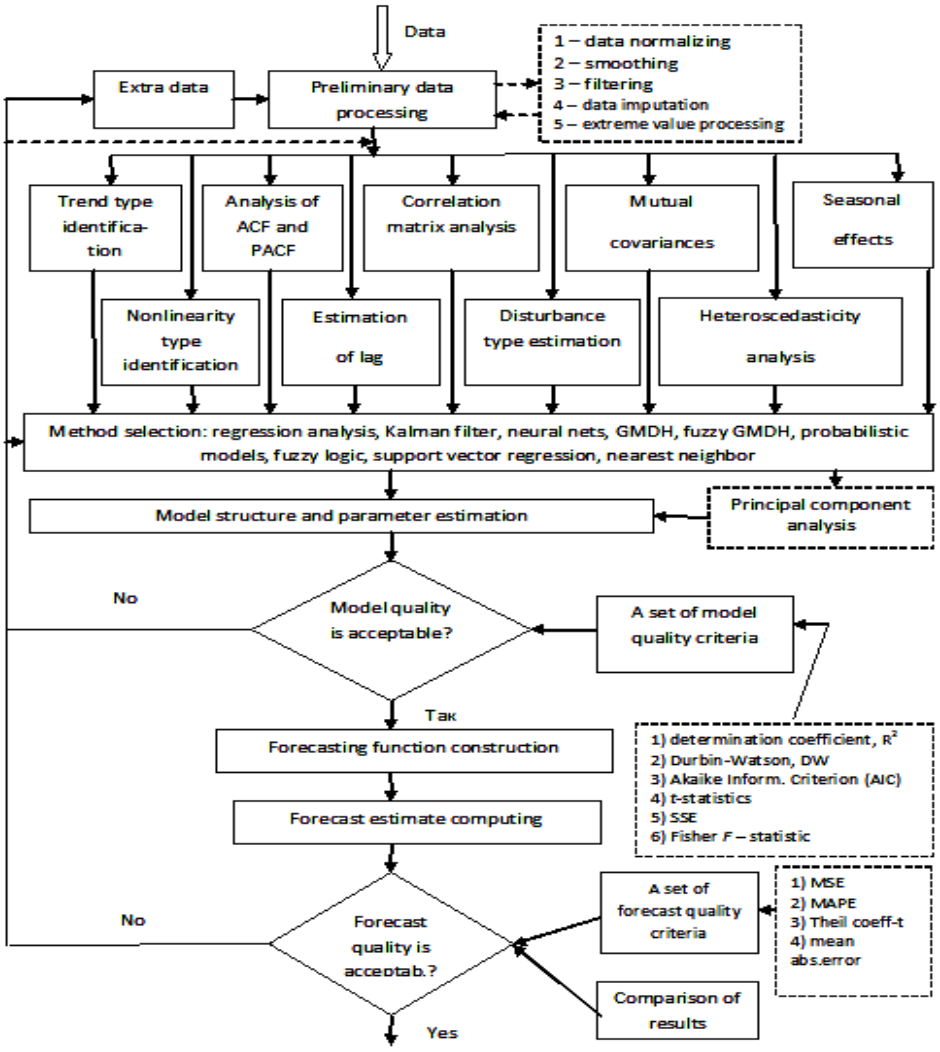

Fig. 1. DSS architecture for forecasting economic and financial process

Coping with uncertainties

The problem of identifying and taking into consideration various uncertainties is a very important one and is practically always present in decision making procedures. Here we do not consider the cases when uncertainties are taken into consideration with expert estimates. Though expert estimates are not excluded from the process of alternative generation. For example, expert estimate can be used for selecting special types of mathematical models that for some reason cannot be chosen automatically due to sophisticated structure or necessity to apply special estimation procedures. Expert judgment can also be useful for final selection of the best alternative from the set of generated decisions.

Statistical data uncertainties such as skipped measurements, extreme values and high level jumps of unknown origin could be processed with appropriately selected statistical procedures. There exist a number of data imputation procedures that help to make collected data complete. For example, very often skipped measurements for time series can be generated with appropriately selected distributions. Processing of jumps and extreme values helps with adjusting data stationarity and to estimate correctly probability distribution.

Application of Kalman filter requires knowledge of covariance for state disturbances and measurement errors. As far as these parameters are often unknown it is useful to apply appropriate adaptive estimation algorithms that provide acceptable estimates for the statistical parameters. An experience of practical application of the filter shows that it better to use separate procedure for covariance estimation to avoid divergence of filtering algorithm.

Fuzzy logic can be hired for coping with the level uncertainties for variables when we consider linguistic variables instead of numerical ones. There are possibilities for transforming fuzzy values into

numerical and vice versa. Thus, there is no problem for processing fuzzy and exact variables in the frames of one computing procedure.

Probabilistic types of uncertainties regarding whether or not some event will happen can be taken into consideration with various probabilistic models. Among them are analysis of distributions, Bayesian networks and other possibilities. From the computational point of view it is easier to process discrete variables as far as they accept a final number of values. In this case probabilities are assigned to outcomes using a probability mass function (PMF). Mass function tells us what "weight" (or a mass) should be assigned to each outcome. The sum for all the masses is 1,0. In a case of dealing with continuous variables, that may accept infinite number of values, we use probability density function (PDF). An integral over the density function should be equal to 1,0. When more than one random variable is considered we have to use joint distribution functions.

Generally speaking the modern instrumentation for coping with uncertainties is very powerful and it should be used in the frames of decision support systems for enhancing their possibilities with respect to reaching the best models and forecasts, and the best possible decisions.

Data, model and forecasts quality criteria

To achieve reliable high quality final result of forecasting at each stage of computational hierarchy separate sets of statistical quality criteria have been used. Data quality control is performed with the following criteria:

– database analysis for missing values using developed logical rules, and imputation of missed values with appropriate techniques;

– analysis of data for availability of outliers with special statistical tests, and processing of outliers to reduce their negative influence on statistical properties of data;

– normalizing of data in a case of necessity;

– application of low-order digital filters (usually that's low-pass filters) for separation of observations from measurement noise;

– application of principal component method to achieve desirable level of orthogonalization between the variables selected;

– computing of extra indicators for the use in regression models.

It is also useful to test how informative is the data collected. Very formal indicator for data informativeness is sample variance. It is considered formally that the higher is the variance the richer is data with information. Another criterion is based on computing derivatives with a polynomial that describes data in the form a time series. For examples, such polynomial may describe rather complex process trend as follows:

$$y(k)=a_0+\sum_{i=1}^{p} a_i y(k-i)+c_1 k+c_2 k^2+...+c_m k^m+e(k), \quad (1)$$

where $y(k)$ is the main variable; a_i, c_i are model parameters; $k=0,1,2,...$ is discrete time which is linked to real continuous time t via data sampling period T_s as follows: $t=kT_s$; $e(k)$ is a random process that integrates influence of external disturbances to the process being modeled as well as model structure and parameters errors. Autoregressive part of the model (1) describes the deviations that are imposed on a trend, and the trend itself is described with the *m*-th order polynomial. In this case maximum number of derivatives can be *m*, though in practice actual number of derivatives is defined by the largest number i of the parameter c_i, that is statistically significant.

To select the best models constructed the following statistical criteria are used: determination coefficient (R^2); Durbin-Watson statistic (DW); Fisher *F*-statistic; Akaike information criterion (*AIC*), and residual sum of squares (*SSE*). The forecasts quality is estimated with hiring the following criteria: mean squared error (*MSE*); mean absolute percentage error (*MAPE*); and Theil inequality coefficient (U). To perform automatic model selection the following combined criterion is proposed:

$$V_N(q, D_N) = e^{|1-R^2|} + \ln(1+\frac{SSE}{N}) + e^{|2-DW|} + \ln(1+MSE) + \ln(MAPE) + e^U, \quad (2)$$

where q is a vector of model parameters; $D_N - '$ is a dataset of power N. The power of the criterion was tested experimentally and proved with a wide set of models and statistical data. Thus, the three sets of quality criteria are used to insure high quality of final result.

Some mathematical models used in DSS

When considering mathematical models it is important to use a unified notion of model structure which we define as follows:

$$S = \{ r,\ p, m, n, d, z, l \},$$

where r is model dimensionality (number of equations); p is model order (maximum order of differential or difference equation in a model); m is a number of independent variables in the right hand side; n is a nonlinearity and its type; d is a lag or output reaction delay time; z is external disturbance and its type; l are possible restrictions for the variables.

Generalized linear models (GLM). GLM can be considered as further enhancement of multiple linear regression (MLR). It is distinguished from MLR with the following elements:

– distribution of dependent variable can be non-Gaussian and not necessarily continuous, say binomial;

– predicted values of dependent variable are computed as linear combination of predictors that are linked to dependent variable via selected link function.

Generally, GLM create a class of statistical models that includes linear regression, variance analysis relations, nonlinear models like logit and probit, Poisson regression and some others [11]. In a general linear model independent variable is supposed to be normally distributed and the link function is called identity function, i.e. linear combination of independent variables is not subjected to any transform.

In regression analysis the value of $h = \mathbf{X}\mathbf{b}$ is called a linear predictor, where $\mathbf{X}$ is independent variables measurement matrix; $\mathbf{b}$ is a vector of model parameters. When creating GLM instead of description in the form of $m = E[y]$ (where E is mathematical expectation; y is dependent variable as a function of linear predictor) some function $g(m)$ is described, i.e.

$$g(m) = h = \mathbf{X}\,\mathbf{b},$$

where $g(\cdot)$ is a link function. Thus, GLM is a model of the following type:

$$y = g^{-1}\left(\sum_{i=1}^{m} \mathbf{b}_i\, g_i(x) \right),$$

where m is a number of independent (explaining) variables. It is usually supposed that dependent variable y belongs to the class of exponential distributions. Thus, characteristics of GLM suppose the knowledge of dependent variable distribution, characteristics and parameters of the link function $g(\cdot)$, and of linear predictor $\mathbf{X}\,\mathbf{b}$. The class of exponential distributions includes the following distribution types: normal, gamma, and beta, and the discrete families – binomial, Poisson, and negative binomial. General representation of PDFs or PMFs for them is as follows:

$$f(x|q) = h(x)\, c(q) \exp\left(\sum_{i=1}^{k} w_i(q) l_i(x) \right),$$

where $h(x) \geq 0$ and $l_1(x),...,l_k(x)$ are real-valued functions of the observation x (they cannot depend on q); $c(q) \geq 0$ and $w_1(x),...,w_k(x)$ are real-valued functions of the possibly vector-valued parameter q (they cannot depend on x).

The following three basic functions are used in practice:

– logit:

$$h = \log\left(\frac{m}{1-m} \right);$$

– probit:

$$h = \Phi^{-1}(m),$$

where $\Phi(\cdot)$ – is a cumulative distribution function (CDF) for normal distribution;

– log-log function:

$$h = \log\{-\log(1-m)\}.$$

This type of link is of importance for short samples with positive mean value; it also can be presented as follows:

$$h = \frac{m^{l} - 1}{l},$$

with limiting value:

$$h = \log m \quad \text{as} \quad l \to 0;$$

or

$$h = \begin{cases} m^{l}, & l \neq 0, \\ \log m, & l = 0. \end{cases}$$

Such forms provide a possibility for performing correct computations when $l \neq 0$, and $l = 0$.

Nonlinear forecasting models logit and probit. To solve the problem of classifying credit borrowers into two groups it is quite logically to use appropriately transformed CDF. CDF belongs to the class of monotonous functions that monotonously decrease or increase on some interval. Suppose that for determining probability of crediting a client p_c it is chosen a normal distribution:

$$p_c = \Phi(\mathbf{b}^T\mathbf{x}) = \int_{-\infty}^{u} j(z)\, dz,$$

where $j(z)$ is a density for standard normal distribution; $u = \mathbf{b}^T\mathbf{x}$ is upper integration limit. This way so called probit model is constructed.

If the probability for successful crediting is determined with logistic distribution function then logit model is constructed. In this case we have:

$$p_c = \Phi(\mathbf{b}^T\mathbf{x}) = \int_{-\infty}^{u} j(z)\, dz = \frac{1}{1 + \exp(-\mathbf{b}^T\mathbf{x})}, \tag{3}$$

or

$$p_c = \frac{\exp(b_1 x_1 + ... + b_m x_m)}{1 + \exp(b_1 x_1 + ... + b_m x_m)}.$$

In contrast to the normal distribution logistic function has so called closed form that provides a possibility for simplified computations in comparison to probit. Parameter estimates for both models can be found with maximum likelihood technique. An alternative possibility is Markov chain Monte Carlo (MCMC) approach that is based on correct generation of pseudorandom sequences, that satisfy certain conditions. Due to availability of multiple alternative techniques for generating pseudorandom sequences MCMC has found wide applications [12]. Classification results achieved with logit and probit are usually acceptable in most cases of application.

Bayesian networks (BN). Bayesian networks are probabilistic and statistical models represented in the form of directed acyclic graphs (DAG) with vertices as variables of an object (system) under study, and arcs showing existing causal relations between the variables. Each variable of BN is characterized with complete finite set of mutually excluding states. The relations between the variables are established via expert estimates or applying special statistical and probabilistic tests to statistical data (when available) characterizing variables dynamics. The process of constructing BN is generally the same as for models of other types, say regression models. For example, as model parameters for BN are unconditional and conditional probabilities for specific values of variables, that are stored in respective tables. For parent variables these are unconditional probabilities and for daughter variables – conditional probability tables (CPT). Unconditional and conditional probabilities are determined by experts (in simpler cases), and by special computational algorithms when appropriate sets of statistical (or experimental) data are available. Thus to each node of DAG is assigned CPT that is used for computing probabilistic inference over the BN [5, 13].

The process of constructing a model in the form of BN can represented with the following steps: 1) – a thorough analysis of the process (object) under study aiming to detecting of its special functioning features and identification of parent and daughter variables; 2) – search and analysis of existing process models and determining the possibility of their usage in DSS; 3) – determining degree of relations between the process variables using special tests and expert estimates; 4) – reduction of the process dimensionality whenever this is possible; 5) – scaling and discretization of the data available when necessary; 6) – determining semantic restrictions on the future model; 7) – estimation of candidate model (directed acyclic graphs) structures using appropriate optimization procedures and score functions; 8) – candidate models analysis and selection of the best one using model quality criteria (including values of score functions); 9) – application of the model(s) constructed to solve the problem stated; 10) – computing inference with the model(s) constructed with regards to the variables selected, quality analysis of the result. In our case the final result of the model application is computing of client default probability with the conditions established by other model variables. According to alternative problem statement BM can be constructed for estimation of operational or other type of financial risks.

Support vector machine. Support vector machine (SVM) belongs to the group of techniques that determines classes with the limits for spaces. It also can be used for constructing SVM based regression models to solve forecasting problem. The support vectors are created with the vectors of data that lay on these limits. The classification result is successful if the space between the limits is empty. Usually SVM is hired to solve the problems of linear classification and regression analysis. The basic idea of SVM is in transformation of input vectors to the space of higher dimension with subsequent search of separating hyperplane with maximum distance in this space. Two parallel hyperplanes are built on both sides of separating hyperplane, and the separating hyperplane is the one that maximizes the distance between the two extra parallel hyperplanes. The algorithm is based on maximization of distance between the parallel hyperplanes what minimizes mean classification error.

The separating hyperplane is described with the equation [14]:

$$\mathbf{w}\,x - b = 0,$$

where $\mathbf{w}$ is a perpendicular to separating hyperplane; x is a normalized real-valued point of data; b is the shortest distance between the separating hyperplane and coordinate origin. The problem of separating hyperplane construction is in minimizing $\|\mathbf{w}\|$ under condition that $c_i(\mathbf{w}\,x_i - b) \ge 1,\ 1 \le i \le n$, where $c_i = 1$, if $(\mathbf{w}\,x_i - b) \ge 1$, and $c_i = -1$, if $(\mathbf{w}\,x_i - b) \le 1$. This is a problem of quadratic optimization of the form:

$$\begin{cases} \|\mathbf{w}\|^2 \to \min, \\ c_i(\mathbf{w}\,x_i - b) \ge 1,\ 1 \le i \le n. \end{cases}$$

This problem can be transformed to equivalent quadratic optimization problem that includes only dual variables:

$$-L(l) = -\sum_{i=1}^{n} l_i + \frac{1}{2}\sum_{i=1}^{n}\sum_{i=1}^{n} l_i\, l_j c_i c_j (\mathbf{x}_i \mathbf{x}_j) \to \min_l,$$

$$l_i \ge 0,\ 1 \le i \le n,$$

$$\sum_{i=1}^{n} l_i c_i = 0.$$

After solving the problem unknown $\mathbf{w}$ and b can be found using the expressions:

$$\mathbf{w} = \sum_{i=1}^{n} l_i c_i x_i\ ; \quad b = \mathbf{w}\mathbf{x}_i - c_i,\quad l_i > 0.$$

The classification itself is performed with the expression:

$$a(x) = sign\left[\left(\sum_{i=1}^{n} l_i c_i x_i\right)\mathbf{x} - b\right].$$

The summing is performed over for the support vectors only for which $l_i \ne 0$.

The problem solution considered is a simpler version of SVM. More sophisticated versions consider the cases of linear non-separation (when the two classes cannot be separated linearly) with weaker restrictions for inequalities, and the case of nonlinear classification with application of so called kernel trick.

Results of forecasting selected financial and economic processes

Below we show some results of forecasting economic and financial processes with the DSS proposed using linear regression models, nonlinear logistic regression, Bayesian networks and SVM. Table 1 shows results of forecasting inflation rate (on test sample) in Ukraine with GDP as predictor using autoregressive moving average and other models.

Table 1

Results of forecasting inflation rate

Model type	Model adequacy			Quality of one-step ahead forecasting			
	R^2	$\sum e^2(k)$	*DW*	*MSE*	*MAE*	*MAPE*	*U*
AR(1)	0,415	141,99	1,931	1,360	1,020	1,008	0,0067
AR(3)	0,317	135,14	1,992	1,360	1,020	1,011	0,0068
AR(7)	0,346	127,24	1,811	1,360	1,012	1,002	0,0067
AR(12)	0,435	97,80	1,941	1,337	1,020	1,013	0,0066
ARMA(1,1)	0,416	141,78	1,996	1,362	1,016	1,005	0,0067
AR(1) + M3	0,419	141,01	1,919	1,340	1,004	0,994	0,0066
AR(1)+GDP	0,419	141,05	1,916	1,335	1,004	0,993	0,0066
SVM	0,438	112,65	1,953	1,341	1,009	1,010	0,0067

Note: here *AR* is autoregression; *ARMA* is autoregression with moving average; R^2 is determination coefficient; $e(k)=y(k)-\hat{y}(k)$ is model error; *DW* is Durbin-Watson statistic; *MSE* is mean squared forecasting error; *MAE* is mean absolute error; *MAPE* is mean absolute percentage error; - *U* is a Theil coefficient; $M3$ is money aggregate; *GDP* is gross domestic product.

These results show that all the models constructed for inflation rate provide high quality short-term forecasting in spite of the fact that the values of determination coefficient R^2 are rather far from ideal. The last model constructed that includes GDP in the right hand side as a predictor showed the best result with $MAPE = 0,993$.

Table 2

Results of volatility forecasting for Microsoft stocks

Model type	Quality of historical forecasting			Quality of forecasting on test sample		
	MAE	*MAPE*	*U*	*MAE*	*MAPE*	*U*
ARCH	1,2756	2387,7	0,389	1,754	3241,2	0,453
GARCH	0,549	24,73	0,115	0,691	33,35	0,138
E-GARCH	0,436	3,45	0,017	0,073	4,75	0,022
SVM	0,517	5,35	0,031	0,295	7,64	0,039

Note: here *ARCH* is conditionally heteroskedastic autoregression; *GARCH* is generalized conditionally heteroskedastic autoregression; $E-GARCH$ is exponential generalized conditionally heteroskedastic autoregression.

Thus, he best result of volatility forecasting for the Microsoft stocks has been reached with the exponential generalized conditionally heteroskedastic autoregression (*MAPE* = 4,75% on test sample). Unacceptable quality of forecasting showed ARCH model with simplest structure (*MAPE* = 2387,7% on learning sample). It also can be seen that the values of mean absolute error and Theil coefficient are in accordance with the mean absolute percentage error that proves correctness of the experiment.

The methodology developed has also been applied to forecasting direction of moving for the stock prices. In this case we have used linear regression model, nonlinear logistic regression with extra market indicators, classification trees, Bayesian network, and some combinations of the models mentioned. The forecasting results for the price evolution direction are given in the Tables 3 and 4 for the maximum and minimum stock price selected.

Table 3

Results of maximum price direction forecasting for the stocks selected

Model type	Direction forecasting quality
Linear regression with indicators	68,95%
Logistic regression with indicators	69,76%
Classification tree with indicators	68,95%
Combination of linear and logistic regression	**74,19%**
Combination of linear regression and classification tree	69,48%
Bayesian network	72,38%

Table 4

Results of minimum price direction forecasting for the stocks selected

Model type	Direction forecasting quality
Linear regression with indicators	73,79%
Logistic regression with indicators	66,13%
Classification tree with indicators	64,66%
Combination of linear and logistic regression	**74,73%**
Combination of linear regression and classification tree	74,62%
Bayesian network	72,65%

Thus, in both cases (forecasting direction of movement for minimum and maximum price value) the best result was achieved with the combination of linear and logistic regression: probability of correct direction forecasting is 74,7 % and 74,19 %, respectively. The role of the linear regression model was to provide the price forecasts, and logistic regression predicted direction of the price evolution. The quality of the forecasts achieved is quite acceptable for their use in the rules of trading. High quality results have also been achieved with the Bayesian networks constructed on the statistical data available with the probability of correct direction forecasting 72,65 % and 72,38 %.

Conclusions

The methodology for constructing of decision support system for mathematical modeling of economic and financial processes that is based on the system analysis principles: hierarchical system structure, taking into consideration of probabilistic and statistical uncertainties, generating of decision alternatives, and tracking of computational processes for all the stages of data processing.

The system proposed has a modular architecture that provides a possibility for the easy extension of its functional possibilities with new model parameter estimation methods, forecasting techniques, and alternative generating. High quality of the final result is achieved thanks to appropriate tracking of the computational processes for all data processing stages: preliminary data processing, model structure and parameter estimation, computing of short- and middle-term forecasts, as well as thanks to convenient for a user intermediate and final results representation. The system is based on different (ideologically different) techniques of modeling and forecasting what creates a good base for combination of various approaches to achieve the best results. The examples of the system application show that it can be used successfully for solving practical forecasting problems.

The DSS can be used for decision making support in various areas of human activities including strategy development for industrial enterprises, transportation and investment companies etc. Further extension of the system functions is planned with new forecasting techniques based on probabilistic technologies.

1. McNeil A.J. Quantitative Risk Management / A.J. McNeil, R. Frey, P. Embrechts. – Princeton (New Jersey): Princeton University Press, 2005. – 538 p. 2. International Convergence of Capital Measurement and Capital Standards. A Revised Framework. Comprehensive Version. – Basel Committee on Banking Supervision, Bank for International Settlements. – Basel, 2006. – 158 p. 3. Mays E. (Ed.) Handbook of Credit Scoring / E. (Ed.) Mays. – Chicago: Glenlake Publishing Company, Ltd., 2001. – 460 p. 4. Neil M. Using Bayesian networks to model expected and unexpected operational losses / Neil M., Fenton N.E., Tailor M. // Risk Analysis. – 2005. – P. 34-57. 5. Zgurovsky M.Z. Method of constructing

Bayesian networks based on scoring functions / M.Z. Zgurovsky, P.I. Bidyuk, O.M. Terentyev // Cybernetics and System Analysis, 2008.- Vol. 44.- No.2.- P. 219-224. 6. Polovcev O.V. A System Approach to Modeling, Forecasting, and Management of Financial and Economic Processes / O.V. Polovcev, P.I. Bidyuk, L.O. Korshevnyuk. – Donetsk: Oriental Publishing House, 2009. – 286 p. 7. Hollsapple C.W. Decision support systems / C.W. Hollsapple, A.B. Winston. – Saint Paul: West Publishing Company, 1996. – 860 p. 8. Tsay R.S. Analysis of financial time series / R.S. Tsay. – Hoboken: Wiley & Sons, Inc., 2010. – 715 p. 9. Bidyuk P.I Methods of Forecasting / P.I. Bidyuk, O.S. Menyailenko, O.V. Polovcev. – Lugansk: Alma Mater, 2008. – 608 p. 10. Burstein F. Handbook of Decision Support Systems / F. Burstein, C.W. Holsapple. – Berlin: Springer-Verlag, 2008. – 908 p. 11. De Jong P. Generalized Linear Models for Insurance Data / P. De Jong, G.Z. Heller. – New York: Cambridge University Press, 2008. – 197 p. 12. Gilks W.R. Markov chain Monte Carlo in practice /W.R. Gilks, S. Richardson, D.J. Spiegelhalter. – New York: Chapman & Hall/CRC, 2000. – 486 p. 13.Jensen F.V. Bayesian Networks and Decision Graphs / F.V. Jensen, Th. D. Nielsen. – New York: Springer, 2007. – 457 p. 14. Smola A.J. A tutorial on support vector regression / A.J. Smola, B. Scholkopf // Statistics and computing, 2004.- Vol. 14.- P. 199–222.

УДК 519.7

V. Lytvynenko
Kherson National Technical University ,
Dept. of Informatics & Computer Science

HYBRID SWARM NEGATIVE SELECTION ALGORITHM FOR DNA-MICROARRAY DATA CLASSIFICATION

В роботі запропоновано метод класифікації. Він заснований на комбінованому алгоритмі негативної селекції, який був спочатку розроблений для задач бінарної класифікації. Точність розробленого алгоритму була перевірена експериментальним шляхом з використанням наборів даних мікрочіпів. Експерименти підтвердили, що напрямок змін, внесених в розроблений алгоритм підвищує точність у порівнянні з іншими алгоритмів класифікації.

Ключові слова: алгоритм вибору, класифікатор, мікрочіп даних, аналіз головних компонентів, імпульсне перетворення, відбір ознак.

In the paper, a classification method is proposed. It is based on Combined Swarm Negative Selection Algorithm, which was originally designed for binary classification problems. The accuracy of developed algorithm was tested in an experimental way with the use of microarray data sets. The experiments confirmed that direction of changes introduced in developed algorithm improves its accuracy in comparison to other classification algorithms.

Key words: Negative Selection Algorithm, Swarm Selection Algorithm, Classifier, DNA-Microarray Data, Principal Component Analysis, Wavelet transformation, Feature reduction, Feature selection.

1. Introduction

DNA microarray technology, introduced in 1995–1996, allows the measurement of thousands of gene expression values simultaneously, providing insight into the global gene expression patterns of cells (tissues) being studied [1,2,3]. Despite the need for further technological developments with microarray assays [4], the approach remains powerful for studying the myriad of transcription-related pathways involved in cellular growth, differentiation, and transformation in various organisms. In particular, the ability to measure thousands of gene expressions simultaneously using DNA microarrays has made it possible to investigate genome-wide objective approaches to molecular cancer classification[5]. Empirical