Optimization of the Bayesian classifier's structure using GMDH algorithm to forecast Internet-clients' preferences

Zhyliaev Sergey[1], Ryabokon Dmitriy[2]

[1]*Dept. of Cybernetics, National Taras Shevchenko University of Kiev, Glushkova 6, Kiev, Ukraine*
[2]*Dept. of Mathematical Methods of System Analyze, Polytechnic university of Kiev, av. Pobedy 37, Kiev, Ukraine*

us5eme@gmail.com, ryabokon_d@rambler.ru

Abstract. *The paper describes an algorithm allowing to raise accuracy of the Bayesian classifier due to optimization of its structure using the GMDH algorithm MULTI. Comparison of accuracy of proposed algorithm vs. regression algorithms is made on the basis of forecasting Internet-clients' preferences. The specific character of this problem lies in the fact that Internet-consumers have a number of attributes. Developed algorithm essentially differs from the Bayesian classifier in that it allows to find out few "strong", statistically stable attributes among large set of properties characterizing object and thus to increase the accuracy of classification.*

Keywords

Bayesian classifier, structure optimization, regression, GMDH

1 Introduction

Ability to establish regularities in statistical arrays is one of necessary skills of a system analyst. Independently of a problem domain (forecasting of economical indexes, estimation of production risks, identifying of potential clients), every analyst must be able to use own mathematical tools and software to solve forecasting problems, discover hidden regularities and recognize patterns. These tools are living essentials for them. Without these instruments, analyst cannot exercise his professional duties as builder cannot build a house without building implements.

The collection of mathematical tools and software described in the paper is meant for solving urgent problems of electronic commerce. While owners of usual supermarkets use old-established strategies of goods promotion [1], marketologists who develop advertising campaigns of electronic storefronts have not any universally recognized strategy because it is impossible to clarify for sure the characteristics of "remote" Internet-clients (gender, age, social status and so on). The largest information companies (Google, Yahoo, MSN) are in extreme need of methods for forecasting purchasing capacity of Internet-clients on the basis of their personal data. These methods must essentially differ from classical marketing techniques. As a result, the companies could give to clients not only search services but advertise tying products also.

The specific character of this problem lies in the fact that consumers have a number of attributes. Meantime, Internet-marketologists want to have not only an algorithm for classification of clients, but, in addition, they want to identify a collection of "strong" attributes that guarantees high accuracy of a forecast and allows to understand sources of this accuracy.

Our algorithms of forecasting of Internet-clients' purchasing capacity are based on the group method of data handling (GMDH) and the Bayesian decision theory. The twist of the paper is using of the Bayesian decision rule as a reference function of GMDH method. It allows constructing a forecasting model with discrete input and estimating a posteriori probability of purchasing for a client with known features.

2 Theoretical Part

2.1 Mathematical problem

Let K be a finite set of goods for sale and T be a m-dimensional set of attribute vectors of Internet-clients $(x_1, x_2, ..., x_m) \in T$, where $x_i \in \{0,1\}$. There is a map $F : T \to K$ mapping every buyer to a purchased product. All goods in K has the same cost. Attribute set consists of two disjoint subsets $T = L \cup E$, $L \cap E = \varnothing$. We shall referee to L as an experimental set and to E as an examining set. It is necessary to determine the goods purchased by customers from the examining set knowing goods purchased by the customers from experimental sets so that the number of mismatches would be minimal. In other words, it is necessary to construct a map $F_E : E \to K$ given a map $F_L = F\big|_L$, which is a narrowing of F on L under the condition $\left|\{x \in E \,|\, F(x) \neq F_E(x)\}\right| \to \min$.

Example:

Let a set of products consists of two elements $K = \{antivirus, systemdoctor\}$ and the number of dimensions m of the attribute space T equals to four.

The map $F : T \to K$ is specified by table I.

Tab.1. Example of map $F : T \to K$

Subset of T	**Attribute vector** $x \in T = L \cup E$	$F(x)$
L	(0,0,0,1)	*systemdoctor*
	(0,0,1,0)	*systemdoctor*
	(1,0,1,0)	*antivirus*
	(1,1,0,0)	*antivirus*
	(0,1,1,1)	*systemdoctor*
	(1,1,1,1)	*antivirus*
	(1,0,1,0)	*antivirus*
E	(0,0,0,0)	*systemdoctor*
	(1,1,0,0)	*antivirus*
	(1,0,0,1)	*antivirus*

It is necessary to determine a map $F_E : E \to K$ mapping customers with attributes vectors from the examining set E into the set of purchased goods given the first seven rows of Table I $F_L = F\big|_L$ so that the number of mismatches with real goods defined by the last three rows of Table I would be minimal.

2.2 Possible approaches to solution

The Bayesian classifier

Let $x \in T, k \in K$ and there exist joint probabilities $p(x,k)$ of purchasing product k by a client with an attribute vector x and a priori probability $p(k)$ of purchasing product k. Denote by $W(k,k^*)$ the function of seller's loss if he proposed product $k^* \in K$ to a client who wished to purchase product k. The Bayesian risk is defined by the following expression.

$$R(F_E) = \sum_{x \in T} \sum_{k \in P} p(x,k) W(k, F_E(x)). \quad (1)$$

The optimal Bayesian strategy is a mapping $F_E^* : E \to K$ that minimizes the risk R.

In considered problem all proposed products have the same cost. Thus, $W(k,k^*) = 1$, if $k \neq k^*$ and $W(k,k^*) = 0$ if $k = k^*$. Then

$$F_E^*(x) = \arg\min_{k^* \in K} \sum_{k \in K} p(x,k) W(k,k^*). \quad (2)$$

Let us execute some equivalent transformations of expression (2).

$$\begin{aligned} F_E^*(x) &= \arg\min_{k^* \in K} \sum_{k \in K} p(x,k) W(k,k^*) = p(x) \arg\min_{k^* \in K} \sum_{k \in K} p(k|x) W(k,k^*) = \\ &= \arg\min_{k^* \in K} \sum_{k \in K} p(k|x) W(k,k^*) = \arg\min_{k^* \in K} \sum_{k \in K \setminus \{k^*\}} p(k|x) = \\ &= \arg\min_{k^* \in K} \left(\sum_{k \in K} p(k|x) - p(k^*|x) \right) = \arg\min_{k^* \in K} (1 - p(k^*|x)) = \arg\max_{k^* \in K} p(k^*|x) \end{aligned} \quad (3)$$

The proposed products have similar functionality and destination, and they are sold to the same category of customers. That is why we may assume that a priori probabilities of their selling are equal to $p(k) = \frac{1}{|K|}, \forall k \in K$. This suggestion is confirmed by marketing investigations. In addition, we use suggestion about conditional independence of attributes in vector $x = (x_1, x_2, ..., x_m) \in T$ when product $k \in K$ is fixed, namely

$$p(x|k) = \prod_{j=1}^{m} p(x_j|k). \quad (4)$$

Note that this does not imply that these attributes are independent a priori. Using of (4) is forced as far as the number of experimental data required for estimation of $p(x|k)$ depends on the number of dimensions of attribute vector x as a power function. Namely due to using of conditionally independent components of attribute vectors in the model such Bayesian strategy is called naive. Granting (3) and (4), we may define

$$F_E^* = \arg\max_{k^* \in K} p(k^*|x) = \arg\max_{k^* \in K} \frac{p(x|k^*) p(k)}{p(x)} =$$

$$= \frac{1}{|K| p(x)} \arg\max_{k^* \in K} p(x|k^*) = \arg\max_{k^* \in K} p(x|k^*) = \arg\max_{k^* \in K} \prod_{j=1}^{m} p(x_j|k^*).$$

The probabilities $p(x_j|k^*)$ are determined on the experimental sample as a ratio of the number of clients purchased product k^* with j th attribute x_j to the number of all clients purchased product k^*. In other words,

$$p(x_j \mid k^*) = \frac{\left|\left\{x \in L \mid x_j \wedge F_L(x) = k^*\right\}\right|}{\left|\left\{x \in L \mid F_L(x) = k^*\right\}\right|}.$$

The naive Bayesian classifier has several deficiencies. First, it does not take into account the dependency between elements of the attribute vector x. Second, when the number of attributes is large, it is necessary to define a subset for forecasting or use the whole attribute set. In turn, this may worsen the accuracy of a forecast as far as for some attributes the priory probabilities may be undefined or the number of statistical data may be insufficient. Also, when the number of used attributes increases, the computational error of the conditional probabilities defined by formula (4) also increases. Third, even the accuracy of forecast is quite high, it is unclear what attributes facilitate the increasing of accuracy and what attributes don't influence on it at all. More detail description of the Bayesian decision theory may be found in [2].

Regression (method of least squares)

It is necessary to determine the dependency of real-valued output $F \in R$ on input parameters $x_j \in R$, $j \in \overline{1,m}$ having the form

$$F = \sum_{j=1}^{m} a_j f_j(x_j), \tag{5}$$

where $a_j \in R$ is unknown coefficients, $f_j : R \to R$ are known maps. In addition, we have output values F_i of n experiments with inputs x_{ij}, where $i \in \overline{1,n}$, $j \in \overline{1,m}$. Let us introduce the following denotations.

Y— column containing outputs with dimension $n \times 1$,

a — column of unknown coefficients from equality (5) with dimension $m \times 1$,

X — matrix with elements $f_j(x_{ij})$ and dimension $n \times m$.

Thus, we may form the following system

$$Xa = Y, \tag{6}$$

As a rule, this system is overdetermined, i.e. the number of equations is much greater than the number of unknowns. Further, we find the solution if this system using the Gauss method

$$X^T Xa = X^T Y, \tag{7}$$

minimizing the sum of squares of errors for system (6). We suppose that the systematical error of measurement of the output value equals to zero. If the determinant of matrix ($X^T X$) equals to zero the active experiment is named poorly defined. In particular, the necessary condition of inequality $\det(X^T X) \neq 0$ is $n \geq m$. Thus, the coefficients in (5) found by method of least square (MLS) are defined as

$$a = (X^T X)^{-1} X^T Y. \tag{8}$$

We can adapt MLS to our problem associating every product in K with a real number, for example, $0,1,2,..,|K|-1$, and determining the coefficients (5) by formula (8). Using the real-valued map $F(x), x \in T$, we can construct the final solution $F_E(x)$ taking values from the set $\{0,1,..,|K|-1\}$ when $x \in E$. To do this, we must select some real values θ_k in intervals $(k-1,k)$, $k \in \overline{1,|K|}$. Consider the family of functions $\varphi(x,\theta_1,..,\theta_{|P|})$ with parameters θ_1, θ_2, .., $\theta_{|P|}$

$\varphi(x,\theta_1,..,\theta_{|P|}) = k \Leftrightarrow F(x) \in (\theta_k, \theta_{k+1}], x \in T$, $\theta_k \in (k-1,k), k \in \overline{1,|P|}$.

Let us referee θ_k^* as an optimal threshold if

$$\left|x \in L \mid F_L(x) \neq \phi(x,\theta_1,..,\theta_k^*,..,\theta_{|P|})\right| = \min_{\theta_k}\left|x \in L \mid F_L(x) \neq \phi(x,\theta_1,..,\theta_k,..,\theta_{|P|})\right|.$$

Then, the desired solution of the problem is $F_E(x) = \phi(x, \theta_1^*, .., \theta_{|P|}^*)\big|_E$.

Every optimal threshold θ_k^* is determined independently from others.

The deficiency of the method described above is using of the Euclidian metrics in set K . For example, in the MLS-criterion product "1" differs from product "3" much more from product "2". In addition, as in the case of the naive Bayesian classifier, it is unclear what attributes facilitate the increasing of accuracy and what attributes decrease it or don't influence on it at all.

GMDH method

The group method of handling is applied for selection of investigated model with optimal structural complexity given functional basis [3] and for discovering of regularities on the basis of small number experimental data. The GMDH method allows to clearly identify the set of significant attribute ultimately facilitating to increasing of the simulation accuracy. For example, the complexity of a polynomial is defined by the number of monomial from which it is consisting of.

The experimental set is divided into two parts: training and examining. Then, on the basis of the training sample we adjust parameters of model with fixed structure using some "internal" criterion. The examining part allows counting quantitative characteristics of required attribute. This constitutes so called "external" criterion. The structure of model is considered as optimal given functional basis if it minimizes external criterion. If several models minimize the criterion simultaneously than we prefer the simplest one. For example, if it is necessary to determine the polynomial dependency of output on input attributes then the MLS-criterion may be used as an internal criterion and the error of model on a training sample may be used as an external criterion. External criteria and their appropriateness in different situations are described in [4].

There is a number of GMDH algorithms allowing to find out a model with optimal structure. These algorithms were developed to solve different applied problem. They are optimized with respect to space and time resources needed for their computer implementation [3-5]. Consider two algorithms used to solve problem formulated in the paper.

Combinatorial multistage algorithm MULTI

Detailed description of this algorithm and its properties may be found in [6, 7]. The search of model with optimal structure consists of several stages. At the first stage we construct all models which depend on unique argument (attribute).

$$y = f_1(x_i), i \in \overline{1,m} .$$

Then, among these models we select l models that minimize the external criteria. Arguments $x_{i_h}^{(1)}, h \in \overline{1,l}$ included in selected models we use to construct models of the second stage. These models depend on two arguments

$$y = f_2(x_{i_h}^{(1)}, x_i), h \in \overline{1,l}, i \in \overline{1,m} .$$

Among constructed model we again select "the best l models". Now, the pairs of attributes $(x_{i_h}^{(2)}, x_{i_p}^{(2)}), h \in \overline{1,l}, p \in \overline{1,l}$ used for construction of the best models of the second stage will create the basis for models construction at the third stage. At this stage we consider dependencies having the form

$$y = f_3(x_{i_h}^{(2)}, x_{i_p}^{(2)}, x_i), h \in \overline{1,l}, p \in \overline{1,l}, i \in \overline{1,m}$$

and so on. Building-up of the stages continues up to that moment when the external criterion computed on the best model of the stage begins to decrease. The maximal possible number of stages is equal to the number of input parameters given to investigator. For sufficiently large l this algorithm provides exhaustive search of all models in the given basis. Identification of dependency of output value given linear basis with MLS as an internal criterion and the error of model on the training sample as external ones is a representative example of application of this method. In this case the partial description of the stages has the following form:

$$y = a_0 + a_1 x_i, i \in \overline{1,m},$$

$$y = a_0 + a_1 x_{i_h}, + a_2 x_i, h \in \overline{1,l}, i \in \overline{1,m},$$

$$y = a_0 + a_1 x_{i_h}, + a_2 x_{i_p} + a_3 x_i, h \in \overline{1,l}, p \in \overline{1,l}, i \in \overline{1,m} \text{ etc.}$$

Iterative multilayered GMDH algorithm

As before, this algorithm means the construction of series of models. At the first layer we construct the models of the form

$$y = f(x_i, x_k), i \in \overline{1,m}, k \in \overline{1,m}.$$

Then among these models we select l "best models", i.e. the models minimizing the external criteria y_j, $j \in \overline{1,l}$. Further we construct the second layer

$$z = f(y_i, y_k), i \in \overline{1,l}, k \in \overline{1,l}$$

To avoid "the multilayered error", at the second and subsequent layers we use as arguments "the best models" of the previous layer and their attributes [5]. Building-up of the layers is stopped when the external criterion computed on the best model of the layer decreases. Function f is called reference, or partial description. Selection of this function closely tied with determining of functional basis. Examples of reference functions are the function with covariance and the quadratic partial description

$$f(x_1, x_2) = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_1 x_2, \quad (9)$$

$$f(x_1, x_2) = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_1 x_2 + a_4 x_1^2 + a_5 x_2^2.$$

These functions allow establishing dependence of simulated parameter in the form of high-order polynomial without a priori information about its structure. The iterative multilayered GMDH algorithm allows establishing complex dependencies using small number of experimental data because every partial description depends on few parameters.

The Bayesian risk as an internal criterion of GMDH

The Bayesian classifier differs advantageously from the MLS in that it does not use a distance between products because this concept is an artificial notion in a finite set of selling goods. However, as mentioned above, using of the Bayesian classifier requires very large number of experimental data to determine a posteriori probabilities $p(x = (x_1, x_2, .., x_m) | k)$. When the number of dimensions of an attribute vector x increases, this number increases as a polynomial function. To solve this problem, we may apply formula (4). But, in this case it is unclear what attributes x_j, $j \in \overline{1,m}$ must be used for classification, because groups of conditionally dependent attributes in combination with formula (4) worsen forecast. Also, there are attributes not having a priori probabilities at all. Such attributes must be excluded from the set of classifying ones. We propose to optimize structure of the naive Bayesian classifier, i.e. to select the set of attributes for predicting with the help of the GMDH algorithm MULTI. In this case the Bayesian risk (1) is used as an internal criterion and the number of erroneous outcomes in training samples is used as external criterion. There is an alternative approach to accounting large numbers of attributes when computation of a priori probabilities $p(x = (x_1, x_2, .., x_m) | k)$ is impossible due to lack of experimental data. This approach is based on the iterative multilayered GMDH algorithm with the reference function

$$f(x_1, x_2) = \arg\min_{k^* \in K} \sum_{k \in K} p(x_1, x_2, k) W(k, k^*)$$

In our case this function has the form

$$f(x_1, x_2) = \arg\max_{k^* \in K} p(x_1, x_2 | k^*). \quad (10)$$

Thus, every partial description in the next layer uses outputs of the previous layer as input parameters. The final solution is expressed as a net.

More detail description of this approach including theorem on convergence to solution corresponding to the minimal Bayesian risk together with practical examples of its application may be found in [8].

2.3 Results

To solve practical problem stated in the paper we used the following algorithms: the Bayesian classifier, the GMDH algorithm MULTI with a linear reference function, and the iterative multilayered GMDH algorithm with a linear, covariance (9), and the Bayesian (10) reference functions. Also, for the first time we used optimization of the structure of the Bayesian classifier with the help of the GMDH algorithm MULTI. All algorithms mentioned above were tested on three samples of buyers with different number of selling products. Sample I contains the attributes of Internet-clients who purchased one of two goods $\{antivirus, systemdoctor\}$, sample II contains the attributes of buyers who purchased three goods $\{antivirus, systemdoctor, drivecleaner\}$, and sample III contains the attributes of buyers who purchased four goods $\{antivirus, systemdoctor, drivecleaner, errorsafe\}$. The experimental sets for every product consist of attributes of 200 buyers and the examining parts consist of attributes of 100 buyers. In every test every buyer has 1281 attributes. Next, we show comparative tables of accuracy reached in recognition of the buyer preferences in every experiment.

Tab.2. Results

Sample	Number of products	Algorithm	Accuracy, %
I	2	the Bayesian classifier	72
		the GMDH algorithm MULTI with a linear reference function	85
		the iterative multilayered GMDH algorithm with a linear reference function	82
		the iterative multilayered GMDH algorithm with a covariance reference function	83
		the iterative multilayered GMDH algorithm with a Bayesian reference function	83
		Bayesian classifier optimized by GMDH MULTI	86
II	3	the Bayesian classifier	66
		the GMDH algorithm MULTI with a linear reference function	66
		the iterative multilayered GMDH algorithm with a linear reference function	64
		the iterative multilayered GMDH algorithm with a covariance reference function	66
		the iterative multilayered GMDH algorithm with a Bayesian reference function	78
		Bayesian classifier optimized by GMDH MULTI	76
III	4	the Bayesian classifier	45
		the GMDH algorithm MULTI with a linear reference function	43
		the iterative multilayered GMDH algorithm with a linear reference function	42
		the iterative multilayered GMDH algorithm with a covariance reference function	44
		the iterative multilayered GMDH algorithm with a Bayesian reference function	52
		Bayesian classifier optimized by GMDH MULTI	54

3 Conclusion

The purpose of the paper is to develop the algorithm for solving the problem of classification of Internet-clients in accordance with preferred goods given statistical data about their personal computers. As possible approaches to solution we investigated the following algorithms: the Bayesian classifier, the GMDH algorithm MULTI with a linear reference function, the iterative multilayered GMDH algorithm with a linear, covariance (9) and the Bayesian (10) reference functions. Every

algorithm must satisfy the following requirements: taking into account specific character of the problem, which is expressed in impossibility of comparison among products and, as a result, incorrectness of introducing of the Euclidian metrics, and necessity to point out few “strong” attributes providing high accuracy of classification.

The novelty of the paper consists in developing the new algorithm of optimization of the Bayesian classifier with the help of the GMDH algorithm MULTI, which essentially differs from the Bayesian classifier in that it allows to find out few “strong”, statistically stable attributes among large set of properties characterizing object and thus to increase the accuracy of classification.

The results of experiment on forecasting buyer preferences of Internet-clients confirm appropriateness and advantages of the Bayesian decision theory over regression methods. The more number of goods laying in the basis of classification, this advantage is more obvious. This is explained by the fact that regression methods use the Euclidian metrics introduced in the set of goods K. For example, in the MLS-criterion product “1” differs from product “3” more than from product “2”.

Optimization of the Bayesian classifier with the help of the GMDH algorithm MULTI allows both to raise accuracy of recognition by 9–14% and to identify attributes influencing on preferences of buyers. If exclude any attribute from the model defined by the algorithm MULTI then accuracy of forecasting decreases. This confirms that the algorithm MULTI select only “useful” attributes.

On the whole, the optimized Bayesian classifier is an alternative to using the iterative GMDH algorithm with the Bayesian reference function to make decision concerning classifying an object with large set of attributes.

References

[1] Kotler, F., Osnovy marketinga (Marketing). – Moskva: Progress, 1990. – 736 p. (In Russian)

[2] Schlesinger, M. and Hlavac, V. Desyat lekciy po statisticheskomu i strukturnomu raspoznavaniyu (Ten Lectures on Statistical and Structural Pattern Recognition). – Kiev: Naukova dumka, 2004. – 535 p. (In Russian)

[3] Ivakhnenko, A.G. Induktivnyi metod samoorganizatsii modelei slozhnykh sistem (An Induktive Method of Self-Organization of Models of Complex Systems). – Kiev: Naukova Dumka, 1987. – 296 p. (In Russian)

[4] Ivakhnenko, A.G. and Stepashko, V.S. Pomekhoustoichivost’ modelirovania (Noise Immunity of Modeling). – Kiev: Naukova Dumka, 1985. – 216 p. (In Russian)

[5] Ivakhnenko, A.G. and Müller J.A. Samoorganizatsiya prognoziruyuschich modeley (Self-Organization of forecasting models). – Kiev: Tekhnika, 1985. – 223 p. (In Russian)

[6] Stepashko V.S. Finite selection procedure for obtaining the exhaustive search result. – Soviet Automatic Control, 14, 3 (1981). – P.18-27.

[7] Stepashko, V.S., and Kostenko, Yu.V. Issledovanie svoystv kombinatorno-selektsionnogo (mnogoetapnogo) algoritma MGUA (Research of properties of combinatorial-selective (multistage) GMDH algorithm). – Modeling and managing of regional ecological and economical systems state – Kiev, 2001. – P.96-100. (In Russian)

[8] Ivakhnenko, A.G., Zaychenko, Y.P., and Dimitrov, V.D. Prinyatie resheniy na osnove samoorganizatsii (Decision Making on the Basis of Self-Organization). – Moscow: Sovetskoye Radio, 1975. – 280 p. (In Russian)