ICSQP '95

Proceedings of International Conference on Statistical Methods and Statistical Computing for Quality and Productivity Improvement

Edited by
Scientific Program Committee

Volume II

Contributed Papers

August 17(Thu.) - 19(Sat.), 1995
The Swiss Grand Hotel, Seoul, Korea

International Conference on Statistical Methods and Statistical Computing
for Quality and Productivity Improvement
August 17-19, 1995, Seoul, Korea

BOOTSTRAP INFORMATION TECHNOLOGY AND DESIGN OF INTELLIGENT INFERENCE MACHINE

LEONID PESHES, ZIGMUND BLUVBAND
A.L.D. Ltd, P.O.Box 679, Rishon Lezion, 75106, Israel

ABSTRACT

Bootstrap technique offers the most powerful and efficient implementation tools for the modern multivariate analysis, test of hypotheses, knowledge acquisition, experimental inference. This paper proposes algorithmic solutions of complex data based problems: selection of distributions, regression models, discrimination analyses, imputations for missing data, diagnostics, predictions, etc., without the need for using unrealistic, unverifiable assumptions. Advanced bootstrap methods allow a full extraction of knowledge from available data in relation to the considered task, automatically build the required knowledge base from the data base and receive responses of interest, i.e. create an automatic generator of inferences. The obtained results define the design of new family of knowledge based intelligent systems.

1. INTRODUCTION

The bootstrap paradigm develops effective automatic inference techniques for complex actual data structures. At the same time the bootstrap algorithms are comparatively simple, flexible, logical, straightforward and to the point, close to natural human thinking.

Historically, this data-based simulation resampling technology was initiated by one of the authors Peshes (1970, 1972). By using resampling procedures (multiple reproduction of the original sample in samples of the same size with identical statistical properties - now called bootstrap samples) the following problems were solved:

- defining required confidence intervals for any characteristic of interest;
- selecting adequate distribution law for observed sampled data, the key for serious errors elimination in the use of statistical models.

The software support for the new method was represented by the packagc known by the name PARUS.

In the late 70s Efron (1979) introduced the second generation of this paradigm and the new term "Bootstrap". He showed that the bootstrap is a special generalized methodology for an applied statistical problem solution. The bootstrap technique was suggested for sampling bias correction, higher order accuracy for any parameter and confidence interval estimation.

Currently it is established that the bootstrap technology defines the rules controlling the actual data sampling process and allows:

- complete information (knowledge) extraction from the observed data in required direction and construction of appropriate inferences, using self resources;
- powerful tool introduction for hypothesis testing and selection of a model most conforming to the observed data.

This concept of entity selection is close to the Fisher's principle of maximum likelihood, which is the keystone of mathematical statistics. The practical applications of this classical pattern are based on the usually unknown hypothetical family distribution of the observed data. Moreover, the maximum likelihood estimations do not allow for bias correction arising at the time of sampled data processing, i.e. they are unable to extract all of the information from the source data. The bootstrap technique guarantees overcoming of these difficulties.

Inference engines for all problems such as estimations, constructions of confidence and prediction regions, hypotheses testing, dependencies finding, multiple imputations, discriminations, classifications, diagnostics, predictions, etc. represent a series of completely automated procedures, which are performed on repeated simulation sets of bootstrap samples.

Thus, instead of the prior hypothetical models (distributions, formulas, equations, heuristics, etc.), the data based bootstrap resampling methods are able to consider the original raw data such as pattern of investigation situation, and extract all necessary information for the given problem.

Further development of these concepts allows to determine the general architechture, performance specification, algorithm description, and user interface design of the universal automatic inferential generator for a new family of self-training systems extracting knowledge from available data which is called Novel Expert System Intelligent Marvel (NESIM). NESIM operates as a rule with a specialzed data base for a specific application and automatically builds a knowledge base according to the considered problem. Thus, NESIM is able to make valid inferences by itself using source data, i.e. by its own bootstrap. This essentially distinguishes NESIM from traditional expert Knowledge Based Systems (KBS). The success of such KBS is to a large extent depends on the one hand, on the abilities of highly qualified experts to express their knowledge and, on the other hand, by the ability of knowledge engineers to create adequate rules and knowledge base from contacts with these experts.

The main goal of this paper is to develop major principles, standard procedures and methods of processing data at hand (similar to the industrial flow of raw material processing) which automatically produces optimal inferences for any complicated testing situations.

The possibilities of self - training (automatic learning) and self - development of NESIM are provided via automatic modification of inference engines and input of new source data. The development of inference mechanisms is realized by means of regular testing of various models and corresponding bootstrap procedures for known situations determined by the available data. This way of self - checking allows to draw conclusions about the best method of providing maximum result precision or minimum risk of any error inferences.

The important advantages of bootstrap technique and particularly its ability to solve numerous previously insolvable problems open new horizons in the modern information technology.

2. BASIC CONCEPTS AND TECHNIQUES OF BOOTSTRAP

The utilization of parametric statistics methods provides prior information about the type of family distributions of considered data. However, the problem of defininig the suitable family distributions from available data, which would allow to avoid serious errors in statistical inferences, has no solution. On the other hand, finding solutions by non-parametric methods independent of distribution causes a portion of information contained in the observed data to be losed. Even though these methods are applicable under more generalized conditions, they do not allow to extract all required information from the available data and their power is lesser than in the parametric approach. Bootstrap paradigm proposes specialized computer-oriented resampling data processing technology to alleviate this problem. Two fundamental theorems serve as the mathematical basis for this technique: uniform convergence of empirical distribution to the theoretical one and artificial reproduction of random varialbles for the given distribution. These theoremes are true without strong restrictions.

The main idea underlying bootstrap is as follows. The data at hand is treated as if it represents the entire population. The available data sample is recreated a multiple number of times in order to obtain a large number of new bootstrap samples similar to the original sample and of the same size. Statistics of interest are defined for each resampled sample. This ensemble of obtained estimates characterizes variability of the corresponding solution for the original sample and allows to evaluate its quality (accuracy), compare differing solutions and make the most plausible inference.

The mathematical essence of the bootstrap paradigm is that an additional randomization axiom is introduced into the data processing algorithms which allows to realize an effective control directly during the solution process and automatically forms corresponding inference engines. The original data is used as a pool from which, with the aid of bootstrap resampling procedures and corresponding plug-in estimates, it is possible to extract all contained information in required direction, i.e. bootstrap technique is capable to consider the original raw data as a pattern of the situation under consideration. Thus, as mentioned in Bluvband and Peshes (1993), bootstrap technique allows to eliminate the principle shortcomings of the nonparametric approach. On the other hand, bootstrap as a powerful tool for testing and selection of adequate models, expands the possibilities of the parametric approach.

Let there be a random observed sample $X=(X_1,X_2,...,X_n)$ of size n drawn from some population F. The simplest reproduction procedure is defined by the random selection of n elements from the original sample X with replacement. The weight (probability) of each variable x_i equals n^{-1} ; $i=1,2,...,n$. Evidently, this method can be used for any arbitrary sample space, whose elements can be numerical or quality varialbes, vectors, images, maps, etc. The observed sample X defines discrete empirical distribution of these units.

If the observed data points X represent numerical values with some continuous variable, then to obtain bootstrap samples, the smoothing procedures can be used. This kind of bootstrap is called smooth. The simple smoothing procedure represents segmented-linear approach by using uniform distribution for jumps of observed empirical law. The standard smoothing procedures apply the kernel density methods of estimation, according to Rozenblat (1956), Parzen

(1962), Cacoullos (1966), Yepanchikov (1969). The smoothing empirical distributions for the source sample X enable to obtain bootstrap samples containing a new real number in the gaps between the variable values of the original sample.

In the process of specific problem solving arises a need to receive bootstrap samples with required properties. For example:

- an equal number of units of the original raw data must be contained in the general population of bootstrap samples (balanced bootstrap);
- among the bootstrap samples there should be no sample coinciding with the original (unrepeated bootstrap);
- the bootstrap procedure can be applied to the specifically selected subset of the original sample (partial bootstrap);
- bootstrap samples are formed according to some generating mechanizm;

A general framework of bootstrap technology for problem solving is described bellow:

- sequential generation of a large number B (10^2-10^7) of bootstrap samples;
- calculation of characteristics of interest for each of the above (e.g., order statistics, moments, model parameters, etc.);
- direct transformation of simulated set results corresponding to the formulated problem and creation of inference engines, checking and control mechanizms;
- obtaining a solution of the considered problem, representing the answer most conforming to the real data.

On one hand, the observed statistical material in the bootstrap serves as the basis for space formation of possible solutions, and on the other hand, defines criteria for decision making.

3. BOOTSTRAP BASED SOLUTION FOR SOME IMPORTANT PROBLEMS

3.1 Selection of Most Adequate Distribution Function

Let there be an X observed sample of limited random real variables and some set of various families of theoretical distributions $F_1(X)$, $F_2(X)$,...,$F_k(X)$. The task is to choose from the above the definition best corresponding to the observed data. The solution of this problem, solved as a rule based on goodness of fit tests, is incorrect. The proposed solution is based on known property of distributions that if two distributions have close values for 4-6 first moments (semiinvariants) then they approximate each other.

The solution of the considered problem is implemented by the following selection algorithm. B bootstrap samples are created for empirical distribution of observed data X. Empirical moments of order i=1,2,...,m are found for each of the B bootstrap samples and for the source sample. Moments $v_{ie}^{(j)}$ of order i are arranged in the variational set $v_{(ie)}^{(1)} \leq v_{(ie)}^{(2)} \leq . . . \leq v_{(ie)}^{(B)}$; i=1,2,...,m. Parameters are estimated and m corresponding moments v_{iT} are calculated for each theoretic distribution $F_r(X)$; r=1,2,...,k by the source sample. A series of nested pivotal confidence intervals of smallest length are constructed around each empirical moment v_{ie} found from the original sample by decreasing the confidence levels $\gamma_1 > \gamma_2 > ...$.

Denote γ100% interval for the i-th moment $[v_{(ie)}^{(j_{low})}, v_{(ie)}^{(j_{up})}]_{\gamma}$ where j_{up}-j_{low}=ent[γB] for {γB-ent[γB]}>0.5, otherwise j_{up}-j_{low}=ent[γB]-1; where ent[·] means integer part of [·]. If one of the moments $v_{iT} < v_{(ie)}^{(1)}$ or $v_{iT} > v_{(ie)}^{(B)}$; i=1,2,...,m of some theoretical distribution then this distribution is not cosidered any longer. The lower and upper limits of specified intervals define an m-dimensional hyperrectangle. Evidently, with decrease of γ level, the new hyperrectangle is nested into the hyperrectangle for greater level, i.e.

$[v_{(ie)}^{(j_{low})}, v_{(ie)}^{(j_{up})}]_{\gamma_2} \subset [v_{(ie)}^{(j_{low})}, v_{(ie)}^{(j_{up})}]_{\gamma_1}$ for $\gamma_2 < \gamma_1$ for all i=1,2,...,m.

The most appropriate theoretical distribution is selected from the considered family based on criteria falling into the smallest hyperrectangle. The considered method does not require assumptions of symmetry for moment distribution the way it used to be at the first application of bootstrap technique for the solution of this task based on central confidence intervals. However, supposition of independent distribution of moments of different order is considered in both solutions.

The algorithm for the required solution which allows to relax the independence restriction is proposed bellow. Ennumerate all empirical moment vectors of bootstrap samples from 1-B. As in above case, the variational sets are formed for each empirical moment. Each member of this set has a vector number to which it belongs. If one of the moments $v_{iT} < v_{(ie)}^{(1)}$ or $v_{iT} > v_{(ie)}^{(B)}$; i=1,2,...,m of some theoretical distribution then this distribution is not cosidered any longer. The minimal confidence regions where all specified moments will be included are constructed for theoretical moments of the rest of the distributions. These regions are defined by confidence intervals for each moment $[\nu_{(ie)}^{(j_{low})}, \nu_{(ie)}^{(j_{up})}]$; i=1,2,...,m where the lower limit is defined as the closest value in i-th variational set which does not exceed the value $\min_{s}[\nu_{(ie)}, \nu_{iT}(s)]$, *i.e.* $\nu_{(ie)}^{(j_{low})} \leq \min_{s}[\nu_{(ie)}, \nu_{iT}(s)]$ and the upper limit as the closest value $\nu_{(ie)}^{(j_{up})} \geq \max_{s}[\nu_{(ie)}, \nu_{iT}(s)]$ where s corresponds to distribution $F_s(X)$. Evidently, level $\gamma_i^{(1)}$ of confidence interval for i-th empirical moment is estimated by the value $[(j_{up}-j_{low}+1)B^{-1}]_i$. For each constructed confidence interval there is a corresponding subset of numbers of empirical moment bootstrap vectors. The quantity $r^{(1)}$ of numbers which form the intersection $S^{(1)}$ of considered subsets allow to estimate a common confidence level $\gamma_{com}^{(1)}=r^{(1)}B^{-1}$ of all constructed confidence intervals.

If $\gamma_{com}^{(1)}=\gamma_1^{(1)}\gamma_2^{(1)}\ldots\gamma_m^{(1)}$, then the specified moments are independent and the appropriate distribution can be selected by the method described above. If $\gamma_{com}^{(1)}<\gamma_1^{(1)}\gamma_2^{(1)}\ldots\gamma_m^{(1)}$, then the considered distributions are not adequate for actual data approximation. Otherwise, distribution selection is as follows. First, theoretical distributions which are worse than the others are removed. These distributions include ones in which at least one of the component vector moments defining some upper or lower confidence limit contains a vector number which does not belong in the intersection $S^{(1)}$. As a result of such exclusion the length of some confidence intervals and their levels will become smaller. The remaining vectors define new confidence levels $\gamma_1^{(2)}, \gamma_2^{(2)}, \ldots, \gamma_m^{(2)}$ and common level $\gamma_{com}^{(2)}$. Thereafter, the distribution where the removal of the theoretical moment vector results in the smallest value $\gamma_1^{(3)}\gamma_2^{(3)}\ldots\gamma_m^{(3)} \geq \gamma_{com}^{(3)}$ is excluded from consideration. This process is repeated until there is left only one most appropriate distribution. The described method determines a general principle for selection of distribution whose theoretical moments fit closest into the confidence intervals for corresponding empirical moments, i.e. most conforming to observed data.

These methods are easy to expand for finding adequate multivariate distributions for observed sample of random vectors. This bootstrap technique allows also to construct confidence intervals for found distributions parameters.

The importance of solving the problem considered above is that it usually constitutes a part of a greater parametric decision making process.

3.2 Regression Problems

These problems are closely related to the previously described cases because they represent tasks of joint distribution of investigated variables. In a traditional sense these problems have two aspects:

- selection of an approriate dependence of some variables (responses) from the others - regressors (regression analysis);
- investigation of mutual association between several variables (correlational and exploratory analysis, reduction of dimensionality, analysis of variance).

The regression curve represents the conditional expectation of response y given x $E(Y|X)=\Psi(X)$ where y and x, generally speaking, are multivariate vectors, $\Psi(X)$ some functional of x. If distribution of y for each x is known, then it is possible to find a mean y given x, i.e. it is possible to find the required regression y given x. In most real-world situations these distributions are known, so the exact regression can be found. In such a case the regression curve must be estimated. In classical regression analysis this problem is solved in the linear model class: $E(y|x)=\sum_{j=1}^{l}\beta_j g_j(x)=G\beta$ where β is the vector of unknown parameters β and G is the vector of nonrandom functions $g_j(X)$, called exploratory variables. The probability structure of this model for observed sample of n vectors is $Y_i=G_i\beta+e_i$ where Y_i is the i-th response, $G_i=G(X_i)$, e_i is an error term, i=1,2,...,n. The task is to find the unknown parameters, their confidence limits and building prediction intervals for responses when new observation exploratory variables exist. In regression theory these problems are solved only in the cases when the random errors e_i are uncorrelated and their distribution has a constant variance. Moreover, it is usually presumed that these errors are normally distributed which leads to the independence of these random values. Bootstrap allows to surmount this limitation. Efron and Tibshirani (1993) give several variations of such solutions.

This task is related to finding regression curve for the given type of linear model G. A more general problem is the selection of a model from some set of dependencies most conforming to actual data. The bootstrap technique enables solutions for those problems. Observed data define empirical densities $f_e(X,Y)$, $f_e(X)$ and correpondingly the conditional empiric density $f_e(y|x)=f_e(x,y)\,f_e^{-1}(x)$. The regressional curve $E(y|x)=\int_{-\infty}^{\infty} yf(y|x)\,dy$ and m-1 moments $E(y^j|x)=\int_{-\infty}^{\infty} y_j f(y|x)\,dy$, j=2,3,...,m are defined as empirical statistical approximation for observed data. The B bootstrap samples are generated from the original data. The specified characteristics are found for each bootstrap sample. The unknown parameters of each considered model G are estimated from the source data and corresponding moments $E(y^j|G)$ are defined. The selection of the regression model depends on the condition of all of its moments belonging in the narrowest bootstrap pivotal confidence regions. Moreover, bootstrap allows to obtain a solution for more general regression models that have no mathematical solution, i.e. when the regression curve is non-linear in the parameters.

3.3 Dependence Problems and Statistical Inferences

These problems are considered most difficult. Solutions are obtained only for some simple data structures or for restrictions of non-adequate real situations. In classical statistical analysis the association between the variables is estimated by correlation coefficient ρ and the correlational ratio η which satisfies the inequality $0\le\rho^2\le\eta^2\le1$. The correlation coefficient ρ characterizes the closeness of linear statistical dependence and does not reflect other forms of mutual association between variables. The condition that some of the random variables are non-correlated $\rho=0$ is a necessary condition of their independence but not sufficient. The correlational ratio η is not connected with the assumption of linearity and evaluates the fact of presence or absence of functional dependency. This coefficient does not reflect the type of this dependency except for statistical or strong linear connection. The difference $\eta^2-\rho^2$ characterizes the non-linearity measure. Therefore, the comparison of values ρ and η allows to estimate only some of the simplest properties of mutual association of considered varialbles.

The effective abilty of bootstrap technique to build the required confidence intervals and to investigate any properties of real empirical distributions allows to utilize the necessary and sufficient performance of mutual association of variables between themselves or dependence of some variables from the rest.

Let there be a matrix of observed data $X=(x_j^{(i)})$ containing n p-dimensional vectors, j=1,2,...,n; i=1,2,...,p. The matrix X defines empirical density of p-variables $f_e(X^{(1)}, X^{(2)},...,X^{(p)})$ which serve as an approximation for their real distribution $f(X^{(1)}, X^{(2)},...,X^{(p)})$. If the varables are independent in aggregate then the necessary and sufficient condition for their function distribution is the equality $f(X^{(1)}, X^{(2)},...,X^{(p)})=f(X^{(1)})f(X^{(2)})...f(X^{(p)})$, i.e. it represents the product of marginal distributions of each variable. The variables $x^{(i_1)},\ldots,x^{(i_k)}$ are independent of the variables $x^{(i_{k+1})},\ldots,x^{(i_p)}$ if $f(x^{(i_1)},\ldots,x^{(i_k)}\,|\,x^{(i_{k+1})},\ldots,x^{(i_p)})=f(x^{(i_1)},\ldots,x^{(i_k)})$.

This criteria allows to establish the connection measure between different groups of variables and to investigate the type of dependence between them. Let the confidence intervals $[a_{low}^{(i)},a_{up}^{(i)}]$ be constructed for any arbitrary group or all variables of observed vectors independently for i-th variable. Each of these intervals represents a segment of the variational set for the corresponding variable with confidence level $\gamma_i<1$ which are estimated by the ratio of the quantity of vectors belonging in this interval, including the boundaries, to n. For the selected group of m variables $i_1,\ldots,i_m$, the indicated intervals form some confidence m-dimensional hyperrectangle. The probability of being included in this region, i.e. the confidence level γ_{com}, is estimated by the quantity of vectors divided by n. The quantity of vectors is determined by the number of vectors falling into all of the considered intervals. If the considered variables are independent, then the common confidence level $\gamma_{com}=\gamma_{i_1}\gamma_{i_2}\cdots\gamma_{i_k}=\gamma_{ind}$ for any population of indicated intervals. If the considered case contains dependent variables, then $\gamma_{com}>\gamma_{ind}$. The possible situation when $\gamma_{com}<\gamma_{ind}$ is not of interest because it indicates that dependence of variables exists in another region of values of those variables. It is easy to see that the maximal value of γ_{com} does not exceed $\min\{\gamma_{i_1},\gamma_{i_2},\ldots,\gamma_{i_m}\}=\gamma_{i_0}=\gamma_{dep}$. Therefore, for some constructed system of confidence intervals, the common confidence level satisfies the inequality $0<\gamma_{ind}\le\gamma_{com}\le\gamma_{dep}<1$. The difference $\gamma_{com}-\gamma_{ind}$ characterizes a measure of association or deviation from independence of population of investigated variables in the corresponding area of their change. In the case when the partial characteristic of association of components $i_1,i_2,\ldots,i_k$ relative to $i_{k+1},i_{k+2},\ldots,i_p$ is estimated, the value $\gamma_{dep}^{(par)}=\min\{\gamma_{i_1},\gamma_{i_2},\ldots,\gamma_{i_k}\}\ge\gamma_{dep}$ is used.

The first time a similar measure of deviation from the variable independence was introduced by Pearson (1904) for two variables of specified frequency contingency table. The proposed characteristic normalized by association coefficients $C_{com}=(\gamma_{com}-\gamma_{ind})(\gamma_{dep}-\gamma_{ind})^{-1}$ for variable population and $C_{com}^{(par)}=(\gamma_{com}-\gamma_{ind})(\gamma_{dep}^{(par)}-\gamma_{ind})^{-1}$ for a particular component group allows to estimate the association of different variables defined by arbitrary distributions of any dimension in the specified area of their change. The indicated coefficients are changing from 0 for independent variables to 1 in the case when at least one variable is completely defined by the others.

Investigation of the intersection of sets of vector numbers falling into different constructed confidence regions for the source data, and similarly for the bootstrap samples, allow to define the presence of any type of dependence and to solve the problem of dimension reduction. This direction considerably expands the possibilities for parametric methods.

Non-parametric solutions. The main advantages of bootstrap technology are determined by the ability to receive non-parameter solutions of most important problems of diagnostics, predictions, finding responses in regressions, imputations for missing data, etc. maximally conforming to the actual data.

The natural approach to problem solving consists of the following. For some tested vector, containing some set of known attributes and one or more unknown components, it is required to find the nearest neighbours from the observed vector sample.. Each observed vector is split into two sub-vectors. One contains the components similar to the ones contained in the tested vector. The other sub-vector contains responses.. The task is to find a subset of observed vectors most similar to the tested vector. The unknown characteristics of interest, i.e. the absent components of the tested vector, are estimated by responses of the indicated subset of vectors.

Let the observed data contain n p-dimensional vectors X_j, i.e. $X_j=\{x_j^{(1)}, x_j^{(2)}, \ldots, x_j^{(p)}\}$; j=1,2,...,n and the tested vector is $Y=\{y^{(i_1)}, y^{(i_2)}, \ldots, y^{(i_k)}\}$; $k<p$, where $i_1, i_2, \ldots, i_k$ defines common components of the observed vectors. If the tested vector contains at least one unit which is bigger or smaller then the corresponding components of observed subvectors or the quality component absent from the actual data, then the problem is not solved. In this case it is said that the given vector does not belong to observation space. This condition can be sometimes corrected, e.g. if the characteristic .of the trend of the compared variables is known.

The algorithm of finding the nearest neighbours follows. Each k-dimensional observed subvector in relation to the tested, forms some confidence region consisting of defined pivotal confidence intervals for each of the k-components. As before, the variational sets are formed for each of the components of the observed subvectors and simultaneously each value of the set is marked by the number of the vector it belongs to. When the qualitative and categorized variables are ordered, their identical elements follow one after the other.

The construction of pivotal confidence intervals for j-th subvector by the i_l component; l=1,2,...,k, is defined by the following rules. If $x_j^{(i_l)}$ is contained in the variational set to the left of the value $y^{(i_l)}$, then the left confidence boundary is defined by $x_j^{(i_l)}$; the right boundary - by the closest to the right or coinsiding with $y^{(i_l)}$ value of this variational set. If $x_j^{(i_l)}$ is to the right of the $y^{(i_l)}$ then it defines the right boundary; the left boundary corresponds to the closest on the left $y^{(i_l)}$ value of the considered set. And finally, if $x_j^{(i_l)}$ coinsides with $y^{(i_l)}$ then the deteriorated pointed confidence interval defined by one value $x_j^{(i_l)}=y^{(i_l)}$ is considered. The vector $y^{(i_l)}$ in which all components i_l coinside with the corresonding components of $x_j^{(i_l)}$ is its twin. The level γ_{i_l} of the constructed confidence interval is estimated by the ratio of quantity of subvectors falling in to it, including the boundaries, divided by n. The common level $\gamma_{com}(j)$ of this region for the j-th object is defined by the ratio of the quantity of subvectors falling into all confidence intervals for each component, to n.

The first nearest neighbour for the tested vector is defined by the number j_1 for which the minimum $\gamma_{ind}(j_1)=\gamma_{i_1}(j_1)\gamma_{i_2}(j_1)\ldots\gamma_{i_k}(j_1)\leq\gamma_{i_1}(j\neq j_1)\gamma_{i2}(j\neq j_1)\ldots\gamma_{i_k}(j\neq j_1)$ and $\gamma_{ind}(j_1)\leq\gamma_{com}(j_1)$. In case when not one of the observed subvectors satisfies the indicated condition, the problem is not solved because the tested vector Y does not correspond to the observed data.

After the first nearest neighbour is found (vector X_{j_1}), all vectors are defined with numbers $j\neq j_1$, in the confidence interval of which falls the vector X_{j_1}. All such vectors are excluded from the number of possible nearest neighbours of vector Y. The significance of this procedure (Pareto-optimal principle, Pareto (1909)) is that all of them have larger deviations in the same direction from the tested vector Y in comparison with the vector X_{j_1}. The second nearest neighbour X_{j_2} is found similarly and later the indicated Pareto procedure is performed for it. Evidently, the level $\gamma_{ind}(j_2) \geq \gamma_{ind}(j_1)$. This process is performed until all of the nearest neighbours are found, i.e. subvectors with numbers $j_1, j_2, ..., j_r$; $r \leq n$.

The (p-k)-dimensional subvectors of responses corresponding to each of the found nearest neighbours are utilized for the problem solving. The value of the levels $\gamma_{ind}(j_v)$; $v=1,2,...,r$ can be considered as a relative estimate of the distances of the subvector j_v from the tested vector Y. Then the weight of each neighbour is estimated by the value $w_v = \left\{ \gamma_{ind}(j_v) \sum_{v=1}^{r} [\gamma_{ind}(j_v)]^{-1} \right\}^{-1}$; $\sum_{v=1}^{r} w_v = 1$. The required empirical distribution of responses is defined by the resulting r subvectors, the weight of each is W_v,$v=1,2,...,r$. The weighted bootstrap of this distribution allows to estimate all characteristics of interest and their confidence regions. It is easy to see that the finding of bootstrap sample responses can be organized according to the algorithm described above which is performed on the source samples of k-dimensional subvectors. The required solution for some simple problems for sufficiently representative source sample may be obtained without bootstrap.

The proposed approach allows to obtain nonparametric solution for the set of multivariate statistical analysis problems indicated above.

4. GENERAL FRAMEWORK OF NEW AUTOMATICALLY SELF-TRAINING INFERENCE MACHINE

The ability of bootstrap technology to extract the necessary knowledge and to obtain the required solutions from adequate source data allows to design a new family of intelligent systems NESIM. Project NESIM allows to realize the idea about the ideal inference machine of the future described by Efron and Tibshirani (1993):
"The theory of the bootstrap is "pre-loaded" into an algorithm and carried out entirely by the computer for any particular application. This doesn't free the statistician from thinking, of course, but it does allow the thinking to concern inferential questions of direct interest to the scientist, rather than a host of small mathematical difficulties.
One can describe the ideal computer-based statistical inference machine of the future. The statistician enters the data, the questions of interest, and the class of allowable probability models. Without further intervention, the machine answers the questions, in a way that is optimal according to statistical theory." The NESIM is able to improve "this ideal" by eliminating the necessity for the entry of the allowable models into the designed inference machine. The characteristic of computer-based products of NESIM family is defined by the aggregate of included subsystems, algorithm of their implementation and cooperation providing industrial technological approach to data processing by bootstrap methods. The main NESIM subsystems include:

- vector data base in the heterogenious attribute space open for additions;
- building block for initial knowledge base providing friendly user interface of problem input and classification of vectors data base in accordance with the type of the given problem;
- building block for working knowledge base realizing automatic control of non-contradiction in separated classes used as training samples;
- problem solving block calculating responses (prediction, diagnosis, estimation, etc.) for a specific vector task;
- self-training block in which the classification analysis is performed by the vectors of the existing data base, the results of the obtained solutions, with expansion of the data base, automatic selection and modification of inference engines, different model testing;
- service block in which the analysis of informational importance of components of vectors of working knowledge base, determination of their dependence, investigation of ability of dimension reduction, graphical imaging, etc. is performed.

The main advantages of the proposed products of NESIM family are as follows:

- industrial technological approach to data processing;
- automatic creation of required knowledge bases from the given data bases;
- automatic generation and modification of inference engines;
- ability of self-analysis, self-training and self-development during the system usage;
- natural empiric way of data presentation;
- possibility of testing various hypotheses, models, concepts and inclusion into the system.

REFERENCES

Bluvband, Z. and Peshes, L. (1993). Bootstrap Technology for RAM Analysis. Procedings of the Conference on New Direction in Military RAM, organized by the American Defense Preparadeness Association. October 19-20, Aberdeen

Proving Ground, MD, 145-155.

Cacoullos, T. (1966). Estimation of Multivariate Density. Ann. Inst. Stat. Math., 18, 179-189.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. Ann. Stat., 7, 1-26.

Efron, B. and Tibshirani, R.J. (1993). An Introduction to the Bootstrap. Chapman & Hall, London.

Pareto, V. (1909). Manuel d'Economie Politique. Gurd, Paris.

Parzen, E. (1962). On Estimation of Probability Density Function and Mode. Ann. Math. Stat., 33, 1065-1076.

Pearson, K. (1904). On the Theory of Contingency and its Relation to Association and Normal Correlation. Drap Co. Memoirs. Biometric Series, 1, London.

Peshes, L. Ya. and Michlukova, L.A. (1970). Confidence Intervals for Parameters of Arbitrary Distribution by Me of Statistical Simulation. Proceedings of the Conference on Reliability Problems of Electro- and Radio-Techni Products. Volume 3, Tbilisi, 73-75.

Peshes, L.Ya. and Stepanova, M.D. (1972). Fundamentals of Accelerated Reliability Testing. Science & Techn Minsk.

Rozenblatt, M. (1956). Remarks on Some Nonparametric Estimates of Density Function. Ann. Math. Stat., 27, 832-8

Yepanchikov, V.A. (1969). Nonparametric Estimation of Multivariate Density. Probability Theory and Its Applicatio Vol. 14, No. 1, 156-160.