UDC 004.272.26: 004.93

Oliinyk A.[1], Subbotin S.[2], Lovkin V.[3], Ilyashenko M.[4], Blagodariov O.[5]

[1]PhD., Associate Professor, Associate Professor of Department of Software Tools, Zaporizhzhia National Technical University, Zaporizhzhia, Ukraine
[2]Dr.Sc, Professor, Head of Department of Software Tools, Zaporizhzhia National Technical University, Zaporizhzhia, Ukraine
[3]PhD, Associate Professor, Associate Professor of Department of Software Tools, Zaporizhzhia National Technical University, Zaporizhzhia, Ukraine
[4]PhD., Associate Professor, Associate Professor of Computer Systems and Networks Department,
[5]Postgraduate student of Department of Software Tools, Zaporizhzhia National Technical University, Zaporizhzhia, Ukraine

PARALLEL METHOD OF BIG DATA REDUCTION BASED ON STOCHASTIC PROGRAMMING APPROACH

Context. The task of automation of big data reduction in diagnostics and pattern recognition problems is solved. The object of the research is the process of big data reduction. The subject of the research are the methods of big data reduction.

Objective. The research objective is to develop parallel method of big data reduction based on stochastic calculations.

Method. The parallel method of big data reduction is proposed. This method is based on the proposed criteria system, which allows to estimate concentration of control points around local extrema. Calculation of solution concentration estimates in the developed criteria system is based on the spatial location of control points in the current solution set. The proposed criteria system can be used in stochastic search methods to monitor situations of excessive solution concentration in the areas of local optima and, as a consequence, to increase the diversity of the solution set in the current population and to cover the search space by control points in a more uniform way during optimization process.

Results. The software which implements the proposed parallel method of big data reduction and allows to select informative features and to reduce the big data for synthesis of recognition models based on the given data samples has been developed.

Conclusions. The conducted experiments have confirmed operability of the proposed parallel method of big data reduction and allow to recommend it for processing of data sets for pattern recognition in practice. The prospects for further researches may include the modification of the known feature selection methods and the development of new ones based on the proposed system of criteria for control points concentration estimation.

Keywords: data sample, pattern recognition, feature selection, parallel computing, informativeness criterion, stochastic programming approach.

ABBREVIATIONS

CMES is a Canonical Method of Evolutionary Search;

GMDH is a Group Method of Data Handling;

MARF is a Method of alternately Adding and Removing of Features;

MMDCA is a Multiagent Method with Direct Connection between Agents;

MMICA is a Multiagent Method with Indirect Connection between Agents;

PCA is a Principal Component Analysis;

PMBDR is a Parallel Method of Big Data Reduction.

NOMENCLATURE

$d(\chi_k, \chi_u)$ is a distance between points χ_k and χ_u of the search space XS;

$d(iter)$ is an average distance between all solutions on the current iteration;

g_{mk} is a m-th coordinate of the k-th solution;

$\overline{g_{mClc}}$ is the m-th coordinate of the c-th cluster center;

$Inform_k$ is an information about the k-th solution;

$InformLI(\chi_k)$ is a flag, which represents presence of solution χ_k in the solution set $R(iter) = \{\chi_1, \chi_2, ..., \chi_{N\chi}\}$ on the last search iteration $iter$;

$InformM(\chi_k)$ is a list of methods, which were used for estimation of solution χ_k;

M is a number of features in the sample of observations S;

$N_{\chi j}$ is a number of control points, which were investigated at the j-th process;

$N(iter)$ is a number of unique sampling points $Xe \in XS$, estimated in the process of feature selection till the current iteration $iter$ inclusively;

$N(XS)$ is a number of discrete space points XS;

P is a set of features (attributes) of observations in the given sample;

p_{qm} is a value of the m-th feature (attribute) of the q-th observation;

Q is a number of observations in the given sample of observations S;

Q_{ic} is a number of incorrectly recognized observations;

$rand[0;1]$ is a randomly generated number from the interval $[0;1]$;

S is a sample of observations (training sample);

$V(p_m)$ is an informativeness of a feature p_m;

$V(\chi_k)$ is a value of objective function of the k-th solution;

t_q is a value of output parameter of the q-th observation;

T is a set of output parameter values;

χ_k is the k-th solution, which corresponds to the k-th investigated control point Xe_k in the search space: $\chi_k \to Xe_k$.

DOI 10.15588/1607-3274-2018-2-7

INTRODUCTION

The investigation of complex technical objects and processes is connected with the necessity of big data processing, particularly with the search of feature set which describes investigated objects and processes in the best way [1–6]. The elimination of non-informative or insignificant features for diagnostic and recognition model synthesis process will allow to reduce model synthesis time, amount of processed data and complexity of the model which was built, but also to improve approximation and generalization abilities of the model [7–14].

As is well known [15–18], feature selection process is a highly iterative and resource-demanding procedure, which makes difficult to execute it in practice for solving of the tasks, where data processing should be performed without significant time delays (in on-line mode). Therefore the development of highly productive data reduction methods based on parallel computing is an actual task.

The object of the research is the process of big data reduction. The subject of the research are the methods of big data reduction. The research objective is to develop PMBDR based on stochastic calculations.

1 PROBLEM STATEMENT

Suppose we have data sample $S=<P,T>$, which consists of Q observations. Every observation is characterized by values of input attributes p_{q1}, p_{q2}, …, p_{qM} and output parameter t_q, where p_{qm} is a value of the m-th input feature of the q-th observation ($q=1, 2, ..., Q$, $m=1, 2, ..., M$); M is a total number of input features in the sample of observations S. Then the problem of informative feature selection can be ideally [1, 7, 19–21] stated as searching for the feature combination P^* from the initial data sample $S=<P,T>$ with minimum value of the given criterion of feature set quality estimation: $V(P^*)=\min\limits_{Xe\in XS} V(Xe)$, where Xe is a member of the set XS; $V(Xe)$ is a criterion of estimation of significance of feature set Xe; XS is a set of all possible feature combinations, which are obtained from the initial feature set P.

2 REVIEW OF THE LITERATURE

At present different methods are used for data reduction by means of informative feature selection. The most frequently used methods are the following ones.

Method of complete enumeration [1, 2, 7] estimates each control point Xe from all possible (2^M-1) control points in the search space XS. Because of complete enumeration of all possible solutions $Xe\in XS$, this method allows to find solution P^*, which has optimal value of objective function $V(P^*)=\min\limits_{Xe\in XS} V(Xe)$. As computing complexity of this method $O\left(2^M\right)$ significantly depends on input feature number M of training sample $S=<P,T>$, this method can be used for selection of features from small data samples. It substantially troubles and makes it impossible to apply this method for possessing of big data samples.

Heuristic methods [2, 7] (method of sequentially feature adding, method of sequentially feature removing) use greedy search strategy, which sequentially add (remove) features to the current feature set. Such approach is more simple in comparison with complete enumeration and demands less computing and time costs. But combinations of features P^*, selected by such methods, are generally characterized by unacceptable values of optimality criterion $V(P^*)$, because heuristic methods investigate very limited areas of search space. As a result feature combinations have optimal (or acceptable) value of objective criterion $V(P^*)$. Computing complexity of such methods is proportional to square of feature number M of input sample $S=<P,T>$: $O\left(M^2\right)$. Therefore application of such approach, when features are selected from big data samples, is also difficult.

Methods of stochastic search are based on application of probabilistic procedures for processing of control points $Xe\in XS$ and generally work with some solution set $R(iter)=\{\chi_1,\chi_2,...,\chi_{N\chi}\}$ on every iteration. Every k-th solution $\chi_k\in R(iter)$ corresponds to the k-th control point Xe_k in the search space XS on the investigated iteration $iter$: $\chi_k\to Xe_k$. Such methods can use evolutionary, multiagent or other approaches of computational intelligence as mathematical basis. Methods of stochastic search during the given number of iterations $Iter$ process $Iter\cdot N_\chi$ control points (where N_χ is a number of solutions, which are processed on every iteration of stochastic search). Therefore computing complexity $O(Iter\cdot N_\chi)$ of this approach does not depend directly on number of features M in the input sample. It allows to apply it for big data reduction. But such methods are given to recirculation in the areas of local optima (during search process some set of control points χ_k is concentrated around local extrema areas). It reduces its application efficiency and raises search time. Therefore expansion of the investigated areas of search space XS is based on usage of big number N_χ on control points χ_k, which are investigated on every iteration. This approach is not effective too because of the low diversity of solutions in the set $R(iter)$. Besides usage of big number of control points N_χ on every iteration increases search time.

Approach, which performs ranking of features p_m according to the values of individual significance $V(p_m)$ regarding output parameter T, can be used for feature selection also. Such approach is computationally simple (its computing complexity is $O(M)$), but it does not take

into account interdependence of features. Therefore in practice when features are interdependent, this approach doesn't allow to select feature sets, which have optimal or acceptable values of criterion of estimation of group informativeness $V(P^*)$.

Thus shortcomings of the existing feature selection methods cause necessity of the development of new method, which is based on stochastic approach and highly productive computings and is free from the described shortcomings.

3 MATERIALS AND METHODS

As was mentioned, application of the known methods of feature selection in practice for big data processing is difficult due to high iterativity and big amount of computings [1–12]. Besides it search strategies, used for feature selection, are also not enough effective for investigation of different areas of search space. Thus greedy strategy, which is used in heuristic methods of feature selection [1, 2, 7], allows to investigate very small part of search space, because in such cases well-defined, determinate action sequence is used, and this sequence performs very limited analysis of feature space (during optimization of objective criterion $V(P^*)$ small number of sampling points is investigated). Method of complete enumeration also applies well-defined action sequence, which investigates all points of search space, and because of significant time costs its application is impossible when there is significant number of features M in initial set $S =< P,T >$.

In stochastic methods (evolutionary, multiagent, etc. [1, 7, 12]) strategies, based on probabilistic search and examination of randomly selected points *Xe* of search space *XS*, are used. It allows to investigate the greater part of search space in comparison with deterministic methods. But methods, which use stochastic strategy, are subjected to recirculation in the areas of local extrema (if local optima areas are found on some iteration, then solutions are subsequently concentrated around such areas). Regardless mechanisms of local extrema leaving (for example, usage of mutation operator in evolutionary search methods or procedure of agent restarting in agent-oriented methods of computational intelligence), concentration of some solutions (control points) around local extrema areas is present on the following search iterations too. It reduces search efficiency (the same areas of feature space are investigated), raises time of its execution on computing machine, and in some cases does not allow to find acceptable solution.

Therefore for elimination of the presented defects in the developed parallel stochastic method of feature selection it is proposed to use combination of different strategies of stochastic search (methods based on evolutionary and multiagent approaches [7, 12]), which should be implemented at different nodes of parallel system. Application of different strategies, based on probabilistic approach, will allow to significantly extend search space coverage in comparison with the existing methods [1, 2, 7]. Application of parallel computings will allow to reduce search time and, as consequence, raise practical threshold of applicability of feature selection methods for big data processing.

In the proposed parallel method of big data reduction during initialization phase at the main core Pr_0 data reduction process is started, input data is read from user (data sample $S =< P,T >$, parameters of method, etc.).

Then feature selection methods are allotted between cores $Pr_1, Pr_2, ..., Pr_{NPr-1}$ of computing system, and also access to the input sample $S =< P,T >$ is passed. At that it is proposed to apply one core Pr_{NPr-2}, Pr_{NPr-1} for low iterative methods (based on decision trees and associative rules) correspondingly. Between the rest of cores $Pr_1, Pr_2, ..., Pr_{NPr-3}$ more complex data reduction methods, which are based on evolutionary and multiagent approaches, are uniformly allotted. Then for example, in the case of system with 24 cores, feature selection methods are allotted between cores of computing system in the following way: Pr_0 – main process, Pr_1 – Pr_6 – feature selection based on evolutionary search with feature grouping [21], Pr_7 – Pr_{11} – feature selection based on evolutionary method with feature clusterization [22], Pr_{12} – Pr_{16} – feature selection based on multiagent search with direct connection between agents [23], Pr_{17} – Pr_{21} – feature selection based on multiagent search with indirect connection between agents [23], Pr_{22} – feature selection based on decision trees [21], Pr_{23} – feature selection based on associative rules [24].

After that at every node $Pr_1, Pr_2, ..., Pr_{NPr-1}$ feature reduction process for the sample $S =< P,T >$ is performed. To raise space search coverage uniformity during feature selection, different methods of stochastic search are used at different nodes of parallel system. For these purposes it is proposed to use the following methods:

– evolutionary search with feature grouping [21] is based on usage of prior information about feature significance during feature selection process. As prior information for evolutionary search, estimations $V(p_m)$ of individual informativeness of features p_m, which are calculated at the method initialization stage, are used in evolutionary operators of crossover and mutation;

– evolutionary method with feature clusterization [22] as the previous method uses estimations $V(p_m)$ of individual informativeness of features p_m for evolutionary optimization. In addition to estimations $V(p_m)$, information about location of features p_m in observation space is also used. It allows to group features during search process and to form control points Xe_k using features, which are located distantly in feature space, eliminating in such a way combinations with interdependent features from consideration;

– multiagent method with direct connection between agents [23] is based on application of agent technologies of computational intelligence without usage of heuristic search procedures, applies agent approach for data exchange, allowing to investigate search space areas with

perspective control points in more detail. This method can be efficiently applied for feature selection during classification model synthesis (when output parameter has discrete values);

– multiagent method with indirect connection between agents [23] applies evolutionary operators of crossover and mutation at agent simulation phase, allowing to investigate search space more efficiently in comparison with the known multiagent methods and to reduce search time. This method allows to select feature combination with the highest significance when features are interdependent, is not subjected to recirculation in local optima, does not use greedy search strategy and does not make additional demands for objective function shape;

– feature selection method based on decision tress [21] which estimates informativeness of feature set Xe_k using decision tress which are synthesized during search process. Method allows to estimate individual and group informativeness of features p_m of training sample $S=<P,T>$ using structure of synthesized tree, performs phases of addition of root features and tree truncation. Such method is not highly iterative and resource-demanding, so it can be applied for finding of combination of the most significant features, when time and computing resources are limited, or can use small number of nodes Pr_j, when parallel systems are applied;

– feature selection method based on associative rules [24] can be efficiently used for informative feature selection from data samples $S=<P,T>$, generated based on transactional data sets $D=\{T_1,T_2,...,T_{N_D}\}$, where every element (transaction) T_j, $j=1,2,...,N_D$ contains information about some interrelated events, objects or processes. At that transactions T_j of data set *D* represent list from some element set. In the feature selection method based on associative rules [24] estimation of feature informativeness $V(p_m)$ is performed using information about interest level of extracted association sets (associative rules).

During feature selection in the proposed PMBDR processes $Pr_1,Pr_2,...,Pr_{NPr-1}$ can exchange signals with the main process Pr_0. At that signal Sgn_{inj} about completion of feature selection on the *j*-th process Pr_j is received from processes $Pr_1,Pr_2,...,Pr_{NPr-1}$ by the main process Pr_0, when one of the given stopping criteria is satisfied. For such purposes the following criteria can be used: $Crit_1$ – successful finding of combination of features P^*, which satisfies the given minimal acceptable search conditions (for example: $V(P^*)\le V_{\min}$, where $V_{\min}$ is a minimal acceptable value of feature set optimality criterion, which was set by user at initialization phase); $Crit_2$ – maximum acceptable number of search iterations; $Crit_3$ – maximum acceptable number of objective function value computing. The other criteria can be also used as stopping criteria.

Signals Sgn_{outj} about the necessity of feature selection procedure completion on the specific process Pr_j are received by the processes $Pr_1,Pr_2,...,Pr_{NPr-1}$ from the main process Pr_0. Signals Sgn_{outj} can be forwarded by the main process in the following situations:

– if signal Sgn_{inj} about successful search completion, when criterion $Crit_1$ is satisfied, is received from any process $Pr_1,Pr_2,...,Pr_{NPr-1}$. In this case the further search at the other processes loses meaning, because acceptable solution is found at the process Pr_j;

– if signal Sgn_{inj} about search completion, when criterion $Crit_2$ or $Crit_3$ is satisfied, is received from the set of processes $Pr_1,Pr_2,...,Pr_{NPr-1}$ (for example, not less than from the half of processes $Pr_1,Pr_2,...,Pr_{NPr-1}$). In this case the further feature selection procedure at the remaining processes is not advisable, because of idle time of the bigger part of computational system nodes, and current information is sent from processes $Pr_1,Pr_2,...,Pr_{NPr-1}$ to the process Pr_0;

– if maximum acceptable search time $Crit_4$ is reached, at every process $Pr_1,Pr_2,...,Pr_{NPr-1}$ current search iteration is finished and information about set of investigated control points $Xe\in XS$ and corresponding values of objective function $V(Xe)$ is sent to the main process.

During search process information $Inf_k=<Xe_k,V(Xe_k)>$ about points $Xe\in XS$ of search space XS which were investigated at every core $Pr_1,Pr_2,...,Pr_{NPr-1}$ is saved. It allows to estimate spatial location of solutions and its movement during search process. Besides it, such approach allows not to perform iterative estimation (calculation of values of objective function $V(Xe)$) of solutions $Xe\in XS$, which were estimated on the previous iterations, reducing search time in such a way.

Processes $Pr_1,Pr_2,...,Pr_{NPr-1}$ during data reduction procedure realization can efficiently exchange information $Inf_k=<Xe_k,V(Xe_k)>$ between each other. It allows to organize parallel search (similarly to island model [6, 7]) at group of processors, which are used for implementation of the same feature selection method, and also not to investigate iteratively control points which have been already estimated.

When feature selection procedure is finished at the nodes $Pr_1,Pr_2,...,Pr_{NPr-1}$, phase of collection and distribution of current information about optimization process is performed. At this phase information $InformPr_j$ about sets of investigated control points is received by the main process Pr_0 from the processes

$Pr_1, Pr_2, ..., Pr_{NPr-1}$. Such information contains coordinates of control points in search space, corresponding values of objective function, and also secondary information about methods, for which these points were estimated):

$$InformPr_j = \{Inform_1, Inform_2, ..., Inform_{N\chi j}\},$$

$$Inform_k = <\chi_k, V(\chi_k), InformLI(\chi_k), InformM(\chi_k)>.$$

After information $InformPr_j$ is received from all processes $Pr_1, Pr_2, ..., Pr_{NPr-1}$, it is combined on the main process Pr_0: $Inform = \bigcup_{j=1}^{N_{Pr}} InformPr_j$. It is significant that during combination of sets $InformPr_j$ situations, when the same solution χ_k is presented in different sets, can happen. In this case list of all methods, where solution χ_k took part, is saved to variable $InformM(\chi_k)$. Value of objective function $V(\chi_k)$ is chosen as the best from estimations, which were obtained at different processes. Different values of objective function $V(\chi_k)$ for the same point χ_k of search space can appear, because of general usage of errors of models, which were built based on feature set, which corresponds to the point χ_k, as objective function. At that artificial neural networks or other models of computational intelligence, can be used as such models. Training of such models is performed using probabilistic procedures, explaining possible differences in estimations $V(\chi_k)$ for the same values of χ_k.

After information $InformPr_j$ about sets of investigated control points χ_k is received, its concentration is estimated for the current solution set $R(iter) = \{\chi_1, \chi_2, ..., \chi_{N\chi}\}$ around local extrema $\nu_{conc}(iter)$ at the main process Pr_0. Calculation of estimations of solution concentration around local extrema $\nu_{conc}(iter)$ is performed for the purpose of defining of uniformity of coverage of search space XS during feature selection process. If there are situations when the majority of solutions χ_k is grouped in small areas of local optima, it is proposed to add extra control points, located outside of local extrema, to new solution set $R(iter+1)$.

For estimation of solution concentration $\nu_{conc}(iter)$ the current solution set $R(iter) = \{\chi_1, \chi_2, ..., \chi_{N\chi}\}$ should be divided into groups (clusters) $Cl(iter) = \{Cl_1, Cl_2, ..., Cl_{NCl}\}$, depending on its spatial location. For this purpose well-known cluster analysis methods should be applied [7].

Then for estimation of solution concentration around local extremum, the following criteria are calculated:

1) average distance $dC(Cl_c)$ between solutions in the specific cluster (1):

$$dC(Cl_c) = \frac{2}{|Cl_c|(|Cl_c|-1)} \sum_{k=1}^{|Cl_c|} \sum_{u=k+1}^{|Cl_c|} d(\chi_k, \chi_u),\ \chi_k, \chi_u \in Cl_c, \quad (1)$$

where distance $d(\chi_k, \chi_u)$ between points χ_k and χ_u of the search space XS, which belong to the cluster Cl_c, is calculated using expression (2):

$$d(\chi_k, \chi_u) = \frac{1}{M} \sum_{m=1}^{M} |g_{mk} - g_{mu}|; \quad (2)$$

2) dispersion $DC(Cl_c)$ of the solution χ_k within the cluster Cl_c represents average distance from the center $\overline{\chi_c}$ to solutions χ_k, belonging to the cluster Cl_c (3):

$$DC(Cl_c) = \frac{1}{|Cl_c|} \sum_{\chi_k \in Cl_c} d(\chi_k, \overline{\chi_c}), \quad (3)$$

where distance $d(\chi_k, \overline{\chi_c})$ between solution χ_k and center of the c-th cluster $\overline{\chi_c} = \{\overline{g_{1Clc}}, \overline{g_{2Clc}}, ..., \overline{g_{MClc}}\}$ is calculated using expression (4):

$$d(\chi_k, \chi_c) = \sqrt{\sum_{m=1}^{M} (g_{mk} - \overline{g_{mClc}})^2}, \quad (4)$$

where the m-th coordinate $\overline{g_{mClc}}$ of the c-th cluster center is calculated using formula (5):

$$\overline{g_{mClc}} = \frac{1}{|Cl_c|} \sum_{\chi_k \in Cl_c} g_{mk}. \quad (5)$$

The lower values of the criteria $dC(Cl_c)$ and $DC(Cl_c)$ corresponds to the higher grouped solutions, located in the c-th cluster Cl_c;

3) average cluster distance $dC(iter)$ between solutions on the current search iteration $iter$ (6):

$$dC(iter) = \frac{\sum_{c=1}^{N_{Cl}} |Cl_c| dC(Cl_c)}{\sum_{c=1}^{N_{Cl}} |Cl_c|} = \frac{1}{N_\chi} \sum_{c=1}^{N_{Cl}} |Cl_c| dC(Cl_c). \quad (6)$$

Criterion $dC(iter)$ characterizes average distance between different control points on the current iteration $iter$ within central cluster;

4) average cluster dispersion $DC(iter)$ of solutions on the current search iteration $iter$ (7):

$$DC(iter) = \frac{1}{N_\chi} \sum_{c=1}^{N_{Cl}} |Cl_c| DC(Cl_c). \quad (7)$$

The lower values of the criteria $dC(iter)$ and $DC(iter)$ corresponds to the solutions (control points), which are higher grouped around local optima on the current iteration $iter$;

5) coefficient of solution concentration on the current iteration (8):

$$\nu_{conc}(iter)=\frac{dC(iter)}{d(iter)}, \quad (8)$$

where average distance $d(iter)$ between all solutions on the current iteration is calculated using expression (9):

$$d(iter)=\frac{2}{N_\chi(N_\chi-1)}\sum_{k=1}^{N_\chi}\sum_{u=k+1}^{N_\chi}d(\chi_k,\chi_u),\ \chi_k,\chi_u\in R(iter). \quad (9)$$

Using estimates of solution dispersion $DC(iter)$, coefficient of solution concentration on the current iteration can be calculated using expression (10):

$$\nu_{conc}(iter)=\frac{DC(iter)}{D(iter)}, \quad (10)$$

where dispersion $D(iter)$ of solutions on the current iteration can be calculated using expression (11):

$$D(iter)=\frac{1}{N_\chi}\sum_{k=1}^{N_\chi}d\left(\chi_k,\overline{\chi}\right), \quad (11)$$

where distance $d\left(\chi_k,\overline{\chi}\right)$ between solution χ_k and central solution $\overline{\chi}$ on iteration $iter$ is calculated using expression (12):

$$d\left(\chi_k,\overline{\chi}\right)=\sqrt{\sum_{m=1}^{M}\left(g_{mk}-\overline{g_m}\right)^2}, \quad (12)$$

where m-th coordinate $\overline{g_m}$ of central solution $\overline{\chi}$ is calculated using expression (13):

$$\overline{g_m}=\frac{1}{N_\chi}\sum_{k=1}^{N_\chi}g_{mk}, \quad (13)$$

Value of criterion $\nu_{conc}(iter)$ belongs to the interval $(0;1)$. The closer the value of this criterion is to 1, the lower grouped solutions are (correspondingly, search space is covered by control points in more uniform way). The values of criterion $\nu_{conc}(iter)$, which are close to zero, evidence significant solution concentration around local extrema.

6) maximum number of control points $\chi_k\in Cl_c$, grouped within one local extremum (in the area of the cluster Cl_c):

$$N_{\max Ncl}(iter)=\max_{c=1,2,...,N_{Cl}}\left(\left|Cl_c\right|\right). \quad (14)$$

The bigger value criterion $N_{\max Ncl}(iter)$ has, the bigger number of solutions is grouped within one local extremum and correspondingly the lower location uniformity solutions have in the search space XS on iteration $iter$.

When decisions about excessive concentration of control points around some areas of local extrema are made, it is proposed to use integral criterion $\nu_{conc}(iter)$ and also criterion $N_{\max Ncl}(iter)$ of maximum number of control points $\chi_k\in Cl_c$, which are grouped within one local extremum, in the developed data reduction method.

If even one criterion has value which is over the given threshold ($\nu_{conc}(iter)>\nu_{concThr}$ or $N_{\max Ncl}(iter)>N_{\max NclThr}$), the decision about excessive concentration of control points within local extrema areas is made. It is proposed to add extra control points $R(iter+1)$, which are located outside of local extrema, to the current solution set to raise uniformity of search space coverage. The number of extra control points is proposed to set equal to the number of solutions in the set $R(iter)$.

For this purpose average values $\overline{g_m}$ of m-th coordinates of central solution $\overline{\chi}$ are calculated using expression (13). Values $\overline{g_m}$ demonstrate local concentration of solutions in projection of m-th feature axis. The closer the value $\overline{g_m}$ is to 1, the bigger the number of solutions χ_k characterizes m-th feature as informative. Similarly when $\overline{g_m}\to 0$, m-th feature is considered as non-informative in solution set $R(iter)=\{\chi_1,\chi_2,...,\chi_{N\chi}\}$.

Then using calculated values $\overline{g_m}$ and randomly generated numbers $rand[0;1]$, new solutions $\chi_a=\{g_{1a},g_{2a},...,g_{Ma}\}$ should be found. m-th coordinate g_{ma} of these solutions can be calculated using expression (15):

$$g_{ma}=\begin{cases}1, & rand[0;1]>\overline{g_m};\\ 0, & rand[0;1]\le\overline{g_m}.\end{cases} \quad (15)$$

Thus m-th coordinate g_{ma} of new control point χ_a will have bigger probability of possessing the value $g_{ma}=b$, if lower number of solutions χ_k in the set $R(iter)=\{\chi_1,\chi_2,...,\chi_{N\chi}\}$ has the same value of the m-th coordinate. Such approach will allow to generate new solutions χ_a, which are significantly distant from the current solution set $R(iter)=\{\chi_1,\chi_2,...,\chi_{N\chi}\}$, in such a way reducing solution concentration during data reduction process and raising uniformity of search space coverage.

It is proposed to use prior information on individual informativeness of features, in such a way allowing to secure values of the m-th coordinates, corresponding to the features and having significant effect on output parameter values. For this purpose at the initialization stage it is proposed to calculate values of individual informativeness $V(p_m)$ of features p_m, which characterize correlation between feature p_m and output parameter T.

Values of pair correlation coefficient, feature entropy, sign correlation criterion can be used as estimates of $V(p_m)$ [7, 21]. If prior information on individual significance of features is used, expression (15) can be modified in the following way:

$$g_{ma} = \begin{cases} 1, \ rand[-1;1] > (\overline{g_m} - V(p_m)); \\ 0, \ rand[-1;1] \le (\overline{g_m} - V(p_m)). \end{cases} \quad (16)$$

Such approach allows to raise probability of generation of new solutions χ_a with genes g_{ma}, corresponding to the highly informative features p_m. At that probabilistic approach maintains possibility of generation of solutions χ_a, which are remotely situated from the current set of control points $R(iter) = \{\chi_1, \chi_2, ..., \chi_{N\chi}\}$.

After the generation of the necessary number of additional control points χ_a, sets $R(iter) = \{\chi_1, \chi_2, ..., \chi_{N\chi}\}$ and $R_a(iter) = \{\chi_{a1}, \chi_{a2}, ..., \chi_{aN\chi}\}$ are united into the set $R(iter+1)$.

After that data reduction procedure is restarted at the nodes $Pr_1, Pr_2, ..., Pr_{NPr-1}$. At that new initial sets of solutions $R_j(iter+1)$ for the corresponding feature selection methods are formed based on the set $R(iter+1)$. The sets $R_j(iter+1)$ are formed at the nodes Pr_j in the following way. At the beginning solutions $\chi_{elit,j}$ with the highest values of objective function $V(\chi_{elit,j}) = \max\limits_{\chi_k \in R_j(iter)} (V(\chi_k))$ on the previous iteration are selected. Thus elite solutions $\chi_{elit,j}$ with the best values of objective function V are automatically transferred into the next population $R_j(iter+1)$, enabling usage of results which were got on the previous iterations and approaching of new initial search points to optimal ones. Number of elite solutions $\chi_{elit,j}$, which are automatically transferred into the next search iteration, is set by user at method initialization stage, and generally is equal to 2–5% of total number of solutions, which are used at the separate node of computation system. Then solutions χ_k are randomly chosen from the set $R(iter+1)$. The overall number of solutions is set according to the requirements of feature selection method, which is used at the *j*-th node Pr_j of computation system.

It is significant that besides the values of coordinates of control points χ_k, mathematical support, which is used at nodes Pr_j, has access to information about all solutions which were estimated earlier and its corresponding values of objective functions $<\chi_k, V(\chi_k)>$, allowing to avoid recurrent estimation of solutions which were estimated earlier and to reduce search time.

Then using initial sets of control points $R_j(iter+1)$ at the nodes $Pr_1, Pr_2, ..., Pr_{NPr-1}$, data reduction procedures are realized.

The described process should be continued till one of the following stopping criteria will be achieved: $Crit_1$ – successful finding of combination of features P^*, which satisfies the given minimal acceptable search conditions; $Crit_5$ – exceeding of total maximum permissible search time on the parallel system; $Crit_6$ – maximum permissible number of restart of data reduction procedure at the nodes $Pr_1, Pr_2, ..., Pr_{NPr-1}$.

Thus the proposed PMBDR proposes to use different strategies of stochastic search, based on evolutionary and multiagent approaches and realized at different nodes of parallel system. Usage of different strategies, based on probabilistic approach, allows to considerably extend coverage of search space. It is proposed to add control points, which are located outside of local optima, to the current solution set in the proposed method for raising of search space coverage uniformity during search process. Application of parallel computing in the proposed method makes it possible to reduce search time and, as consequence, to raise practical threshold of feature selection methods applicability for big data processing.

The criteria system, which enables to estimate concentration of control points around local extrema, was proposed. Calculation of solution concentration estimates in the developed criteria system is based on the spatial location of control points in the current solution set. The proposed criteria system can be used in stochastic search methods to monitor situations of excessive solution concentration in the areas of local optima and, as a consequence, to increase the diversity of the solution set in the current population and to cover the search space by control points in a more uniform way during optimization process.

4 EXPERIMENTS

For experimental investigation of the efficiency of the proposed method application for feature selection and pattern recognition problems solving, the vehicle recognition task [21], which is characterized by the data sample containing 10000 observations, was used. Every sample observation presents vehicle image and is formed by values of 26 features and 1 output parameter which defines, if observation belongs to the considered class.

At the beginning of the pattern recognition problem solution process concerning the considered task, feature selection methods were applied. These methods allowed to get informative feature set, which was considered as the most informative. It allowed to solve feature selection problem on the one hand, and on the other hand to select informative feature set, which then was used for model synthesis realization. Every such model was then used for vehicle recognition based on the classification with two classes: if observation belongs to the corresponding class (motorcyclist, passenger car, truck, bus, minivan or an object which is not recognized) or no. That is totally 5 such models were synthesized.

The following methods besides PMBDR, proposed in the paper, were considered as feature selection methods: PCA, GMDH, CMES, MARF, MMICA, MMDCA.

Let's consider criteria, which were used for investigation of the obtained results of feature selection problem solution.

Number of features k, which formed informative feature set, was considered as basic estimation criterion. Taking into account that the problem was solved for 5 variants of the problem statement separately, the number of selected features was presented as rounded average value, as well as interval of these values (minimum and maximum values, that is the lowest and the largest cardinal number of the set of informative features selected as a result of the corresponding method application).

Taking into account shortcomings which should be eliminated in the proposed method, it is necessary not only to estimate obtained results for the given conditions, but also to estimate the depth of coverage of search space XS. Therefore the corresponding criterion $\nu_{Deep}(iter, XS)$ should be calculated using expression (17):

$$\nu_{Deep}(iter, XS) = \frac{N(iter)}{N(XS)}. \quad (17)$$

As a set XS presents collection of all possible feature combinations p_m ($m = 1, 2, ..., M$), obtained from the initial feature set P ($|P| = M$), quantity $N(iter)$ can be calculated in the following way (18):

$$N(XS) = |XS| = 2^M - 1. \quad (18)$$

The number of unique sampling points, estimated on the current iteration, can be considered as alternative to the criterion presented above. It makes possible to demonstrate convergence of the method in absolute representation and in particular to compare it with methods, which are characterized by finding of local optima instead of global ones.

The results of the feature selection phase directly influence on the quality of pattern recognition solutions, therefore the following criteria were set as investigation criteria for the obtained pattern recognition problem solutions:

– recognition error E, which is defined in the following way:

$$E = \frac{Q_{ic}}{Q}. \quad (19)$$

– method operating time T_c, which is needed by method to achieve an acceptable solution.

The software based on the proposed method was written in C language using MPI and CUDA libraries: data exchange between the core and the rest of the cluster nodes was performed using multiple MPI exchange functions (Bcast, Gather, Scatter, Reduce).

For realization of parallel computing in experimental investigation, hardware of Software Tools Department of Zaporizhzhia National Technical University was used.

For investigation of PBMDR, evolutionary search with feature grouping, multiagent method with indirect and direct connection between agents and also method of feature selection based on associative rules were used at different nodes of parallel system as the most suitable for the considered task solving based on the preliminary comparison.

5 RESULTS

Table 1 presents results of informative feature set selection based on feature selection methods, expressed as interval and average value of cardinal number of such set (for alternatives of the forecasted recognition class).

Table 1 – Number of features which were selected by feature selection methods during vehicle recognition

№	Feature selection method	Values of comparison criteria		
		K_{min}	K_{max}	K
1	PCA	12	13	12,4
2	GMDH	11	12	11,2
3	CMES	10	11	10,4
4	MARF	11	13	12,2
5	MMICA	10	11	10,2
6	MMDCA	10	11	10,4
7	PMBDR	10	11	10,2

The dependence of number of unique sampling points on the current iteration number for CMES is presented in the Figure 1.

The analogous presentation of the number of unique sampling points, investigated on the current iteration, for MMDCA is showed in the Figure 2.

The change of unique sampling points number during execution of parallel method of big data reduction is presented in the Figure 3.

In the Figure 4 the diagram, which presents distribution of vehicle recognition error level during investigation of pattern recognition problem depending on the feature selection method which was applied on the corresponding stage, is presented.

The diagram, which presents ratio of vehicle recognition operation time on application of different feature selection methods, is presented in the Figure 5.

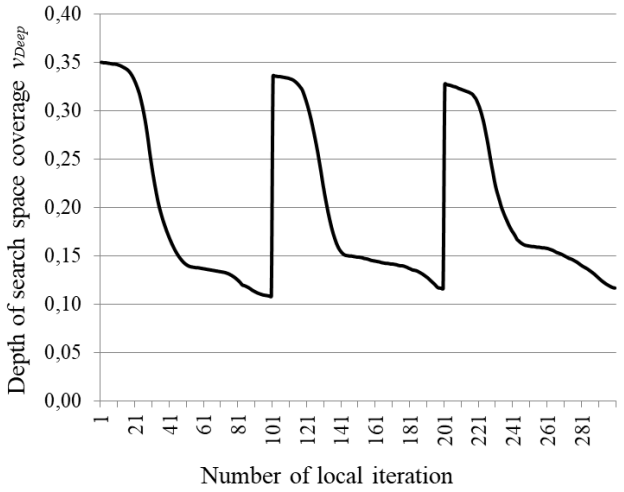

Figure 1 – Graph of dependence between number of unique sampling points and number of CMES iteration

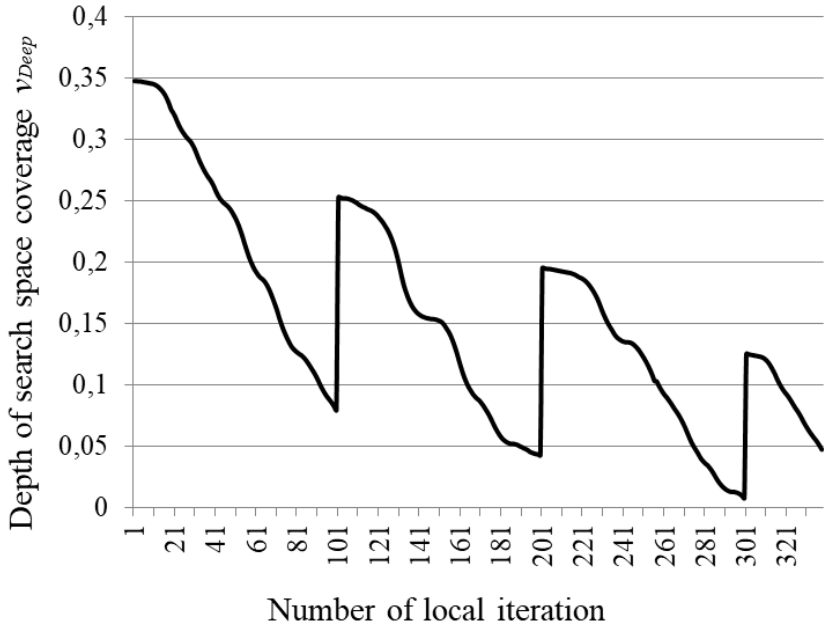

Figure 2 – Graph of dependence between number of unique sampling points and number of MMDCA iteration

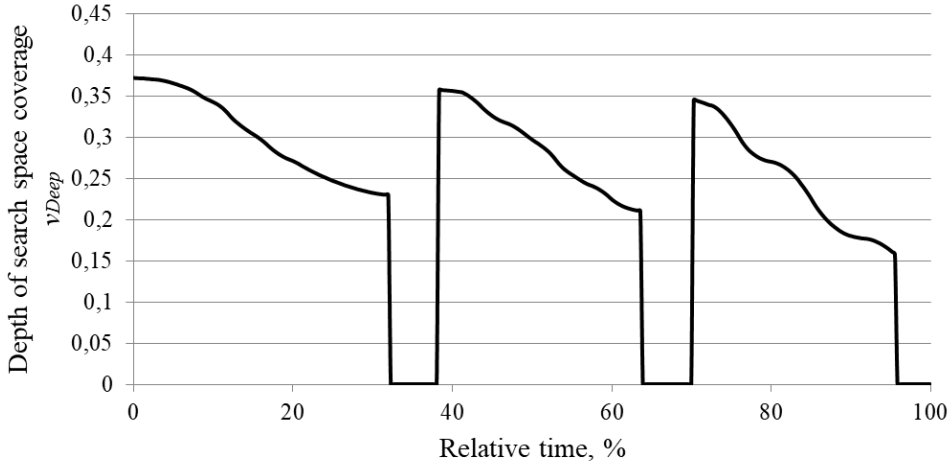

Figure 3 – Graph of dependence between number of unique sampling points and number of iteration of parallel method of big data reduction

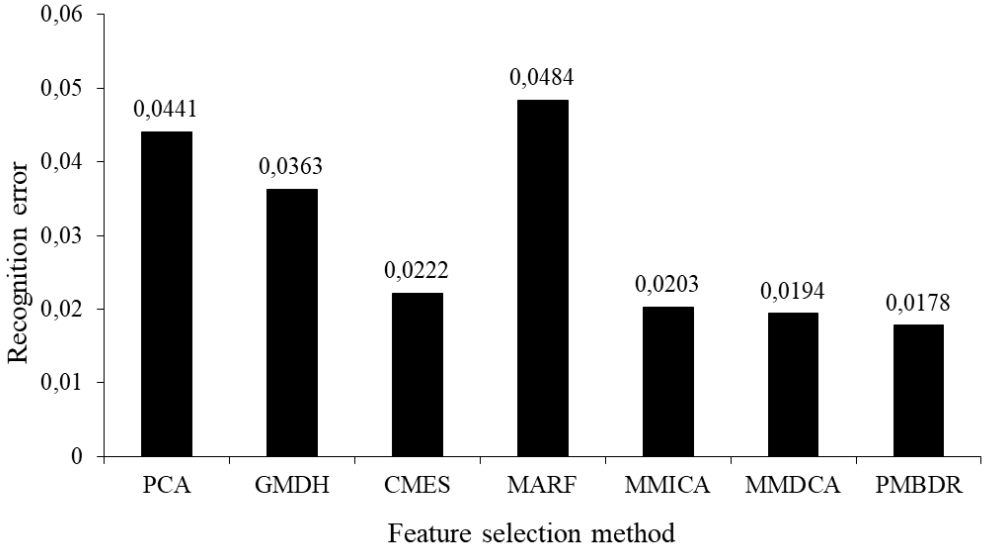

Figure 4 – Diagram of vehicle recognition error distribution

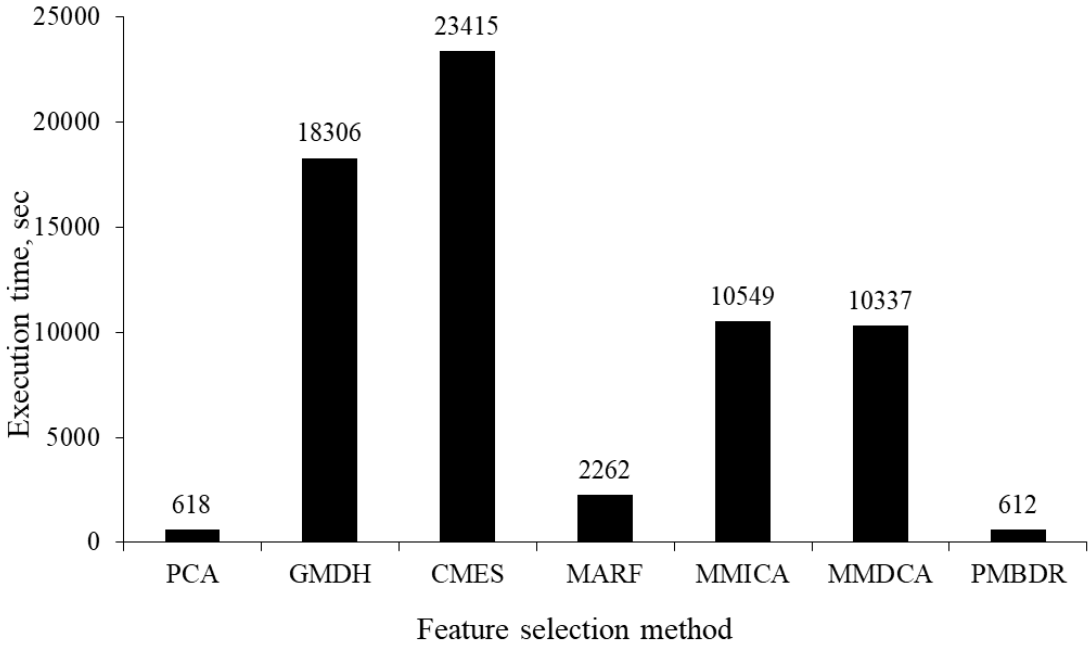

Figure 5 – Diagram of vehicle recognition operation time distribution

6 DISCUSSION

The results of experiments, shown in the Table 1, demonstrated that the set with the lowest number of informative features was selected by MMICA and PMBDR (10.2 on average). MMDCA and CMES were characterized by almost the same set (10.4).

Figures 1–3 show the dependence of the search space coverage depth on the current iteration number for 3 methods with the best recognition results (Table 1, Figures 4–5).

As can be seen from the Figure 1, during CMES execution process high initial values of search space coverage decrease rapidly (2.36 times during 30 iterations), leading to shortcoming when significant coverage of the search space is implemented only at the initial stage, so almost half of iterations is performed over a set of unique points that cover only 15% of the search space. At the same time the initial coverage of the search space does not decrease significantly after iteration set is repeated, that is the repeated implementation of iterations begins with almost the same set of unique points (33–35 %).

Fig. 2 demonstrates the same presentation of the search space for MMDCA. In this case, there is no quick reduce of the number of unique points as it was in the evolutionary search (search space coverage is decreased till 15% during 3/4 of iterations). However, this method leads to the following situation: when iteration set is repeated, the initial coverage of the search space is constantly reduced, and more than 30% of unique points is considered only during the first 27 iterations, leading to the fact that at the last stage of this method implementation a small set of points (compared with initial set) is considered.

PMBDR (Fig. 3) actually allowed to inherit positive characteristics of search space investigation obtained by the methods considered above. Besides it PMBDR extends these advantages, adding extra control points. To represent the depth of search space coverage using parallel computing additional indicator, relative time, was used. It is caused by the fact, that realization of iterations of different strategies at different nodes of parallel system lasts different time, and so it is necessary to normalize this presentation for illustration of overall coverage of search space. The relative time is expressed as a percentage ratio of current time to total running time of the entire parallel system.

As can be seen from the Figure 3, when iteration set is repeated, the initial search space is comparable each time (34.5–37.2 %). If realization of each such repetition is considered separately, it is noticeable that the coverage depth does not reduce as quickly as in the evolutionary search, as a result reducing probability of falling into local optima. Every repetition of iteration set ends with reduction of search space depth to 0, because computations on the main core Pr_0 are realized at that period of time, thus search is not performed.

The results of vehicle recognition problem solving (recognition error and execution time), presented in the Figures 4 and 5, demonstrated, that the best values corresponded to the PMBDR, proposed in the paper.

The developed method allowed to get recognition error of 0.0178, which is 8.2% and 12.3 % more accurate than MMDCA and MMICA correspondingly, 19.8% more accurate than CMES. Thus the best recognition results in terms of accuracy were demonstrated by the methods, which selected feature sets with the lowest cardinal number for the given task.

At the same time, the proposed method proved to be the best in terms of execution time, the value of which was 612 sec. PCA demonstrated comparable speed of work: it has performed recognition process 6 sec. faster. However, its recognition error was almost 2.5 times higher than

recognition error of the proposed method. The next result in these terms showed MARF, which made it possible to perform recognition 3.7 times slower than the proposed method, but was characterized by the largest (among the considered methods) recognition error (0.0484). MMDCA and MMICA, having recognition error which is comparable with the proposed method, realized recognition 17.24 and 16.89 times slower.

Thus it can be argued that the proposed parallel method of large data reduction allows to effectively solve the informative features selection problem, which leads to an effective solution of the problem of pattern recognition, besides in comparison with the other methods of informative feature selection the proposed method is implemented faster with the lowest recognition error.

CONCLUSIONS

In this paper the actual task of automation of feature informativeness estimation process in diagnostics and pattern recognition problems was solved.

Scientific novelty of the paper is in the proposed parallel method of big data reduction. This method is based on the proposed criteria system, which allows to estimate concentration of control points around local extrema. Calculation of solution concentration estimates in the developed criteria system is based on the spatial location of control points in the current solution set. The proposed criteria system can be used in stochastic search methods to monitor situations of excessive solution concentration in the areas of local optima and, as a consequence, to increase the diversity of the solution set in the current population and to cover the search space by control points in a more uniform way during optimization process.

Practical significance of the paper consists in the solution of practical problems of pattern recognition. Experimental results showed that the proposed method allowed to select informative feature set and it could be used in practice for solving of practical tasks of diagnostics and pattern recognition.

ACKNOWLEDGMENTS

The work was performed as part of the research work "Methods and means of decision-making for data processing in intellectual recognition systems" (number of state registration 0117U003920) of software tools department of Zaporizhzhia National Technical University and was partially supported by the international project "Internet of Things: Emerging Curriculum for Industry and Human Applications" (ALIOT, registration number 573818-EPP-1-2016-1-UK-EPPKA2-CBHE-JP) funded by the Erasmus+ programme of the European Union.

REFERENCES

1. Jensen R. Computational intelligence and feature selection: rough and fuzzy approaches / R. Jensen, Q. Shen. – Hoboken: John Wiley & Sons, 2008. – 339 p. DOI: 10.1002/9780470377888.
2. Lee J. A. Nonlinear dimensionality reduction / J. A. Lee, M. Verleysen. – New York : Springer, 2007. – 308 p. DOI: 10.1007/978-0-387-39351-3.
3. Mulaik S. A. Foundations of Factor Analysis / S. A. Mulaik. – Boca Raton, Florida : CRC Press. – 2009. – 548 p.
4. Oliinyk A. Production rules extraction based on negative selection / A. Oliinyk // Radio Electronics, Computer Science, Control. – 2016. – Vol. 1. – P. 40–49. DOI: 10.15588/1607-3274-2016-1-5.
5. McLachlan G. Discriminant Analysis and Statistical Pattern Recognition / G. McLachlan. – New Jersey : John Wiley & Sons, 2004. – 526 p. DOI: 10.1002/0471725293.
6. Bow S. Pattern recognition and image preprocessing / S. Bow. – New York : Marcel Dekker Inc., 2002. – 698 p. DOI: 10.1201/9780203903896.
7. Encyclopedia of machine learning / [eds. C. Sammut, G. I. Webb]. – New York : Springer, 2011. – 1031 p. DOI: 10.1007/978-0-387-30164-8.
8. A comparison of approaches to large-scale data analysis / [A. Pavlo, E. Paulson, A. Rasin, D. J. Abadi et al] // International Conference on Management of Data. – 2009. – P. 165–178. DOI: 10.1145/1559845.1559865.
9. The model for estimation of computer system used resources while extracting production rules based on parallel computations / [A. A. Oliinyk, S. Yu. Skrupsky, V. V. Shkarupylo, S. A. Subbotin] // Радіоелектроніка, інформатика, управління. – 2017. – № 1. – С. 142–152. DOI: 10.15588/1607-3274-2017-1-16.
10. Sulistio A. Simulation of Parallel and Distributed Systems: A Taxonomy - and Survey of Tools / A. Sulistio, C. S. Yeo, R. Buyya // International Journal of Software Practice and Experience. Wiley Press. – 2002. – P. 1–19.
11. Shin Y.C. Intelligent systems : modeling, optimization, and control / C. Y. Shin, C. Xu. – Boca Raton: CRC Press, 2009. – 456 p. DOI: 10.1201/9781420051773.
12. Oliinyk A. A. Information Technology of Diagnosis Model Synthesis Based on Parallel Computing / [A. A. Oliinyk, S. A. Subbotin, S. Yu. Skrupsky et al] // Радіоелектроніка, інформатика, управління. – 2017. – № 3. – С. 139–151.
13. Kira K. A practical approach to feature selection / K. Kira, L. Rendell // Machine Learning : International Conference on Machine Learning ML92, Aberdeen, 1–3 July 1992 : proceedings of the conference. – New York : Morgan Kaufmann, 1992. – P. 249–256. DOI: 10.1016/B978-1-55860-247-2.50037-1.
14. Shitikova O. V. Method of Managing Uncertainty in Resource-Limited Settings / O. V. Shitikova, G. V. Tabunshchyk // Радіоелектроніка, інформатика, управління. – 2015. – № 2. – С. 87–95. DOI: 10.15588/1607-3274-2015-2-11.
15. Guyon I. An introduction to variable and feature selection / I. Guyon, A. Elisseeff // Journal of machine learning research. – 2003. – № 3. – P. 1157–1182.
16. Hyvarinen A. Independent component analysis / A. Hyvarinen, J. Karhunen, E. Oja. – New York: John Wiley & Sons, 2001. – 481 p. DOI: 10.1002/0471221317.
17. Oliinyk A. A. Parallel multiagent method of big data reduction for pattern recognition / A. A. Oliinyk, S. Yu. Skrupsky, V. V. Shkarupylo, O. Blagodariov // Радіоелектроніка, інформатика, управління. – 2017. – № 2. – С. 82–92.
18. Bezdek J. C. Pattern Recognition with Fuzzy Objective Function Algorithms / J. C. Bezdek. – N.Y. : Plenum Press, 1981. – 272 p. DOI: 10.1007/978-1-4757-0450-1.
19. Oliinyk A. Parallel computing system resources planning for neuro-fuzzy models synthesis and big data processing / A. Oliinyk, S. Skrupsky, S. Subbotin, O. Blagodariov, Ye. Gofman // Радіоелектроніка, інформатика, управління. – 2016. – № 4. – С. 61–69. DOI: 10.15588/1607-3274-2016-4-8.
20. Zaigham Mahmood Data Science and Big Data Computing: Frameworks and Methodologies / Zaigham Mahmood // Springer International Publishing. – 2016. – P. 332. DOI: 10.1007/978-3-319-31861-5.
21. Субботін С. О. Неітеративні, еволюційні та мультиагентні методи синтезу нечіткологічних і нейромережних моделей : мо-

нографія / С. О. Субботін, А. О. Олійник, О. О. Олійник ; під заг. ред. С. О. Субботіна. – Запоріжжя : ЗНТУ, 2009. – 375 с.
22. Subbotin S. Entropy Based Evolutionary Search for Feature Selection / S. Subbotin, A. Oleynik // The experience of designing and application of CAD systems in Microelectronics : IX International Conference CADSM-2007, 20–24 February 2007 : proceedings of the conference. – Lviv, 2007. – P. 442–443. DOI: 10.1109/CADSM.2007.4297612.
23. Oliinyk A. O. Agent technologies for feature selection / A. O. Oliinyk, O. O. Oliinyk and S. A. Subbotin // Cybernetics and Systems Analysis. – 2012. – Vol. 48, Issue 2. – P. 257–267. DOI: 10.1007/s10559-012-9405-z.
24. Oliinyk A. Training Sample Reduction Based on Association Rules for Neuro-Fuzzy Networks Synthesis / A. Oliinyk, T. Zaiko, S. Subbotin // Optical Memory and Neural Networks (Information Optics). – 2014. – Vol. 23, № 2. – P. 89–95. DOI: 10.3103/S1060992X14020039.

Article was submitted 25.03.2018.
After revision 17.04.2018.

Олійник А. О.[1], Субботін С. О.[2], Льовкін В. М.[3], Ильяшенко М. Б.[4], Благодарьов О. Ю.[5]

[1]Канд.техн.наук, доцент, доцент кафедри програмних засобів, Запорізький національний технічний університет, Запоріжжя, Україна

[2]Д-р техн. наук, професор, завідувач кафедри програних засобів, Запорізький національний технічний університет, Запоріжжя, Україна

[3]Канд.техн.наук, доцент, доцент кафедри програних засобів, Запорізький національний технічний університет, Запоріжжя, Україна

[4]Канд. техн. наук, доцент, доцент кафедри комп'ютерних систем та мереж, Запорізький національний технічний університет, Запоріжжя, Україна

[5]Аспірант кафедри програних засобів, Запорізький національний технічний університет, Запоріжжя, Україна

ПАРАЛЕЛЬНИЙ МЕТОД РЕДУКЦІЇ ВЕЛИКИХ ДАНИХ НА ОСНОВІ СТОХАСТИЧНОГО ПРОГРАМУВАННЯ

Актуальність. Вирішено задачу автоматизації задача автоматизації процесу редукції великих даних при діагностуванні та розпізнаванні образів. Об'єкт дослідження – процес редукції великих даних. Предмет дослідження – методи редукції великих даних.

Мета роботи полягає в створенні паралельного методу редукції даних на основі стохастичних обчислень..

Метод. Запропоновано паралельний метод редукції великих даних. Даний метод ґрунтується на запропонованій системі критеріїв, що дозволяють оцінювати концентрованість контрольних точок близько локальних екстремумів. Обчислення оцінок концентрованості рішень в розробленій системі критеріїв засноване на просторовому розташуванні контрольних точок в поточній множині рішень. Запропонована система критеріїв може використовуватися в методах стохастичного пошуку для відстеження ситуацій надмірної концентрації рішень в областях локальних оптимумів, і, як наслідок, для підвищення різноманітності множини рішень в поточній популяції і більш рівномірного покриття простору пошуку контрольними точками в процесі оптимізації.

Результати. Розроблено програмне забезпечення, яке реалізує запропонований паралельний метод редукції великих даних і дозволяє виконувати відбір інформативних ознак і скорочення великих вибірок даних при синтезі розпізнавальних моделей.

Висновки. Проведені експерименти підтвердили працездатність запропонованого паралельного методу редукції великих даних і дозволяють рекомендувати його для використання на практиці при обробці масивів великих даних для розпізнавання образів. Перспективи подальших досліджень можуть полягати в модифікації існуючих і розробки нових методів відбору ознак на основі розробленої системи критеріїв оцінювання концентрованості контрольних точок близько локальних екстремумів.

Ключові слова: вибірка даних, розпізнавання образів, відбір ознак, паралельні обчислення, критерій інформативності, стохастичний підхід.

Олейник А. А.[1], Субботин С. А.[2], Левкин В. Н.[3], Ильяшенко М. Б.[4], Благодарев А. Ю.[5]

[1]Канд. техн. наук, доцент, доцент кафедры программных средств, Запорожский национальный технический университет, Запорожье, Украина

[2]Д-р техн. наук, професоор, заведующий кафедрой программных средств, Запорожский национальный технический университет, Запорожье, Украина

[3]Канд. техн.наук, доцент, доцент кафедры программных средств, Запорожский национальный технический университет, Запорожье, Украина

[4]Канд. техн. наук, доцент, доцент кафедры компьютерных систем и сетей,Запорожский национальный технический университет, Запорожье, Украина

[5]Аспирант кафедры программных средств,Запорожский национальный технический университет, Запорожье, Украина

ПАРАЛЛЕЛЬНЫЙ МЕТОД РЕДУКЦИИ БОЛЬШИХ ДАННЫХ НА ОСНОВЕ СТОХАСТИЧЕСКОГО ПРОГРАММИРОВАНИЯ

Актуальность. Решена задача автоматизации процесса редукции больших данных при диагностировании и распознавании образов. Объект исследования – процесс редукции больших данных. Предмет исследования – методы редукции больших данных.

Цель работы заключается в создании параллельного метода редукции данных на основе стохастических вычислений.

Метод. Предложен параллельный метод редукции больших данных. Данный метод основывается на предложенной системе критериев, позволяющих оценивать концентрированность контрольных точек около локальных экстремумов. Вычисление оценок концентрированности решений в разработанной системе критериев основано на пространственном расположении контрольных точек в текущем множестве решений. Предложенная система критериев может использоваться в методах стохастического поиска для отслеживания ситуаций чрезмерной концентрации решений в областях локальных оптимумов, и, как следствие, для повышения разнообразия множества решений в текущей популяции и более равномерного покрытия пространства поиска контрольными точками в процессе оптимизации.

Результаты. Разработано программное обеспечение, которое реализует предложенный параллельный метод редукции больших данных и позволяет выполнять отбор информативных признаков и сокращение больших выборок данных при синтезе распознающих моделей.

Выводы. Проведенные эксперименты подтвердили работоспособность предложенного параллельного метода редукции больших данных и позволяют рекомендовать его для использования на практике при обработке массивов больших данных для распознавания образов. Перспективы дальнейших исследований могут заключаться в модификации существующих и разработки новых методов отбора признаков на основе разработанной системы критериев оценивания концентрированности контрольных точек около локальных экстремумов.

Ключевые слова: выборка данных, распознавание образов, отбор признаков, параллельные вычисления, критерий информативности, стохастический подход.

REFERENCES

1. Jensen R., Shen Q. Computational intelligence and feature selection: rough and fuzzy approaches. Hoboken, John Wiley & Sons, 2008, 339 p. DOI: 10.1002/9780470377888.
2. Lee J. A., Verleysen M. Nonlinear dimensionality reduction. New York, Springer, 2007, 308 p. DOI: 10.1007/978-0-387-39351-3.
3. Mulaik S. A. Foundations of Factor Analysis. Boca Raton, Florida, CRC Press, 2009, 548 p.
4. Oliinyk A. Production rules extraction based on negative selection, *Radio Electronics, Computer Science, Control*, 2016, Vol. 1, pp. 40–49. DOI: 10.15588/1607-3274-2016-1-5.
5. McLachlan G. Discriminant Analysis and Statistical Pattern Recognition. New Jersey, John Wiley & Sons, 2004, 526 p. DOI: 10.1002/0471725293.
6. Bow S. Pattern recognition and image preprocessing. New York, Marcel Dekker Inc., 2002, 698 p. DOI: 10.1201/9780203903896.
7. eds. Sammut C., Webb G. I. Encyclopedia of machine learning. New York, Springer, 2011, 1031 p. DOI: 10.1007/978-0-387-30164-8.
8. Andrew Pavlo, Paulson E., Rasin A., Abadi D. J., DeWitt D. J. A comparison of approaches to large-scale data analysis, *International Conference on Management of Data*, 2009, pp. 165–178. DOI: 10.1145/1559845.1559865.
9. Oliinyk A. A., Skrupsky S. Yu., Shkarupylo V. V., Subbotin S. A. The model for estimation of computer system used resources while extracting production rules based on parallel computations, *Radio Electronics, Computer Science, Control,* 2017, No. 1, pp. 142–152. DOI: 10.15588/1607-3274-2017-1-16.
10. Sulistio A., Yeo C. S., Buyya R. Simulation of Parallel and Distributed Systems: A Taxonomy – and Survey of Tools, *International Journal of Software Practice and Experience. Wiley Press*, 2002, pp. 1–19.
11. Shin Y. C., Xu C. Intelligent systems : modeling, optimization, and control. Boca Raton, CRC Press, 2009, 456 p. DOI: 10.1201/9781420051773.
12. Oliinyk A. A., Subbotin S. A., Skrupsky S. Yu., Lovkin V. M., Zaiko T. A. Information Technology of Diagnosis Model Synthesis Based on Parallel Computing, *Radio Electronics, Computer Science, Control*, 2017, No. 3, pp. 139–151.
13. Kira K., Rendell L. A practical approach to feature selection, *Machine Learning : International Conference on Machine Learning ML92, Aberdeen, 1–3 July 1992 : proceedings of the conference*. New York, Morgan Kaufmann, 1992, pp. 249–256. DOI: 10.1016/B978-1-55860-247-2.50037-1.
14. Shitikova O. V., Tabunshchyk G. V. Method of Managing Uncertainty in Resource-Limited Settings, *Radio Electronics, Computer Science, Control*, 2015, No. 2, pp. 87–95. DOI: 10.15588/1607-3274-2015-2-11.
15. Guyon I., Elisseeff A. An introduction to variable and feature selection, *Journal of machine learning research,* 2003, No. 3, pp. 1157–1182.
16. Hyvarinen A., Karhunen J., Oja E. Independent component analysis. New York, John Wiley & Sons, 2001, 481 p. DOI: 10.1002/0471221317.
17. Oliinyk A. A., Skrupsky S. Yu., Shkarupylo V. V., Blagodariov O. Parallel multiagent method of big data reduction for pattern recognition, *Radio Electronics, Computer Science, Control*, 2017, No. 2, pp. 82–92.
18. Bezdek J. C. Pattern Recognition with Fuzzy Objective Function Algorithms. N.Y., Plenum Press, 1981, 272 p. DOI: 10.1007/978-1-4757-0450-1.
19. Oliinyk A., Skrupsky S., Subbotin S., Blagodariov O., Gofman Ye. Parallel computing system resources planning for neuro-fuzzy models synthesis and big data processing*, Radio Electronics, Computer Science, Control*, 2016, Vol. 4, pp. 61–69. DOI: 10.15588/1607-3274-2016-4-8.
20. Zaigham Mahmood Data Science and Big Data Computing: Frameworks and Methodologies, *Springer International Publishing*, 2016, pp. 332. DOI: 10.1007/978-3-319-31861-5.
21. Subbotin S., Oliinyk A., Oliinyk O. Noniterative, evolutionary and multi-agent methods of fuzzy and neural network models synthesis : monograph. Zaporizhzhya, ZNTU, 2009, 375 p. (In Ukrainian).
22. Subbotin S., Oleynik A. Entropy Based Evolutionary Search for Feature Selection, *The experience of designing and application of CAD systems in Microelectronics : IX International Conference CADSM-2007, 20–24 February 2007 : proceedings of the conference*. Lviv, 2007, pp. 442–443. DOI: 10.1109/CADSM.2007.4297612.
23. Oliinyk A. O., Oliinyk O. O. and Subbotin S. A. Agent technologies for feature selection, *Cybernetics and Systems Analysis*, 2012, Vol. 48, Issue 2, pp. 257–267. DOI: 10.1007/s10559-012-9405-z.
24. Oliinyk A., Zaiko T., Subbotin S. Training Sample Reduction Based on Association Rules for Neuro-Fuzzy Networks Synthesis, *Optical Memory and Neural Networks (Information Optics),* 2014, Vol. 23, No. 2, pp. 89–95. DOI: 10.3103/S1060992X14020039.