Розроблено алгоритм навчання системи керування датацентром з використанням системи допусків на значення ознак для кожного з класів розпізнавання. Це дозволяє застосувати нормовані статистики кількості потраплянь ознак у поля допусків для визначення моменту перенавчання системи та підвищити достовірність рішень. Досліджено ефективність використання адитивно-мультиплікативної та ентропійної згорток частинних критеріїв якості функціонування датацентру

Ключові слова: хмарні сервіси, датацентр, інформаційний критерій, машинне навчання, ройовий алгоритм

Разработан алгоритм обучения системы управления датацентров с использованием системы допусков на значения признаков для каждого из классов распознавания. Это позволяет применить нормированные статистики количества попаданий признаков в поля допусков для определения момента переобучения системы и повысить достоверность решений. Исследована эффективность использования аддитивно-мультипликативной и энтропийной сверток частных критериев качества функционирования датацентра

Ключевые слова: облачные сервисы, датацентр, информационный критерий, машинное обучение, роевой алгоритм

OPTIMIZING THE PARAMETERS OF FUNCTIONING OF THE SYSTEM OF MANAGEMENT OF DATA CENTER IT INFRASTRUCTURE

V. Moskalenko
PhD, Senior Lecturer*
E-mail: systemscoders@gmail.com
S. Pimonenko
Postgraduate student*
E-mail: pstsnet@gmail.com
*Department of Computer Science
Sumy State University
Rimsky-Korsakov str., 2, Sumy, Ukraine, 40007

1. Introduction

As of late, corporate applications, e-mail, search engines and e-commerce have been increasingly deployed in the computing environment of cloud data centers. The competitiveness of cloud providers is determined by the capacities of non-failure operation of a data center in a 24/7 format. In this case, as a result of high level of energy consumption in cloud data centers, the providers are trying to maximize efficiency of the use of electricity by redistributing virtual resources and disabling idling physical servers. However, when minimizing costs, the operator of the cloud must ensure compliance of the service quality metrics with the requirements of the service level agreement (Service Level Agreement, SLA) since the violation of the SLA terms results in penalties and outflow of customers [1].

Development of efficient algorithms for resource allocation in a cloud-based environment is complicated by the lack or inaccuracy of information about resource requirements and the performance of a specific task in the loaded heterogeneous nodes of the data center. A priori uncertainty of functional condition of the node in the course of fulfilling the task and inability to accurately estimate the time of its execution may lead to allocation of excess resources that will stay idle, reducing the workload of computing environment, or allocation of insufficient resources, which leads to the overhead costs associated with the process of introduction of additional resources or with migration of the tasks to another node. In addition, monitoring the virtual machines by means of their operating system or a hypervisor is linked to consumption of extra resources and incorrect work of the means of monitoring with significant decrease in performance.

The most efficient tools in maintaining failure-free and efficient performance of a data center are the analytical tools for analysis of archival data of the monitoring of the components of IT-infrastructure and subjectively-statistical studies of assessment of services quality [2]. The purpose of the analytical tools is formation, in the process of machine learning or self-learning, of decisive rules for timely detection (active detection) or forecasting (proactive detection) of abnormal functional states of the components of IT-infrastructure of a data center and abnormal behavior of users or the cloud-based applications. The obtained decision rules allow, in the process of functioning of the data center management system, removing the uncertainty regarding the functional state of the data center's IT-infrastructure and, as a result, increasing the efficiency of actions on reconfiguration and reallocation of resources.

The task of allocation of resources of a data center is multicriterial because it is necessary to simultaneously ensure minimum power consumption, the volume of unused resources and violations of the SLA terms. However, these partial criteria are pairwise conflicting; they have different dimensionality and are non-linear functions of controlled characteristics and configurations of IT-infrastructure of a cloud-based data center. Similar problems in a compromise area have a multitude of the Pareto-optimal solutions. In this case, for the problems of finding the Pareto-optimal solutions for multi-criteria optimization problems, the most promising is to use the ideas and methods of swarm intelligence, which allows increasing the efficiency of solutions for the dynamically changing conditions of functioning.

Thus, the development of new schemes of encoding the attributes of the SLA breach in the algorithms of information-extreme machine learning and the ways of convolution

of the partial criteria of efficiency of functioning of a data center in the population search algorithms are relevant fields of research, aimed at decreasing energy costs and increasing service quality of the end-users.

2. Literature review and problem statement

In the course of designing autonomous data centers with the properties of self-configuration and self-diagnosis, the task of forecasting the functional state of the components of IT-infrastructure and services acquires great importance. To predict the functional state of physical servers when deploying a new virtual machine on it, papers [3, 4] proposed to use a multilayer neural network with computation of the attributes of recognition based on the principle of the bag of words (Bag-of-Words). The input mathematical description of such a predictive model is formed by searching in the archives history of functioning of the data center for the moments of decreased performance of its nodes and the compilation, by the results of cluster analysis of archival data of monitoring, of the dictionary of virtual machines with different templates of resource consumption. The main shortcomings of this approach are ignoring category contextual data and noticeable decrease of efficiency of learning and recognition when expanding the dictionary of attributes and the alphabet of classes. In this case, as shown in articles [5, 6], the reduction in the performance of virtual machines often leads to delays and incorrect work of the monitoring tools embedded in a hypervisor or operating system of the virtual machine that is the reason of false decisions. Papers [7, 8] consider the use of the Bayes classifier to predict reduction in productivity of the nodes of a cloud cluster when assigning the tasks from the queue to it. To avoid overloading and incorrect work of the monitoring system, the authors propose to set a configuration threshold of available resources below the maximum capacity of the physical node. In this case, the set of features of the classifier contains static and dynamic resource usage features of a problem and the node of assignment while the set of classes describes the successful execution of the task. However, the statistical method of machine learning of the classifier limits its effectiveness under conditions of imbalanced and heterogeneous sets of initial data, which occurs in practice. In addition, the task of predicting the decrease in the functional effectiveness of the formed decision rules has not been examined up to now.

Articles [9, 10] explore the use of rough binary coding of observations that enables to unify the presentation of different types of attributes and speed up the processing of input data in the modes of learning and decision making. In this case, paper [10] substantiated the use of logarithmic information criterion of functional efficiency (CFE) of learning for building up highly reliable decisive rules by small training samples. The proposed scheme of encoding the features allows using normalized statistics of the number of entries of the values of features to their fields of control tolerances as a predictive function of efficiency of decision rules [10]. However, this approach implies the selection of a single class, base class, relative to which the upper and lower limits of the control tolerances on the features of recognition is determined. The coded observations of classes characterize deviations of certain level and direction only from the base class, which leads to the loss of part of the meaningful statistical infor-

mation about mutual positioning in space of the attributes of observations of any pair of classes. In order to increase reliability and noise immunity of decision rules, one may consider an alternative scheme of encoding, in which the system of the control tolerances for the meanings of features is constructed relative to each class of functional statuses, but the peculiarities of implementation and efficiency of this approach have not been studied up to now. In this case, the increase in information capability of the decision rules is a relevant task since a number of partial criteria of efficiency of functioning of a data center are calculated by the results of forecasting the functional state of its components and services.

To simplify the multi-criteria problem of control of the parameters of data center operation, it is as a rule reduced to an one-criterion multiparametric optimization problem. Paper [11] proposed heuristic algorithm for the placement of virtual machines on physical servers, based on maximizing the additive convolution of the partial criteria, however, applying this convolution makes sense only in the case of convexity of the set of feasible solutions and, as shown in article [12], in practice this often leads to instability of the decisions. Paper [13] proposes a multiplicative convolution of the partial criteria, which is successfully used in many economic problems; however, the condition for the Pareto-optimal decisions under this convolution, in addition to convexity of the set of feasible solutions, is the concavity of logarithm function of the convolution relative to each of the partial criteria. Articles [14, 15] demonstrated that the adequacy of the method of convolution of the partial criteria is affected by topological distribution of the analyzed alternatives in the space of the partial criteria – convexity or concavity of the Pareto area relative to each pair of coordinates. In this case, different ways of convolution of the partial criteria can lead to significantly different results, and the choice of procedure of the convolution cannot be formalized in full and is determined by the specific of the task, by goals, experiences and intuition of the researcher. Therefore, examination and analysis of the algorithms of formation of a generalized index of efficiency of data center operation are important tasks of information synthesis of the management system of a data center's IT-infrastructure.

3. Aims and objectives of the research

The aim of this work consists in increasing the efficiency of operation of management system, capable of learning, of a cloud-based data center under conditions of heterogeneity of physical nodes and services.

To achieve the set aim, the following tasks are to be solved:

– to design algorithms of information-extreme machine learning of the system of data center management using the system of control tolerances for the values of the features, which are separately determined for each class of functional statuses;

– to develop an algorithm of predicting the moment of reduction in the functional efficiency of prognostic decision rules to determine the moment of re-training the system of data center management;

– to determine the optimal method of convolution of the partial criteria for the task of placement of virtual machines on physical servers in a cloud data center.

4. Algorithms of functioning of intellectual management system of a data center's IT infrastructure

The algorithm of functioning of intellectual system of data center management should include the procedures of monitoring the computing environment and accumulation of knowledge on interrelation between unwanted functional states of the environment with its characteristics and those events in it that are registered and archived. In this case, the main source of information is the key performance indicators (Key Performance Indicators, KPI) key quality indicators KQI (Key Quality Indicator, KQI) and the system messages that are read at the different levels of a cloud system. Fig. 1 shows a generalized structural scheme of the system of data center management.
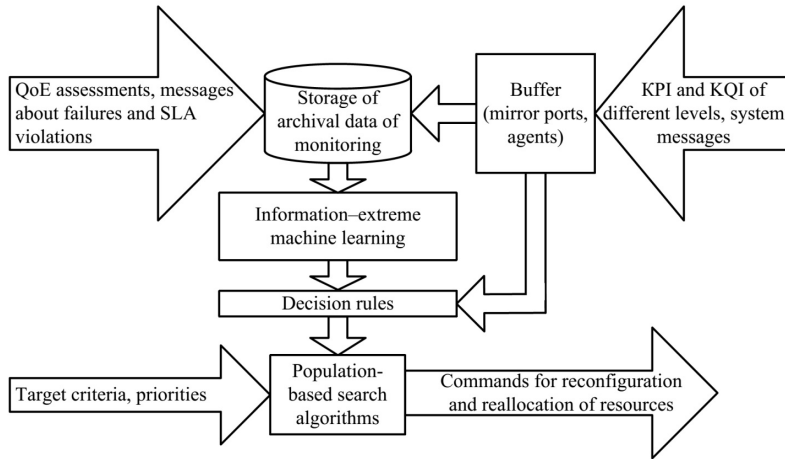

Fig. 1. A generalized structural scheme of data center management system

The assessment of the current functional status can be carried out by checking the fulfillment of conditions of the service level agreement (Service Level Agreement, SLA) containing boundary values of the target parameters (Service Level Objectives, SLO), which include perception of the service quality (Quality of Experience QoE). When using categorical features with the purpose of taking into account the frequency of their occurrence in a functional state of the component of a data center, their frequency re-coding is performed, at which each value of a categorical attribute is represented by the frequencies of their occurrence in each class of functional statuses. In this case, the process of encoding the feature vector is proposed to perform by comparing the value of the i-th attribute to the corresponding lower $A_{L,m,i}$ and upper $A_{U,m,i}$ control tolerances of the class of functional statuses, which are calculated by formulas

$$A_{L,m,i} = \overline{y}_{m,i}\left[1 - \frac{\delta_{m,i}}{\delta_{max}}\right], \qquad (1)$$

$$A_{U,m,i} = \overline{y}_{m,i}\left[1 + \frac{\delta_{m,i}}{\delta_{max}}\right], \qquad (2)$$

where $\overline{y}_{m,i}$ is the averaged value of the feature in the base class; $\delta_{m,i}$ is the parameter of the field of control tolerances for the i-th attribute of recognition; δ_{max} is the maximum value of parameter of the field of control tolerances.

The formation of binary learning matrix $\{x_{m,i}^{(j)}\,|\,i=\overline{1,N};\ j=\overline{1,n};\ m=\overline{1,M}\}$, where N is the number of attributes of recognition, n is the number of feature vectors in in the class and M is the number of classes of functional state, is carried out by the rule

$$x_{m,M^*(i-1)+k}^{(j)} = \begin{cases} 1, & \text{if } A_{L,k,i} \le y_{m,i}^{(j)} \le A_{U,k,i}; \\ 0, & \text{else;} \end{cases} \quad k=\overline{1,M}. \qquad (3)$$

The proposed scheme of encoding (3) allows increasing the variety of binary feature vectors and take into account the level and direction of deviation of distribution of the sampled feature vectors of each pair of patterns among themselves.

Iterative optimization procedure of the parameter of the fields of control tolerances on the attributes of recognition is performed by maximizing information criterion of functional efficiency of learning (CFE), averaged by the alphabet of the classes of recognition, in the couse of training the management system.

$$\{\delta_i^*\,|\,i=\overline{1,n}\} = \arg\max_{G_\delta}\left\{\max_{G_E \cap G_d}\overline{E}\right\}, \qquad (4)$$

where $\overline{E}^{(k)}$ is the averaged value of informational CFE; G_δ is the area of feasible values of the parameter of fields of control tolerances on the values of attributes G_E is the permissible area of defining the function of informational CFE; G_d is the area of feasible values of radius of hyper spherical containers of the classes of functional status that are builted in the binary Hamming space.

The procedure of optimization of the radius of containers is embedded in the process (4) and can be implemented by direct exhaustive search with a given step, as the number of steps of this search is relatively small.

As the criterion of efficiency of machine learning of the classifier, we consider modification of the Kullback information measure [10]:

$$E_m = \left[1 - (\alpha_m^{(k)} + \beta_m^{(k)})\right] \times \log_2\left[\frac{2 - (\alpha_m^{(k)} + \beta_m^{(k)})}{(\alpha_m^{(k)} + \beta_m^{(k)})}\right], \qquad (5)$$

where α_m is the error of the second kind during recognition of observation of the class X_m^o; β_m is the error of the second kind.

The permissible area of defining the function of informational CFE (5) is limited by the inequalities $\alpha_m < 0{,}5$, $\beta_m < 0{,}5$ and $d_m < d(x_m \oplus x_c)$, where d_m is the radius of hyper spherical container of the class $\tilde{O}_m^o$, $d(x_m \oplus x_c)$ is the coded distance between the averaged vector of the class $\tilde{O}_m^o$ and the nearest to it averaged vector of the class $\tilde{O}_n^o$.

To forecast the moment of reduction in the functional efficiency of the information-extreme decisicion rules, at the last stage of machine learning they carry out reflection of the binary matrix $\|x_{m,i}^{(j)}\|$ to the set of free statistics, invariant to a broad family of laws of probability distribution, with the subsequent formation of variational series of extreme order statistics (EOS) $<\{S_{m,n}^*\}>$. Herewith as one-dimensional statistical characteristics of the sample set we consider normalized statistics of the number of entries of the attributes to their fields of control tolerances for n trials [10]

$$S_{m,n} = \sum_{j=1}^{n} \left(\frac{k_{m,j} - \overline{k}_{m,n}}{s_{m,n}} \right)^2, \quad m = \overline{1, M}, \qquad (6)$$

where $k_{m,j}$ is the number of successes at the j-th trial; $\overline{k}_{m,n}$ is the sample mean of the number of successes after n trials; $s_{m,n}^2$ is the sample unbiased dispersion for n trials.

The statistics (6) has distribution χ^2 and depends only on the volume of trials n. Free statistics $S_{m,n}$ is a member of variational series – order statistics, whose rank is determined by the number of the learning step. In this case, the lower and upper confidence limits of the blocks of variational series $< \{S_m^*\} >$ are recommended to calculate by dividing the distances between the neighboring EOS by equal parts.

In the operating mode of the management system, the current class of the functional state of a virtual machine is determined and the current EOS forms. If the EOS extends beyond the variational blocks, then the decision is made about the necessity of retraining the system, as a result of changes in the structure of consumption of resources by virtual machines of a data center.

Each z-th decision P_z regarding distribution of physical resources of a data center between R virtual machines that await a decision in a time window Δt is encoded by the vector of natural numbers

$$P_z = <p_{z,1}, ..., p_{z,r}, ..., p_{z,R}>, \qquad (7)$$

where $p_{z,r}$ is the number of the physical node, to which the r-th virtual machine is assigned.

In operating mode, the management system must take the optimal operational decision P_z^* regarding the allocation of resources of a data center. The estimation of optimality of decision is performed at every step of finding the global extremum of convolution of the partial criteria, which include the level of power consumption and predicted breaches of the SLA conditions, the amount of unused resources and others.

In a general case, for provision of uniformity of the influence from each of the partial criteria to the value of convolution, it is necessary to align the ranges of change in the values of partial criteria by scaling and reducing their values to the dimensionless scale [0.1] by the rule

$$k_i' = \begin{cases} 0, k_i \leq k_i^{min}; \\ \dfrac{(k_i - k_i^{min})}{(k_i^{max} - k_i^{min})}; \quad k_i^{min} < k_i < k_i^{max}, \\ 1, k_i > k_i^{max}; \end{cases}$$

where k_i^{min}, k_i^{max} are the lower and upper limits of permissible area of values of the i-th partial criterion, respectively.

During normalization and formation of the convolution formula, one should take into account that the partial criteria are not unidirectional: part of the partial criteria must be maximised, part – minimized. That is why the partial criteria are divided into stimulants (that must be maximized) and destimulants (which should be minimized). The formula of stimulants normalization can be simplified and have the form

$$k_i' = \frac{k_i}{k_i^{max}}. \qquad (8)$$

The formula of destimulants normalization can be similarly simplified

$$k_i' = \frac{k_i^{min}}{k_i}. \qquad (9)$$

To cover a wide range of problems, we designed a large number of modifications of additive-multiplicative convolutions of both classical form and those based on the Kolmogorov-Gabor polynomial [16]. Papers [15, 16] proposed exponential and entropy convolutions that cover a wide range of tasks, have simple formula for calculation and the minimum number of parameters. With regard to (8) and (9), the additive-multiplicative convolution can be represented in the form

$$F = \sum_{i=1}^{K_1} \omega_i \frac{k_i}{k_i^{max}} + \sum_{i=K_1+1}^{K_2} \omega_i \frac{k_i^{min}}{k_i} + \prod_{i=1}^{K_1} \left(\frac{k_i}{k_i^{max}} \right)^{\omega_i} \prod_{i=K_1+1}^{K_2} \left(\frac{k_i^{min}}{k_i} \right)^{\omega_i}, (10)$$

where ω_i is the weight (priority) of the i-st criterion, for which the condition must be valid

$$\sum_{i=1}^{K_1+K_2} \omega_i = 1.$$

The formula of convolution, built by the principle of information entropy, reflects substantive content of the concept of usefulness as information category and has the following form

$$F = \sum_{i=1}^{K_1} \omega_i \left(\frac{k_i}{k_i^{max}} \right)^{\omega_i} + \sum_{i=K_1+1}^{K_2} \omega_i \left(\frac{k_i^{min}}{k_i} \right)^{\omega_i}. \qquad (11)$$

Weight value of each partial criterion can be calculated based on the methods that are based on the pairwise comparison of criteria or analytical dependence of indicators of the importance of criteria, and formal methods, such as the method of basic criteria or the Churchman-Ackoff method [17]. However, under conditions of a priori uncertainty, according to the Bernoulli-Laplace principle, one may accept the weight of the criteria as the same and equal to

$$\omega_i = \frac{1}{K_1 + K_2}.$$

Heterogeneity, hierarchy of information and telecommunications environment, multifactorial and nonstationary processes of resources consumption predetermine nonlinearity, multi-extremality and high dimensionality of optimization problem. To improve the efficiency of the search for both the optimal solution for allocation of resources and the optimal vector of parameters of predictive decisive rules, it makes sense to use population algorithms, based on the ideas borrowed from nature, as well as basic postulates of universality, solidity, inherent to the self-organisation of natural systems. In this work we propose to choose the population algorithm of search by a shoal of fish (Fish School Search, FSS), which is characterized by simplicity of implementation, interpretability and high speed of convergence [18]. In this algorithm, the position of agent in an N-dimensional space of solutions is presented in the form of the numeric vector $P^{|N|}$ of length N, which corresponds to the vector of parameters that is optimized.

In the algorithm of FSS, a shoal of fish is the aggregation of agents of the population that move with approximately

the same speed and orientation, keeping approximately the same distance among them. The individual success of each fish in the process of finding a solution is characterized by its weight that plays the role of memory. Each iteration of the search performs two groups of operators – operators of feeding and operators of swimming.

The operator of feeding formalizes the success of research by the agents of those or other areas of "aquarium" and is in the calculation of the weight of the z-th agent, which is proportional to the normalized difference in the values of the fitness function in the next and current iterations

$$w_z[k+1] = w_z[k] + \frac{J(P_z[k+1]) - J(P_z[k])}{\max(J(P_z[k+1]), J(P_z[k]))}, \ z = \overline{1, Z},$$

where $P_z[k+1]$, $P_z[k]$ is the position of the z-th agent in the multi-dimensional space of solutions in the k-th and $(k+1)$-th iteration of the FSS algorithm.

Maximum possible value of weight of the agent w_z in the FSS algorithm is limited by the value $w_{max} > 0$. In this case, during initializing the population, all the agents are assigned the weight that equals $w_{max} \cdot 0,5$.

In the FSS algorithm they distinguish three types of swimming – individual, instinctive-collective and collective-willed. These types of swimming are performed sequentially one after the other at certain intervals

$$(t_1, t_2], (t_2, t_3], (t_3, t_4), \ \ t_1 < t_2 < t_3 < t_4, t_4 = t_1 + 1.$$

During individual swimming of the agents, their displacement occurs, which has equiprobable random character. In this case, in one iteration of the FSS algorithm, a step of individual swimming is performed a fixed number of times. The components of the displacement step are evenly distributed in the given interval v_{max}^{ind}

$$V_z^{ind} = U(0;1) v_{max}^{ind}, \ z = \overline{1, Z},$$

where $U(0;1)$ is a random number from the range $(0;1)$.

In the process of instinctive-collective swimming, each of the agents is affected by all other agents of the population and this effect is proportional to individual successes of the agents. In this case, positions of the agents are calculated by the formula

$$P_z^{t_3} = P_z^{t_2} + \frac{\sum_j V_j^{ind}(t_2)(J(P_j^{t_2}) - J(P_j^{t_1}))}{\sum_j J(P_j^{t_2}) - J(P_j^{t_1})}, \ j = \overline{1, Z}. \quad (12)$$

Collective-willed swimming comes down to shifting all the agents in the direction of the current center of gravity of the population under conditions of increase in the total weight of the shoal of fish as a result of individual and instinctive-collective swimming. If the total weight is decreased, then the shift moves in the opposite direction. Collective-willed swimming is performed according to the rules

$$P_z^{'} = P_z^{t_3} \pm v^{vol}(P_z^{t_3} - P_c^{t_3}), z = \overline{1, Z}, \quad (13)$$

where $P_c^{t_3}$ are the coordinates of the center of gravity of the shoal of fish that are defined by formula

$$P_c^{t_3} = \frac{\sum_z w_z^{t_3} P_z^{t_3}}{\sum_z w_z^{t_3}}.$$

In the formula (6), a plus sign is used under conditions

$$\sum_z w_z^{t_3} > \sum_z w_z^{t_3 - 1},$$

and a minus sign – in the opposite case. In this case, the dimension of the step of the agents' displacement v^{vol} is a random mangnitude

$$v^{vol} = v_{max}^{vol} U(0;1),$$

where v_{max}^{vol} is the positive value of the maximally permissible length of the step of displacement during collective-willed swimming.

Thus, we proposed the algorithm of functioning of the management system of a data center based on the reduction of multi-criteria optimization problem to the one-criterion and the search by the population algoritm of the shoal of fish for the global maximum of generalized criterion of efficiency of functioning of a data center. In this case, such partial criteria as the level of compliance with the terms of SLA are found by the forecasting, where prognostic decision rules are formed in the process of information-extreme machine learning by the training data of monitoring and are automatically corrected when changing the structure of the data center's physical resources consumption by virtual machines.

5. Results of physical modeling the intellectual management system of IT infrastructure of a data center

For physical modeling, we set up a data center based on 20 servers with processors of the Intel type (including Atom and Xeon), which differ by the number of cores, the volume of memory, clock frequency and energy consumption. Virtual machines are deployed on the servers, one part of which performs the functions of the virtual nodes of Hadoop cluster for processing heterogeneous tasks of distributed data processing using MapReduce such as PiEstimator (estimation of the number Pi with accuracy 1st million mark), WordCount (calculation of frequency of occurrence of words in 15GB of data); Sort (sorting 18GB of data), Grep (search for matches by a randomly selected regular expression in 6GB of data), TeraSort (sorting 1GB of data), Kmeans (cluster analysis of 6GB of numerical data). Another part of the virtual machines performs the functions of web servers, based on which the online services of access to books and applications of the Micro Web App or PHP/MySQL type operate. The work load on the servers is generated by clients that are programmed to form a total minimum demand in resources, which varies over time according to the law shown in Fig. 2.

As a tool to deploy a cloud-based platform of the data center we used free software Apache CloudStack written in Java [19]. The function of a hypervisor is performed by XenServer [19]. The data warehouse is controlled by using NetApp and Cumulus [19]. Apache CloudStack supports a variety of algorithms to manage the placement of virtual machines, for example, firstfit, random, worstfit and others, however, in this study we developed our own scheduler by way of inheriting from the class nova.scheduler.driver.

Scheduler and overriding methods with the purpose of research into the following methods schedule_run_instance and select_destinations.
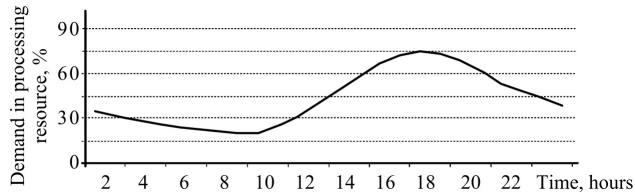

Fig. 2. Graph of load change of data center within 24 hours

The task of consolidating virtual machines of the data center is divided into four subtasks: detection of underloaded servers; detection of overloaded servers; selection of virtual machines for replacement; placement of virtual machines on servers. To simplify the comparison, initially we implemented the algorithm of placement of virtual machines MBFD (Modified Best Fit Decreasing), the main idea of which is to sort virtual machines that are subject to migration, in decreasing order of their resource demands and their assignment to the most energy efficient servers that have a sufficient amount of resources [20]. In this case, by default, the algorithm of detection of the underload finds the least loaded servers and tries to relocate its virtual machines by the MBFD algorithm to other servers. To examine the proposed algorithms, their efficiency will be compared to the MBFD algorithm.

In this study we propose to carry out optimization of distribution of virtual machines by way of population search by integral criteria (10) and (11). In this case, with the purpose of taking into account the heterogeneity of virtual machines and servers, it is proposed, before displacement, to predict the SLA terms violations as a result of competition of virtual machines for shared physical resources.

The training of management system to predict the SLA terms violations as a result of competition of virtual machines for resources of physical servers is carried out both for the level of IaaS (Infrastructure-as-a-Service) and for the level of SaaS (Software-as-a-Service). Therefore, for the clients' requests we attached the terms of the user's SLA in the form of the following indicators:

– maximum period, during which a user agrees to expect the result, Td;

– the price a user is ready to pay for the service, Bd;

– amount of compensation for violating the deadline, Cp;

– volume of files that are sent by users, Fs;

– length of the user's request (in millions of instructions that would be run on a virtual node), Mi.

Each physical node in an experimental environment was described by resource SLA in the form of metrics:

– deployment time of virtual machine, Tvm;

– the cost of one hour of using a virtual machine, Cvm;

– the cost of data transfer between users and a virtual machine, Cfs;

– the speed of processing the user's tasks in millions of instructions per second, M;

– the speed of data transfer between a user and a physical node, Sfs.

Due to the limited resources of the data center, predicting the SLA violations as a result of competition of virtual machines for shared physical resources makes it possible to make decisions that are optimal in terms of cost. In this case, the repetitive nature of tasks, which are solved by applications of virtual machines, provides for the possibility of using machine learning to analyze the log-data of tracing virtual machines for the synthesis of predictive model. The input mathematical description of such a model includes the results of cluster analysis of the data of tracing virtual machines by the k-means algorithm. The groups (classes) of virtual machines, formed in this way, characterize the pattern that exist in a data center of consumption of various types of resources of the physical server. In this case, the feature set of the classifier of virtual machines includes mean volume of the processors' resource usage, operating memory usage, swap file usage, network channel usage, disk space usage, mean intensity of input/output operations with the disk memory.

For training the prognostic classifier of the SLA terms violation due to placement of the selected virtual machine on the selected physical node, we propose to use the following feature set:

– normalized, relative to the requirements of the selected virtual machine, volumes of free resources (CPU, RAM, I/O Disk, Network) of the selected physical node;

– normalized numbers of virtual machines of each class (6 clusters) on the selected physical node taking into account the selected virtual machine, which is planned to be placed on it;

– estimated time of query execution $Te=Td-Tvm-Fs/Sfs-Mi/M$;

– estimated budget remainder $Br=Bd-Cfs \times Fs/Sfs-Cvm \times Mi/M$.

Determining the belonging of feature vector to one of the classes of the SLA violation with the aim of forming the learning sample is carried out by the results of registration in the monitoring data of such events as: the task is forwarded by the scheduler to the hosting machine for starting; re-planning (migration) is executed for the task; the task is fulfilled without violations of the SLA terms.

Fig. 3 displays a graph of dependency of averaged value of normalized information CFE (5) on the number of iterations of parameter optimization of the field of control tolerances on the attributes value (4) by the swarm algorithm for search.
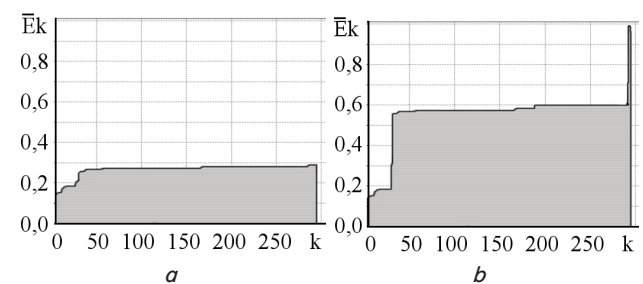

Fig. 3. Graph of change in maximums of value of the normalized criterion (5), averaged by set of classes, in the process of swarm optimization of the system of control tolerances: *a* — one-level system of control tolerances; *b* — multilevel system of control tolerances

The analysis of Fig. 3, *a* shows that information-extreme machine learning by traditional algorithm with a single basic class does not allow obtaining highly reliable decision rules and the corresponding value of the global maximum of averaged normalized CFE of training the classifier is $\overline{E}_*^* = 0,29$,

that corresponds to the value of accuracy $\overline{P_{true}}^{*} = 0,92$. The analysis of Fig. 3, *b* demonstrates that the construction of the system of control tolerances for each class of recognition $\{\delta_{i,m}^{*}, i = \overline{1,N}; m = \overline{1,M}\}$ allows receiving the boundary value of the information criteria $\overline{E} = 1,0$ and obtaining decision rules, containing no mistakes with regard to learning matrix, i. e., $\overline{P_{true}}^{*} = 1,0$.

Fig. 4 displays the charts of dependency of the normalized CFE (5) on the radii of containers of the corresponding classes at the optimal system of control tolerances to the values of feature set.
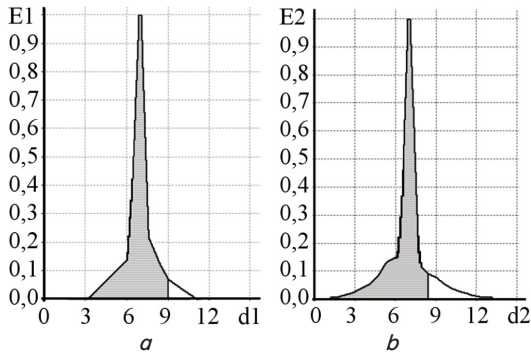

Fig. 4. Charts of dependency of the normalized CFE (5) on the radii of containers of classes $a — X_1^{\circ}$; $b — X_2^{\circ}$

The analysis of Fig. 4 shows that the maximum values of CFE of learning for the classes X_1° and X_2° are equal to $E_1^{*} = 1,0$ and $E_2^{*} = 1,0$, respectively, and the optimal values of radii of the corresponding containers of the classes of recognition $d_1^{*} = 7$, $d_2^{*} = 7$ (in code units). The CFE boundary values for the classes X_1° and X_2° testify to building decisive rules for them, containing no mistakes with regard to learning matrix. In this case, the inter-center code distance is $d(x_1 \otimes x_2) = 14$, indicating compactness of the feature vectors and the clarity of partition in the binary Hamming space.

Fig. 5 shows dependence of the operating expenses (in relative units), which include the total cost of migration of virtual machines, compensation for SLA violations and electricity costs, on every step of decision-making by the MBFD algorithm and its modifications with forecasting the SLA violations and optimization by criteria (10) and (11), respectively.
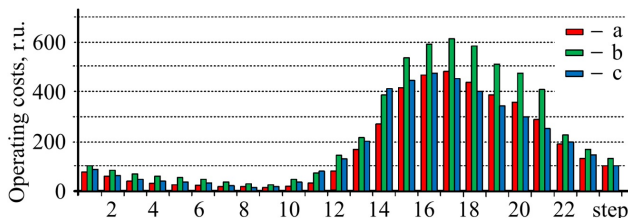

Fig. 5. Histogram of change in operating costs for different planning algorithms depending on the load of a data center on the corresponding step of decision making: *a* — MBFD; *b* — predictive modification of MBFD by criterion (10); *c* — predictive modification of MBFD by criterion (11)

The analysis of Fig. 5 reveals that the use of prognostic decision rules makes it possible to reduce operating expenses due to decreased number of violations of the SLA terms and reduction of idling physical resources without overloading the loaded physical nodes. In this case, when increasing the load of a data center, the convolution of criteria (10) is more optimal, and while decreasing the load of a data center, the best results can be obtained by using convolution (11).

Thus, the developed algorithm of information-extreme machine learning to predict the SLA violations allows increasing the efficiency of minimization of operating costs on resource management of a data center. In this case, one of the considered convolutions of the partial criteria of optimization is more efficient when increasing a data center load, and the other one – while decreasing the demands in physical resources of a data center.

---

6. Discussion of results of the simulation modeling

As seen in Fig. 2, 5, the rise in operating costs of the data center management system is proportional to the total resource needs of users in services; however, the use of population search with forecasting SLA violations allows reducing the level of expenditures in comparison with the MBFD algorithm. In this case, the convolution of criteria (10) proved to be more sensitive to the level of power consumption, which make it possible to increase the efficiency of planning at the stage of growth in the load of a data center, whereas convolution (11) is more sensitive to the level of idling physical resources, which predetermines its efficiency at decreasing the load of a data center.

The proposed modification of the algorithm of information-extreme machine learning by the sample that is formed in the process of monitoring the work of a data center, managed by the MBFD algorithm, in comparison with traditional algorithm, allows increasing the functional efficiency of the management system and obtaining decision rules, containing no mistakes with regard to learning matrix. However, as a result of expanding the set of services that are deployed based on virtual machines of a data center, new templates of consumption of physical resources emerge, which leads to the change in statistical characteristics of patterns and, accordingly, decrease in the functional efficiency of the decision rules formed previously. Fig. 6, *a* displays serial statistics, obtained at the optimal parameters of learning, by the first n=100 feature vectors (observations) of a learning matrix for a given alphabet of the classes. Fig. 6, *b* demonstrates the change in time of examination serial statistics in their variational blocks during n recognition with the growth in demand for new services that were not used when collecting the learning sample. In this case, the curves of EOS of red color match statistics for the class X_1°, and the curve of dark blue color matches statistics for the class X_2°.

The analysis of Fig. 6, *a* shows that the class X_1°, that describes fulfillment of the SLA conditions, has the smallest value of EOS, equal to $S_1^{*} = 98,5$. The class X_2°, that describes functional condition of the SLA terms breach, is more appropriately characterized by the value of EOS equal to $S_2^{*} = 131$. The analysis of Fig. 6*b* reveals that after the change in the structure of consumption of resources within 72 recognitions of the functional state, the one–dimensional static characteristic S_1 of implementations of the class X_1° transfers to the variation block of the class X_2°, that creates statistical uncertainty. In order to prevent reduction in func-

tional efficiency of the management system, at this point one should initiate its retraining.
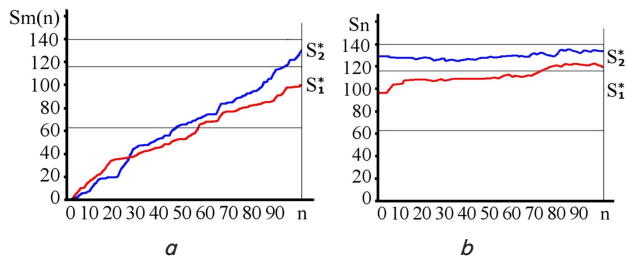

Fig. 6. Graphs of dependence of EOS: *a* – on the number of trials at the optimal parameters for learning;
b – on the number of periods of examination in the process of growing demand for new services

Thus, the developed information and algorithmic provision of intellectual system of data center mangement allows determining the moment of reduction in its functional efficiency and adapting to the new conditions of operation.

7. Conclusions

1. The advantage of applying in the information-extreme learning of the proposed algorithm of binary encoding of feature set was demonstrated, in which, in contrast to the traditional algorithm, the system of control tolerances is defined for each class of recognition, not only for the base one, which makes it possible to enlarge the code distance between the centers of containers of the classes and to increase the accuracy of decision rules.

2. The results of physical simulation proved efficiency of the proposed algorithm for planning the placements of virtual machines on physical servers using integral optimization criterion derived with the help of convolutions of the partial criteria. It was observed that the additive-multiplicative convolution is more sensitive to the level of power consumption, which enables to increase the efficiency of planning at the stage of growth in the load of a data center, while the entropic convolution is more sensitive to the level of idling physical resources, which predetermines its larger efficiency at the reduction in the load of a data center.

3. It was demonstrated that the use of extreme serial statistics in the form of normalized statistics of the numbers of attributes' entry into their fields of control tolerances allows, in operating mode of the management systems, detecting occurence in a datacenter of new patterns of consumption of physical resources due to active use of new services of a data center, allowing timely training or retraining of the management system and maintaining its functional efficiency at high level.

References

1. Cao, Z. Dynamic VM consolidation for energy-aware and SLA violation reduction in cloud computing [Text] / Z. Cao, S. Dong // Proceedings of the 13th International Conference on Parallel and Distributed Computing, Applications and Technologies, 2012. – P. 363–369. doi: 10.1109/pdcat.2012.68

2. Sharma, B. Applications of Data Mining in the Management of Performance and Power in Data Centers [Text] / B. Sharma // Technical Report, Department of Computer Science and Engineering. The Pennsylvania State University, 2009. – P. 1–5.

3. Caglar, F. Towards a performance interference-aware virtual machine placement strategy for supporting soft real-time applications in the cloud [Text] / F. Caglar, S. Shekhar, A. Gokhale // Proceedings of the 3rd International Workshop on Real-time and Distributed Computing in Emerging Applications, 2014. – P. 15–20.

4. Delimitrou, C. Paragon: QoS-aware scheduling for heterogeneous datacenters [Text] / C. Delimitrou, C. Kozyrakis // Proceedings of the 18th international conference on Architectural support for programming languages and operating systems. – 2013. – Vol. 41. – P. 77–88. doi: 10.1145/2451116.2451125

5. Hayashi, T. Performance Degradation Detection of Virtual Machines via Passive Measurement and Machine Learning [Text] / T. Hayashi, S. Ohta // International Journal of Adaptive, Resilient and Autonomic Systems. – 2014. – Vol. 5, Issue 2. – P. 40–56. doi: 10.4018/ijaras.2014040103

6. Bodik, P. Fingerprinting the Datacenter: Automated Classification of Performance Crises [Text] / P. Bodık, M. Goldszmidt, A. Fox, D. B. Woodard, H. Andersen // Proceedings of the 5th European conference on Computer systems, 2010. – P. 111–124. doi: 10.1145/1755913.1755926

7. Nanduri, R. Job Aware Scheduling Algorithm for MapReduce Framework [Text] / R. Nanduri, N. Maheshwari, R. Raja, V. Varma // 2011 IEEE Third International Conference on Cloud Computing Technology and Science, 2011. – P. 724–729. doi: 10.1109/cloudcom.2011.112

8. Kandalintsev, A. Profiling Cloud Applications with Hardware Performance Counters [Text] / A. Kandalintsev, R. L. Cigno, D. Kliazovich, P. Bouvry // The International Conference on Information Networking, 2014. – P. 52–57. doi: 10.1109/icoin.2014.6799664

9. Dovbysh, A. S. Information-Extreme Method for Classification of Observations with Categorical Attributes [Text] / A. S. Dovbysh, V. V. Moskalenko, A. S. Rizhova // Cybernetics and Systems Analysis. – 2016. – Vol. 52, Issue 2. – P. 224–231. doi: 10.1007/s10559-016-9818-1

10. Dovbysh, A. S. Learning decision making support system for control of nonstationary technological process [Text] / A. S. Dovbysh, V. V. Moskalenko, A. S. Rizhova // Journal of automation and information sciences. – 2016. – Vol. 48, Issue 6. – P. 39–48. doi: 10.1615/jautomatinfscien.v48.i6.40

11. Chen, L. MTAD: A Multitarget Heuristic Algorithm for Virtual Machine Placement [Text] / L. Chen, J. Zhang, L. Cai, R. Li, T. He, T. Meng // International Journal of Distributed Sensor Networks. – 2015. – Vol. 2015. – P. 1–14. doi: 10.1155/2015/679170

12. Salmasnia, A. A new desirability function-based method for correlated multiple response optimization [Text] / A. Salmasnia, M. Bashiri // The International Journal of Advanced Manufacturing Technology. – 2015. – Vol. 76, Issue 5-8. – P. 1047–1062. doi: 10.1007/s00170-014-6265-x

13. Altinoz, O. T. A multiobjective optimization approach via systematical modification of the desirability function shapes [Text] / O. T. Altinoz, A. E. Yilmaz, G. Ciuprina // 8th International symposium on advanced topics in electrical engineering, 2013. – P. 1–6. doi: 10.1109/atee.2013.6563481

14. Sanginova, O. Comparative analysis of some computional schemes for obtaining a compromise solution [Text] / O. Sanginova // Eastern-European Journal of Enterprise Technologies. – 2015. – Vol. 1, Issue 4 (73). – P. 10–18. doi: 10.15587/1729-4061.2015.35607

15. Shengnan, Z. Multi-response robust design based on improved desirability function [Text] / Z. Shengnan, W. Jianjun // International Conference on Grey Systems and Intelligent Services, 2015. – P. 515–520. doi: 10.1109/gsis.2015.7301911

16. Kushwaha, S. A Modified Desirability Function Approach for Mean-Variance Optimization of Multiple Responses [Text] / S. Kushwaha, S. Sikdar, I. Mukherjee, P. K. Ray // International Journal of Software Science and Computational Intelligence. – 2013. – Vol. 5, Issue 3. – P. 7–21. doi: 10.4018/ijssci.2013070101

17. Yoo, D. G. Rehabilitation Priority Determination of Water Pipes Based on Hydraulic Importance [Text] / D. G. Yoo, D. Kang, H. Jun, J. H. Kim // Water. – 2014. – Vol. 6, Issue 12. – P. 3864–3887. doi: 10.3390/w6123864

18. Parpinelli, R. Theory and New Applications of Swarm Intelligence [Text] / R. Parpinelli. – InTech, 2012. – 204 p. doi: 10.5772/1405

19. Jain, S. A Comparative Study for Cloud Computing Platform on Open Source Software [Text] / S. Jain, R. Kumar, Anamika, S. K. Jangir // An International Journal of Engineering & Technology. – 2014. – Vol. 1, Issue 2. – P. 28–34.

20. Kaur, A. Energy optimized VM placement in cloud environment [Text] / A. Kaur, M. Kalra // 2016 6th International Conference – Cloud System and Big Data Engineering (Confluence), 2016. – P. 141–145. doi: 10.1109/confluence.2016.7508103