Cloud Computing Support for Diagnosing Researches

A. Amirov, O. Gerget, V. Kochegurov

Abstract**—**One of the main biomedical problem lies in detecting dependencies in semi structured data. Solution includes biomedical portal and algorithms (integral rating health criteria, multidimensional data visualization methods). Biomedical portal allows to process diagnostic and research data in parallel mode using Microsoft System Center 2012, Windows HPC Server cloud technologies. Service does not allow user to see internal calculations instead it provides practical interface. When data is sent for processing user may track status of task and will achieve results as soon as computation is completed. Service includes own algorithms and allows diagnosing and predicating medical cases. Approved methods are based on complex system entropy methods, algorithms for determining the energy patterns of development and trajectory models of biological systems and logical–probabilistic approach with the blurring of images.

Keywords**—**Biomedical portal, cloud computing, diagnostic and prognostic research, mathematical data analysis.

I. Introduction

MODERN biology and medicine rush away from verbal description towards formalized processes, mathematical models and information technologies. Significant difficulties in researching quantitative biomedical systems are based on their features, including structural and functional complexity, variability of expository indexes, nonlinear curves, and fuzziness of research objects. Diagnostic and prognostic tasks cannot be solved without appropriate informative environment being created. Its creation enables to settle problems of data and knowledge representation, seeking dependencies, creating decision rules, decision making. One the most intensive bioinformative environment designing process is seeking dependencies in data arrays. It is not always successful due to database containing various types of contradictory and incomplete information. Each scale has its own way of mathematical processing. There are procedures for cases when features are measured in different scales. This field of science is constantly evolving, new data is rapidly streaming, and new algorithms are developed to obtain information. But most of the existing biomedical portals are designed for solving limited amount of tasks, these portals are complex, expensive and narrow what makes them inapplicable for common use in hospitals. Therefore, for diagnostic and prognostic studies in polytypic information processing the medical portal was developed that includes such important services as search for hidden patterns, estimation of functional reserves, adaptive capacity, human body tension degree, as well as diagnosis and prognosis of disease. When service portal based on the original algorithm was created, it was taken into the fact that the human body is a complex dynamic biological system, which contains all the elements necessary for sustainable operation in ambient conditions. In order to exist, it needs to communicate with the outside by information, energy and matter exchange. Metabolic processes in the body are subject to the fundamental laws of science and its operation can be considered as a change in the internal state of the internal and external forces.

II. Biomedical Portal Structure

Portal advances diagnostic and prognostic researches with the help of portal services and parallel computing.

1. Service determines patterns of development in semi-structured data on the basis of energy data. Changing of the human state is based on energy sharing processes occurring within and supported by the arrival of external energy. Degradation of dynamic systems is associated with the impaired functioning of energy-exchange processes, supporting changing states within limits. This means that the generic assessment criteria of dynamic bio-system functioning are to be based on energy indexes.

In order to apply energy indexes, a system-wide formalism description was used. To this end, we introduced generalized indicators:

$$q(x,t), \dot{q}(x,t) \qquad (1)$$

In generalized coordinates, the state of the system is determined by the level of kinetic energy$W_k(q,\dot{q})$and potential energy$W_п(q,\dot{q})$.In this case, the partial derivatives are, respectively, the generalized momentum and generalized force, determining the temporal processes in a dynamic system:

$$\frac{\partial W_k(q,\dot{q})}{\partial \dot{q}} = P(q,\dot{q}) \qquad (2)$$

$$\frac{\partial W_п(q,\dot{q})}{\partial q} = F(q,\dot{q}) \qquad (3)$$

Metabolic processes are accompanied by the consumption of kinetic energy and to perform or to recover potential energy. When considering the difference of kinetic and potential energy, it is possible to take into account the internal and external costs of energy supply. Metabolism contains slowly varying components (that characterize pattern) and relatively rapidly changing cyclical components. Based on the study, to assess the functioning of the human body it is important to control not only the levels of state but their

A. Amirov is with the Karaganda State Technical University, Karaganda, Bulvarmira avenue, 56, Kazakhstan (phone: 7212-424972; e-mail: azamat-amirov@mail.ru).

O. Gerget and V. Kochegurov are with the National Research Tomsk Polytechnic University, Tomsk, Lenin Avenue, 30, Russia (phone: 3822-426-100; e-mail: olgagerget@mail.ru, kva06@mail.ru).

dynamic relationship. Different observations enable different processing methods.

2. The service allocation of main paths in the normal functioning of the body. The development of dynamic processes follows a trajectory (magistral) providing a balanced equilibrium state with variables that change over time. To monitor the magistral state (evaluating system properties related to the regular changes of bio-object state variables) it is appropriate to use a geometric mean ratio [1]:

$$\Gamma_x(t) = \sqrt[n]{\prod_{i=1}^{n} x_i}\,(t) \qquad (4)$$

For the equilibrium state (functioning within the normal range) geometric mean value must be equal to:

$$\Gamma_{x_0}(t) = \sqrt[n]{\prod_{i=1}^{n} x_{i_0}}\,(t) \qquad (5)$$

A geometric mean relative deviation from the equilibrium values of biological object is defined as:

$$\frac{\Delta\Gamma_x(t)}{\Gamma_{x_0}(t)} = \frac{1}{n} * \sum_{i=1}^{n} \frac{\Delta x_i(t)}{x_{i_0}(t)} \qquad (6)$$

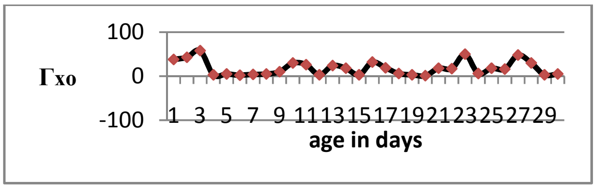
Under certain tolerances the stress state of the system can be determined on the basis of geometric mean indicators for relative change:

$$\gamma = \frac{\alpha^2}{1-\alpha^2} \qquad (7)$$

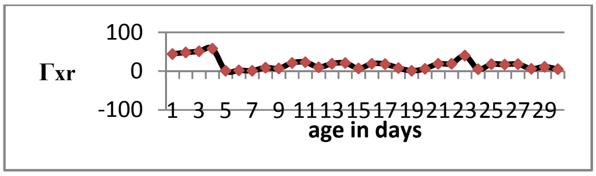
$$\alpha^2 = \frac{\Delta x^{\mathrm{T}}(t)\Delta x(t)}{\Delta x_m^{\mathrm{T}}\Delta x_m} \qquad (8)$$

where $0 \leq \alpha \leq 1$ and $0 \leq \gamma \leq \infty$.

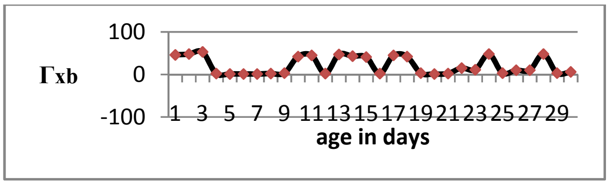
We present the results of blood indices processing that aim to identify the link between the child's body and living conditions. For the study, experts formed three groups of children: group of healthy children, children at risk for variations in health status, and sick children (diagnosed with damage of perinatal central nervous system). During the research, average values for the following teams were simulated. There are trajectory model laws for different groups of children in Fig. 1



(a)



(b)



(c)

Fig. 1 (a) Healthy children; (b) children with the risk of variations in health status; (c) sick children

In this case, the approach proposed in [2] and [3] was in use, in particular, considering the information as a measure of biological object preference behavior:

$$I = \sum_{j=1}^{n} P_0(x_j) ln \frac{P_0(x_j)}{P_1(x_j)} \qquad (9)$$

where *n* is the number of informative features; *$P_0(x_j)$* is the probability determining the bio–object state i.e. the case of *j* variable deviations from the "norm" being 0. The state of the biological entity will be treated as the human condition. In assessing the adaptability of biological objects, a condition of all variables being the average values will be treated as the normal condition. *$P_1(x_j)$*is the probability that feature value x corresponds to what is "normal." The probability is calculated by the following formula:

$$P_1(x_j) = P(|x_j - \bar{x}| < \delta) = 2\Phi\left(\frac{\delta}{\sigma}\right) - 1 \qquad (10)$$

where x_j is a mean value of $\bar{x}$ feature, δ is the value of a given deviation, σ is standard deviation of the trait, and Φ is Laplace function.

This criterion provides a deviation measure of a person’s current state from the "preferred". Thus, we have an integrated assessment of the adaptive capabilities of an organism which makes it possible to identify patterns in complex processes due to the influence of external factors on the organism’s functional state. If measured parameters vary randomly without any noticeable dependence then the system does not change its state and information indicators do not exceed the specified level. If the effect of the environment and conditions of the state of the body leads to a change in the state information, indicators will exceed the performance baseline depending on the change in the state of biological systems.

3. Service evaluations of functional reserves and the degree of tension of the human body based on the entropy approach.

In contrast to classical physics that considers deterministic and reversible processes in the human body, and any open system as well, there may be stable imbalance. Entropy of living systems being measures of uncertainty enables a comparison of their order, disorder, and uncertainty for different states of the human body. With equal probabilities of all possible states a system is completely disorganized because it can change its state at any moment. Such systems possess maximum energy. Increasing order (decreasing in entropy) means an increasing relationship between factors determining the system's state that leads to the predictability of the system's behavior.

4. To carry out diagnostic and prognostic studies the portal service provides construction of a neural network. A neural network is a set of neurons that are interconnected in some way. Neurons and interneuron communication is set programmatically. The computing of the neuron resembles the work of the biological neuron. The functioning of the formal neuron is as follows. At the current time a neuron receives input signals from other neurons through dendrites. The signal from each input is multiplied by a weighting factor of the input and added to other signals multiplied by a weighting factor corresponding inputs as well. Depending on the measured value, an output signal is calculated and transmitted to other neurons:

$$S = \sum_{i=1}^{n} x_i w_i \tag{11}$$

where n – number of inputs of the neuron; x_i– the value of i_{th} input neurons; w_i – weight of the i_{th} synapse. Thus, the neural network takes some signal as its input; after it passes neurons the network becomes available to output the definite answer that depends on the weights of all neurons. Network learning is the process of finding the neural weights that minimize the error [4]. Choosing the correct number of neurons in the hidden layers is very important. A small number of neurons may lead to problems with learning, while too much will make the learning process too time–consuming. The number of neurons in the hidden layer was calculated:

$$N = \frac{1}{2}(N_{\text{вх}} + N_{\text{вых}}) + \sqrt{Q} \tag{12}$$

where $N_{вх}$, $N_{вых}$ — dimension of the input and output signals respectively; Q is the number of elements in the training set.

The number of neurons in the input layer is determined by the number of input factors, the output layer – the number of output factors.

Sigmoid function was the same as the activation function:

$$f(x) = \left(\frac{1}{(1+exp(\alpha x))}\right) \tag{13}$$

where α– is the slope parameter of the sigmoid function. Multilayer neural networks possess greater abilities than single–layer networks only if the activation function is non–linear. There are limitations for the back–propagation method: function must be differentiable everywhere. Sigmoid function satisfies this requirement. It is worth mentioning that it automatically controls the amplifying. For weak signals (f(x) close to zero) the "input – output" curve has a significant slope that causes strong amplification. When the signal becomes larger the gain decreases. Thus, large signals are perceived by networks without saturation, and weak signals pass through the network without excessive attenuation.

Testing the service portal was carried out for the diagnosis of perinatal damage to the central nervous system in children. The neural network used normalized data and blood parameters neurosonography as the input data that was presented as a 16x330 matrix. The amount of rows determines the neural network input layer, and the amount of columns determines tge set of research objects. Accuracy is 94%.

5. Diagnostic service with incomplete and vague information.

The concept of complete and incomplete information exists because it is not always possible to measure the object's whole set of attributes. If determining the values of features is available only with some degree of confidence, or in a range of features, then the information may be treated as vague. In such situations, it is advisable to use the probabilistic methods. The concept of logical and probabilistic decision–making is not fresh and was used in the evaluation of the probabilistic characteristics of complex systems [5], reliability studies in engineering systems [6], and technical diagnosis and assessment of the probability of events based on belief networks [7]. Although these problems were solved by logic and probabilistic approaches, none of them considered situations when decision–making was carried out by assigning values to variables in a certain range or the equally probable assignment of different images (diseases).

The portal service is based on probabilistic decision–making algorithms [8]. The matrix method mentioned in [9] was proposed as a basic method for data and knowledge construction. In order to represent the training set of objects that the Q matrix was used, each row of the matrix submitted a single object from the set, and columns corresponded to features. Training objects' accessory to a pattern was set by an R distinctions matrix. R matrix rows correlate to Q matrix rows and R columns to the classification criteria. Objects that match up to the same combination of classification criteria are considered to be in the same class, and multiple descriptions of these objects determine the class. Description of an object submitted by a set of criteria is set by conjunction z, and what is more each variable may be set with some degree of confidence (probability) or interval. Testing of probabilistic algorithm conducted for assess the functional state of the human body. Rows of Q collated to objects descriptions (a set of attribute values obtained from experimental and laboratory studies conducted on humans) and columns – diagnostic features. A matrix specifying the partition of the training objects on non–overlapping disorders (classes) was used as an R matrix. A description of each class included 30 to 35 rows of the matrix Q in the space of 26 characteristic features.

The state of the human body, described by a set of characteristic features, was set by the conjunction of z and the range of attribute values using probability values. In incomplete information, reliability of the results obtained using this service was 87% and confirmed the highly qualified status of the experts.

6. In conditions of incomplete and inaccurate information diagnostic problems were solved with an algorithm of heterogeneously consistent recognition procedures [10]. Inhomogeneous consecutive recognition procedures were evident in the designation of the object to one of two non–overlapping classes. The object of study is set by characteristic features $x_1,\ldots,x_n$. In order to increase decision–making and avoid error these features were arranged in descending order of information provided. To determine each feature's information measure, Kullback–Leibler divergence was used:

$$lg\frac{\alpha}{1-\beta} < k(x_1) + k(x_2) + \cdots + k(x_n) < lg\frac{1-\alpha}{\beta} \qquad (14)$$

$$k(x_i) = lg\frac{P(x_i/A_1)}{P(x_i/A_2)} \qquad (15)$$

To find whether the object belongs to one of the classes available the formula was used: where$k(x_i)$ – diagnostic factor for each feature. P_1 – possibility of belonging to A_1 class, P_2 – possibility of belonging to A_2 class.$P(x_i/A_j)$ - conditional probability ($j\in\{1,2\}$).

An inhomogeneous consistent recognition procedure aimed at solving the problem of diagnosing the most common neonatal period children illnesses. For this purpose, two groups of children (training and control) were formed and surveyed. The training group received coefficients for each of the states – "sick", and "healthy". The quality of recognizing with diagnostic tables was checked using the control group of children. As the scope of the report does not allow presenting the actual results of performed diseases' diagnosis, only a summary of the analysis will be mentioned. Fault diagnosis: neural system damage, anemia, pneumonia, meningitis, lies within 3.2–6.4%. A number of indeterminate findings were observed in 12–14% of cases, which is satisfactory in practice. The developed bioinformatics medical portal allows the processing of diagnostic and scientific medical data in a parallel mode using a private cloud platform and such software as Microsoft System Center 2012 and Windows HPC Server. The implementation of such a service allows administrators to hide internal processing logic from the user by providing a user–friendly interface. Sending data to the calculation allows users to monitor the status of the task and get them to the end of the treatment process. The parallel data processing mode provides a high load of computational resources by distributing complex tasks into multiple computing units and cloud computing is moving towards running multiple tasks on a single server as a virtual machine. Virtualization is achieved by separating logical and physical levels, which solves a lot of problems. Programs and data are permanently stored in a remote data center and access to a computing service is transparent for the user via the Internet and use of a personal computer. A private cloud is a concept that provides multiple advantages for bioinformatics and biomedicine. Computing with cloud technology enables engagement with a necessary amount of hardware and software recourses at any moment of processing. Thus we obtain the incessancy for performing long-term settlements that suppose an unpredictable burst or a cyclical load. Transferring the load from a limited-recourses platform to a virtual cloud enables us to acquire unlimited computing power and allows developers to get rid of the routine to install, update, support, and adjust. Such a model is named Science-as-a-Service. Thus, parallel processes based on the private cloud include all the benefits of parallel and cloud computing complementing each other.

The portal service includes original design and allows successful diagnosis and prediction of diseases. The complexity of tasks to be solved causes a portal service to be organized as a hierarchical structure, allowing increase of computing powers and modification of service architecture by adding or removing components. One of the key portal features enables every service to have two function modes: autonomous (single application for solving concrete tasks) or part of an IDE (packaged diagnosis). When developing bio–informatics, portal cloud technologies were used due to the fact that there are no clouds available that will allow the analysis of data for medical researches. Developing a cloud system of medical data processing allows a significant amount of scientists to get access to mathematical models and methods without being forced to start developing their own computational platform and software.

Having modern operating systems, which enables the creation of unique cloud solutions and widespread internet implementation, allows a remote cloud service to be a powerful tool for doctors that will assist them in solving tasks without additional software being required.

The application of neural network technology in conjunction with the use of parallelized computing and combined with accomplished experience of diagnosing various diseases enables the creation of a diagnostic platform with free access via the Internet.

When developing a medical portal, the private cloud concept was approved due to having a high degree of control over the performance, reliability, and security. Nowadays there are several implementations of the private cloud concept combined with parallel computing (http://www.activeeon.com) but they are not aimed at analyzing medical and confidential data thus being general purpose systems. The development of a medical portal will implement a comprehensive approach to the diagnosis and prediction of human body health, by combining the analysis and control of information and the organization of operational data into a single information space.

III. Conclusion

Biomedical portal realization for diagnostic and prognostic research is based on concept of cloud technology and parallel computing.

Authors proposed and tested entropy methods of modeling complex systems, energy algorithms for determining the energy patterns of development, approaches for building models of biological systems development trajectory.

Portal includes such important services as hidden dependency finding, assessing adaptive capabilities, degree of tension and morbidity forecast. Approbation of the portal using non simulated data displayed that quality of solution satisfy practitioner demands.

Further improvements will include designing proper interface and adding new services.

IV. Conclusion

Biomedical portal realization for diagnostic and prognostic research is based on concept of cloud technology and parallel computing.

Authors proposed and tested entropy methods of modeling complex systems, energy algorithms for determining the energy patterns of development, approaches for building models of biological systems development trajectory.

Portal includes such important services as hidden dependency finding, assessing adaptive capabilities, degree of tension and morbidity forecast. Approbation of the portal using non simulated data displayed that quality of solution satisfy practitioner demands.

Further improvements will include designing proper interface and adding new services.

References

[1] Konstantinova L.I. Kochegurov V.A.,Shumilov B.M. Parametric identifying non–linear differential equations based on spline schemeand polynomial// 1997.– №5.– P. 15–20.

[2] Rotov A.V., Pekker, I.S., Medvedev M.A., Berestneva O.G. Adaptive characteristics of a human (assess and prognose). –Tomsk:Pub. TPU, 1997.

[3] Gerget O.M., Kochegurov V.A. Actual medical provlems solving using mathmatical methods/ LAP LAMBERT Academic Publishing GmbH & Co. KG, LAP, 2012, 1 – 145.

[4] Neural networks advantages // Artificial intelligence portal. URL: http://www.aiportal.ru/articles/neural–networks/advantages.html.

[5] Zagoruiko N.G., Samokhvalov K.F., Sviridenko D.I. Empiric research logic. Novosibirk: Nayka, 1985.

[6] Gelfand I.M., Rosenfeld B.I., Shifrin M.A. Essays about working together mathematicians and physicians.–N.:Nayka,1989.

[7] Yankovskaya A.E., Matrosova A. Yu., Strizhov M.A. The Logical Probabilistic System of Pettern Recognition// Proceedings of the Pettern Recognition and Image Understanding . 5th Open German- Russian Workshop.– Germany. Herrshing.–1999.– pp. 298-305.

[8] Yankovskaya A.E., Gerget O.M. Intelligent subsystem logic and probabilistic recognition of blurring of images. Informational systems and technologies.–Novosibirsk,2000.–P. 548–551.

[9] Yankovskaya A.E. The degree of implication and partial orthogonalization disjunctive normal Boolean functions in connection with the problem of decision–making // All–Siberian Readings on Mathematics and Mechanics, 1997.– T.1.– Tomsk: Pub. TGU, 1997.– P.225–231.

[10] Genkin A.V., Dubner P.N., Petergao E.V. Subsystem forecasting indicators of child health in medical information systems / ACS designing problems.– Minsk, 1979.– № 37/3.– P.123–125.