Architecture of the Platform for Big Data Preprocessing and Processing in Medical Sector

Nataliya Shakhovska ^[0000-0002-6875-8534], Solomiia Fedushko ^[0000-0001-7548-5856].

Eugen Zasoba, Vladyslav Mykhaikyshyn, Mykhailo Osypov, Olena Vovk ^[0000-0001-5523-0901]

Lviv Polytechnic National University, Lviv, Ukraine

nataliya.b.shakhovska@lpnu.ua, solomiia.s.fedushko@lpnu.ua

Abstract. The paper presents the architecture of the platform for Big data preprocessing and processing. The method for prevention of the risk disease based on Probabilistic Production Dependencies is developed. The investigation is started from the platform for Big data in medical domain analysis. The platform consists of 6 layers: Data layer, Communication layers, Preprocessing layer, Data processing layer, User layer, Integrational layer. The platform architecture for Big data processing in medical sector is developed. The accuracy of the proposed method is estimated.

Keywords: EHealth, Probabilistic Production Dependencies, Big data, Data Mining, Medical Sector, Big Data Preprocessing, Big Data Processing

1 Introdution

In recent years, the health policies are transforming form acoute intervention driven towards prevention and self-responsibility of the individuals by implementing measures for increasing health literacy. Particularly important for the implementation of the prevention programs are the literacy of young medical staff, the completeness and timeliness of receiving information about patients, the ability to observe the patient, not only in the hospital, but also at home. That is why integration of information technologies, intelligent devices and systems into all spheres of life takes place. It is important that the technologies used by the patient be created taking into account the requirements of the universal design, that is, without the need for adaptation or design changes if used by a person with special needs. The engineering branch is engaged in the design of "smart" homes, businesses and cities. Such complex intelligent systems collect process and transmit large amounts of data during their operation, including in the medical sector. In this process, microcontrollers, communication equipment, sensors, as well as ambulances, the next branch of a health facility, family doctors, etc.,

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). IREHI-2019: International Conference on rural and elderly health Informatics, Dakar, Sénégal, December 04-06, 2019 process the information coming from patients. Devices may have failures, as well as errors of the first and second kind, which leads to incomplete or inaccurate information about the patient, and, consequently, adversely affects the adoption of medical decisions.

The aim of the paper is analysis of the existing stage of data mining technics and Big data approach in EHealth and development of the novel technic for prevention of the risk disease based on Probabilistic Production Dependencies.

2 State of art

Data mining is widely used in medicine. Norman et al [1] is proposed to analyze data gaps and fill them based on decision trees, EM algorithm, and regression approach to predict missing data using prediction functions. Similar results were obtained by Khanmohammadi [2] for associative classification of medical data and Constantinou [3] for comprehensive questioning and interviewing of data on Bayesian models to support medical decision making. The Bayesian network was also used to reformulate the QMR model in decision theory. However, with the spread of Big data technology, Bayesian networks were not fast enough. Therefore, Tang et al [4] developed a method for parallelizing Bayesian networks. Similar results were obtained by Anders L.Madse [5]. Bayesian networks are also used to diagnose diseases, such as Lakho and Seixas [6]. However, even in the case of parallelism, it is advisable to use Bayesian networks in combination with other methods of machine learning for multiparameter, large-scale and dynamic medical data flows [7].

Currently, Hybrid Systems of Computational Intelligence are widely used to solve a wide class of Medical Data Mining problems, combining the approximating and extrapolating properties of artificial neural networks (ANNs), both traditional shallow ones, and more advanced deep neural networks (DNNs), interpretability of the results of fuzzy reasoning systems (FRS), the ability to find the best architectures for solving specific problems, provided by using the apparatus of evolving systems [8]. Medical Data Mining considers the most complex tasks of classical Data Mining due to the fact that the generated clusters are usually nonconvex, overlapped, nonstationary; initial data are corrupted by "gaps" - omissions and abnormal outliers, the initial samples to be processed can be short, which leads to undesirable overfitting, as well as extra-long (Big Data), which leads to the inability to work in batch mode, and requires the use of Data Stream Mining methods. It is clear that in this situation, recurrent online data processing methods come to the forefront - training, which allows not to accumulate huge amounts of information, but to "forget" the data after it has been processed.

In modified applications, the most problematic are the problems of diagnostics and early fault detection of controlling biomedical signals, in doing so comes to the fore not only the accuracy of the results obtained but also the time required to obtain them come to the fore, which leads to the need to abandon traditional mutiepoch learning.

Currently, to solve the mentioned problems, the most effective from the accuracy of the obtained results are deep neural networks, which are however completely ineffective in the conditions of short learning sets and unsuitable for operating in online mode, when data are fed to processing in real time.

Probabilistic neural networks (PNN) introduced by D. Specht [9] and trained with the so-called lazy learning according to principle "neurons in data points" can serve as an alternative to DNN in diagnostic-recognition tasks. PNN training is very fast ("just-in-time-models"), however, these networks suffer from "curse of dimensionality", and besides, these crisp systems are inefficient in the context of overlapping classes. In connection with this, it is advisable to develop the evolutionary PNN architecture, which allows limiting the network dimension to the conditions of the data flow arriving for processing.

Bhatt [10] developed an approach to collecting medical data using the Internet of Things and a neural network ensemble to detect unusual human movements. However, in the case of solving highly specialized problems, which is characteristic of the field of medicine, the error of training of the neural network ensemble is much higher than the error of the operation of one network. In addition, the training of the neural network ensemble is quite time-consuming and costly. In Silva-Ramírez E. L [11], a multilayer perceptron method is used to solve the classification and fill in data gaps, and is based on a combination of multilayer perceptrons and k-nearest neighbors.

Therefore, data mining techniques are used to solve many of the problems of processing and analyzing medical information. However, there are no comprehensive studies aimed at identifying the patient's condition without specifying the type of anamnesis. Specific tasks have been addressed in this area, but studies have only partially addressed the phenomena of Big Data and the Internet of Things, in-depth analysis, and visualization of accumulated data to support personalized treatment decisions.

Big Data has attracted much attention in academia and industry fields in the last few years. The trend of utilizing medical big data has increased tremendously as well. The medical big data is generated from medical records, medicine, research, etc. but it is not well connected due to its long-term development in the last decades in most hospitals without high-level strategic guide and plan [12]. They propose a platform called Data Mid-Platform (DMP) to solve the above problems. The feasibility of this mid-platform is highly recommended and certified by several domain experts and it is under construction in a real project.

In the context of "Big Data +", people began to study the application of data visualization to medical data. Data visualization can make full use of the human sensory vision system to guide users through data analysis and present information hidden behind the data in an intuitive and easy-to-use manner. There is the problem of visualizing such a large amount of data in a way that is understandable to humans: doctors, patients. In addition, proper visualization allows you to extract general knowledge and track trends and dependencies. An example of a platform for the visualization of large medical data sets is presented in [13].

Healthcare (Cardiovascular diseases) has now become a major concern in the economic as well social regiment of almost everyone. The medical system should be able to handle long-term and continuous monitoring, inconspicuous monitoring and should be highly sensitive to the patient's ECG signal. IoT helps the ECG signals to get transmitted through the sensor via a gateway using communication protocols like Bluetooth, Zigbee, 3G/4G, Wi-Fi, LAN, etc. Then this data can be sent to numerous places like the doctor's end, caregiver or the cloud-storage for analysis or processing, etc. The doctor at the remote location can view the patient's report on various smart devices, all thanks to the computing technology. This helps in dealing with the emergencies. Big Data plays a pivotal role in this system as it provides with data analysis, decision-making, extracting useful information via algorithms, intelligent storage etc [14. 15]. The combination of these three technologies will make the world a more secure place to live. Also medical imaging is often used and gives big data medical records [16]. When we are able to analyze all these diagnostic data, it becomes possible to predict the patient's condition based on previous medical history [17, 18]. Table 1 presents comparison of Big data representation.

Table 1. Comparison of models of representation of Big data	

Name	Authors	Advantages	Disadvantages	
Multidimension- al model	Maté, Alejandro [19]	It is used for solv- ing problems of data visualization and analysis	In connection with the heter- ogeneity of the hypercube with heterogeneous data, their volume increases, which is unacceptable for Big data	
Object model	Chang, M [20]; Papakonstan- tinou, Y [21]	For a particular modification, it can be used to represent Big data	Unresolved is the task of displaying certain types of data in an object model	
Graph model	Zhou Feng [22]	Convenient when analyzing links between objects	Computational complexity of search algorithms provided a large number of objects	

So, for the healthcare domain all of Big data representation models can be used.

3 Methods and results

3.1 Big data platform development

The investigation is started from the platform for Big data in medical domain analysis. The platform consists of 6 layers (Fig. 1):

- Data layer,
- Communication layers,
- Preprocessing layer,
- Data processing layer,
- User layer,
- Integrational layer.

The first layer consists of all possible sources of information. Such approach allows usage of this platform as part of smart house too. Combination of indoor and outdoor parameters allows finding dependencies between patient's conditions and the weather and using these dependencies for the prediction of the new state of the patient.



Fig. 1. Platform architecture for Big data processing in medical sector.

The second layer is used for the communication between the rests of layers. It is very important to provide the secure transmission and saving of the personal data. In this paper, we are not considered the security aspects. More details about security features of medical data is given in [23].

The next layer is preprocessing layer. The medical data is collected from various sources given at data layer. That is why the filtering, cleaning, wrapping and missed data imputation methods are used. The next important thing is that medical data is collected from different countries. Each country has own medical protocols description written in different languages. In addition, different medical devices are used in hospitals from different countries. That is why the main tasks of preprocessing layers are:

• Anonimazation and Data protection - for medical data privacy,

- Missing data recovery [24, 25] for recovery the gaps from devices and medical records,
- Medical records preprocessing for semi structural data processing and comparison of the protocols in different countries,
- Pharmacian data preprocessing [26] for semi structural data processing and avoidance of language barriers.

Data processing layer consists of the following modules:

- Online diagnostic based on fuzzy neural networks [27] for fast training process,
- Personalization of the treatment based on the decision tree [28] for the most appropriative medications selection,
- The risk prevention model see below in more details,
- Statistical model for the risk prediction [29] for dependencies mining between parameters of the pation and the rest of smart home parameters.
- The ontology development [30] for increasing of the literacy of the young doctors.

The user layer represents the different groups of user interested in the platform usage.

The integration layer represents the list of standards and open platforms related to the proposed platform.

3.2 Big data platform development

The usage of a variety of processing technologies of emerging Big Data will allow to study causal mechanisms of modelling and predicting treatment stages, taking into account individual peculiarities of the patient, to analyze medicines' data and find their key characteristics. This information we use for development of innovative approaches to improving the methodology of risk stratification, to model the therapeutic schemes, to improve medical care quality through personalization of treatment regimens for patients.

The analysis of patients' information will help to identify common social characteristics and thus contribute to the development of recommendations for disease prevention

This also will stimulate the healthy people to monitor their health parameters.

Analyzing large amounts of data requires identifying attribute groups that form functional dependencies. However, in the real world, data sets are much more common, with important dependencies defined only on a subset of key attribute group values; we will call such dependencies partial functional dependencies. The partial functional dependencies are modified associative rules but for part of the data. The main idea of this method is to process data given in different structure or semistructured data. The algorithm for the Probabilistic Production Dependencies building is built on lazy calculation approach (modification of FP-tree). It allows reducing the time complexity and using the parallel and distributed mode for calculation.

Probabilistic Production Dependency (PPD) is the kind of functional dependency (F-dependency) arized in relational databases. This is also similar dependency to as-

sociative rules built on Aproory algorithm or something similar. The main difference between associative rules and PPD is that PPD will generate from existing functional dependencies in dataset. For this purpose, the Armstrong rules are used.

PPD:
$$K = \{a_i\}, a_i \in A, D = \{a_i\}, a_i \in A, : P(k \in K \to d \in D) = p,$$
 (1)

where k and d are the tuples of groups of attributes K and D respectively.

The main indicator of the reliability of such a dependency is the ratio of the number of objects that such PPD has to the number of objects in the selection:

$$P(\text{PPD}) = \frac{|\sigma_{k \in K \land d \in D}(R)|}{|\sigma_{k \in K}(R)|}$$
(2)

The calculation of the reliability of the implementation of such dependence is based on the possibility of decomposition of such dependence into components of the PPD:

$$P(s \in S \to t \in T) = \sum_{t_i \in T} P(s \in S \to t = t_i) = \sum_{t_i \in T} \frac{\sum_{j \mid s = s_j \land t = t_i \mid}}{\sum_{j \mid s = s_j \mid}}$$
(3)

As in the case of F-dependencies (functional dependencies), the set of classification rules that take place in a given relation can be represented by some subset of them, from which by means of output rules all classification rules of a given relation can be obtained. Since classification rules are an extension of F-dependencies, it is worth considering axiom transformations for functional dependencies for classification rules.

Therefore, the algorithm for PPD mining consists of two steps:

- 1. Frequent pattern mining based on Apriory algorithm.
- 2. The existing pattern splitting and decomposition based on reliability ration.

In the patient data analysis, the sequence of events is often of interest. When detecting regularities in such sequences, it is possible to predict with some degree the occurrence of events in the future, which allows us to make more correct decisions. A sequence is called an ordered set of objects. Using the hierarchy allows you to determine the connection that goes into higher levels of the hierarchy, since the support for the set can increase if the entry of the group, and not its object, is counted. In the hierarchical structure of objects, you can change the nature of the search by changing the analyzed level. Moving up the hierarchy, we summarize the data and reduce their number, and vice versa. So, Probabilistic Production Dependencies will be a combination of associative rules and sequential rules.

The proposed algorithm makes it possible to assert that the task of detecting Probabilistic Production Dependencies in distributed databases belongs to the class of P- tasks. Low asymptotic complexity of the established association rules mining algorithm and a wide set of data types supported for analysis allow to apply the established algorithm in practically all subject areas working with association dependencies in data. Algorithm for finding association dependencies is well-solved with MapReduce.

4 **Results**

The proposed algorithm is tested on open data set for associative riles mining [31]. The data set used for this application is the Adult data set in the Machine Learning Repository UCI.

The AOG model is an oriented acyclic graph, where each vertex of a graph corresponds to a variable with assigned parameters. In Bayesian networks, parameters are given as local conditional distribution of probabilities of the values of the variables P (Xi | F (Xi)).

And in Gaussian networks - as the coefficients of linear equations (for edges) and dispersion of deviations (for vertices).

The construction of the AOG-models meets the problem of reproducing the model from the statistical data. These include methods for restoring the "Collifinder" and "Proliferator-C" AOG model, generalizing the Chow & Liu method. The use of Collifinder and Proliferator-C allows recognition of transitive, synergistic, and combined associations, and thus provides a reliable and effective method for reproducing structures of single-threaded dependency models without first-level tests.

The problem of the above-described methods is the need to specify elementary conditional relationships for constructing a graph of the AOG-model.

Finding such dependencies goes beyond these algorithms.

Own algorithm. The paper proposes an algorithm for extracting PPD. For the relationship with the scheme $R = \{A_i, dom(A_i)\}, i = \overline{1, m}$, it allows you to find statisti-

cally significant rules that reflect the dependence of the attribute A_m on the utes $A_1, A_2, \ldots, A_{m-1}$, that is, the dependence of the species $A_1, A_2, \ldots, A_{m-1} \rightarrow A_m$.

As a measure of statistical significance, the Kullbach-Leibler information measure is used.

The algorithm allows you to look for only dependencies defined on the whole set of input data; in addition, it has a high computational complexity if there are many classification rules.

The result of own algorithm is given on fig. 2.





Parallel coordinates plot for 10 rules

Fig. 2. The result of created algorithm for associative rules building

If we need to search for associative rules among individual subjects, you can find that in many cases, associations with high support for individual events are practically absent. Therefore, in spite of the fact that, in general, event association support1 -> Event2 can be very high, association support between individual types of events is likely to be low. Thus, such associations, although they may be of interest, will be excluded from consideration as they will not meet a certain minimum support threshold S_{\min}

To solve this problem when seeking associative rules, not individual subjects, but their hierarchy is considered. If there are no such interesting associations on the lower hierarchical levels, then they may occur at higher levels. In other words, support for an individual object will always be less than the support of the group to which it belongs:

$$S(I) > S(i_i),$$

Where I the group is in the hierarchy; i_i j - this item is included in the given group. The reasons for this are obvious: the total support for the group is equal to the amount of support for the items included in it:

$$S(I) = \sum_{i=1}^{n} i_{j},$$

where n - the number of items in the group.

Associative rules found for objects or events located at different hierarchical levels are called multilevel rules.

Going down to the lower levels of abstraction, the descendants of only those categories and subcategories that are frequent sets are analyzed, that is, there are at least a predetermined number of times, where k - the number of the level.

Searching for frequent sequences runs from level 1 to the maximum possible. The results of successive passes will be presented in the Table 2.

1- sequend	ce	2- sequen	ce	3- sequenc	e	4- sequence	
F_1	Support	F_2	Support	F ₃	Support	F_4	Support
<1>	4	<1; 2>	2	<1; 2:3>	2	<1; 2; 3; 4>	2
<2>	2	<1; 3>	4	<1; 2:4>	2		
<3>	4	<1; 4>	3	<1; 3; 4>	3	-	
<4>	4	<1; 5>	3	<1; 3; 5>	2	-	
<5>	4	<2; 3>	2	<2; 3:4>	2	-	
		<2; 4>	2			-	
		<3; 4>	3	-			
		<3; 5>	2	-			
		<4; 5>	2	-			

Table 2. Prediction matrix

Thus, the maximum sequences are <1; 2; 3; 4>, <1; 3; 5> and <4; 5> because they are not contained in sequences of greater length. They will be sought after by successive templates.

5 Conclusion

The study presents the architecture of the platform for Big data preprocessing and processing. The state of art in this domain is presented. The method for prevention of the risk disease based on Probabilistic Production Dependencies is developed. The accuracy of the proposed method is estimated.

References

- 1. Norman, G. R., et al. (2018). Expertise in medicine and surgery
- Khanmohammadi, S., Adibeig, N., Shanehbandy, S. (2017). An improved overlapping kmeans clustering method for medical applications. Expert Systems with Applications, 67, 12-18
- Constantinou, A. C., et al. (2016). From complex questionnaire and interviewing data to intelligent Bayesian network models for medical decision support. Artificial intelligence in medicine, 67, 75-93
- 4. Y. Tang, Y. Wang, K. M. Cooper, L. Li, Towards big data Bayesian network learning-an ensemble learning based approach, International Congress on Big Data. pp.355-357, 2014
- Kjaerulff, U. B., Madsen, A. L. (2008). Bayesian networks and influence diagrams. Springer Science+ Business Media, 200, 114.
- Lakho, Shamshad, et al. Decision Support System for Hepatitis Disease Diagnosis using Bayesian Network. Sukkur IBA Journal of Computing and Mathematical Sciences, 2017, 1.2: 11-19
- Seixas, Flávio Luiz, et al. A Bayesian network decision model for supporting the diagnosis, 2014, 51: 140-158

- Bodyanskiy, Y., Dolotov, A. (2010). Hybrid systems of computational intelligence evolved from self-learning spiking neural network. Methods and Instruments of Artificial Intelligence, 17.
- Specht, D. F., & Romsdahl, H. (1994). Experience with adaptive probabilistic neural networks and adaptive general regression neural networks. International Conference on Neural Networks, Vol. 2, pp. 1203-1208.
- 10. Bhatt, Chintan; Dey, Nilanjan; Ashour, Amira S. (ed.). Internet of things and big data technologies for next generation healthcare. 2017
- E. L. Silva-Ramírez, R. Pino-Mejías, M. López-Coello, "Single imputation with multilayer perceptron and multiple imputation combining multilayer perceptron and k-nearest neighbours for monotone patterns", Applied Soft Computing, vol.29, pp.65-74, 2015
- D. Li, et al. Practical Data Mid-Platform Design and Implementation for Medical Big Data, Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), China, 2019, pp. 1042-1045.
- Yang and T. Chen, Analysis and Visualization Implementation of Medical Big Data Resource Sharing Mechanism Based on Deep Learning, vol. 7, pp. 156077-156088, 2019.
- M. Bansal and B. Gandhi, "IoT & Big Data in Smart Healthcare (ECG Monitoring)," 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, 2019, pp. 390-396.
- Suinesiaputra, P. Medrano-Gracia, B. R. Cowan and A. A. Young, "Big Heart Data: Advancing Health Informatics Through Data Sharing in Cardiovascular Imaging," in IEEE Journal of Biomedical and Health Informatics, vol. 19, no. 4, pp. 1283-1290, 2015.
- 16. S. Xu, et al. Cardiovascular risk prediction method based on test analysis and data mining ensemble system, International Conference on Big Data Analysis, 2016, pp. 1-5.
- Ruano M. G., et al, Reliability of medical databases for the use of real word data and data mining techniques for cardiovascular diseases progression in diabetic patients, Global Medical Engineering Physics Exchanges/Pan American Health Care Exchanges, Porto, 2018, pp. 1-6.
- Melnykova, N., Shakhovska, N., & Melnykov, V. (2019). Using Big Data for Formalization the Patient's Personalized Data. Proceedia Computer Science, 155, 624-629.
- Maté, A., Llorens, H., & de Gregorio, E. (2012, October). An integrated multidimensional modeling approach to access big data in business intelligence platforms. In International Conference on Conceptual Modeling (pp. 111-120). Springer, Berlin, Heidelberg.
- Veerapaneni, R., Co-Reyes, J. D., Chang, M., Janner, M., Finn, C., Wu, J., ... & Levine, S. (2020, May). Entity abstraction in visual model-based reinforcement learning. In Conference on Robot Learning (pp. 1439-1456).
- Jagadish, H. V., et al (2014). Big data and its technical challenges. Communications of the ACM, 57(7), 86-94
- Yang, Z., Zhou, F., Yang, L., & Zhang, Q. (2020). A new prediction method for recommendation system based on sampling reconstruction of signal on graph. Expert Systems with Applications, 113587.
- 23. Pirbhulal, S., et al (2019). A joint resource-aware and medical data security framework for wearable healthcare systems. Future Generation Computer Systems, 95, 382-391.
- Tkachenko, R., Izonin, I., Kryvinska, N., Dronyuk, I., Zub, K. (2020). An Approach towards Increasing Prediction Accuracy for the Recovery of Missing IoT Data based on the GRNN-SGTM Ensemble. Sensors, 20(9), 2625
- Shakhovska, N., Vovk, O., Kryvenchuk, Y.: Uncertainty reduction in Big data catalogue for information product quality evaluation, Eastern-European Journal of Enterprise Technologies, Vol 1, No 2 (91), 12-20 (2018).

- 26. Bodyanskiy, Y., Tyshchenko, O. (2019). A Hybrid Cascade Neuro–Fuzzy Network with Pools of Extended Neo–Fuzzy Neurons and its Deep Learning. International Journal of Applied Mathematics and Computer Science, 29(3), 477-488.
- Shakhovska, N., Fedushko, S., Shvorob, I., & Syerov, Y. (2019). Development of mobile system for medical recommendations. Procedia Computer Science, 155, 43-50. https://doi.org/10.1016/j.procs.2019.08.010
- Melnykova, N. (2017). Semantic search personalized data as special method of processing medical information. In Advances in Intelligent Systems and Computing (pp. 315-325). Springer, Cham.
- 29. Fedushko, S., Ustyianovych, T., & Gregus, M. (2020). Real-time high-load infrastructure transaction status output prediction using operational intelligence and big data technologies. Electronics, 9(4), 668. https://doi.org/10.3390/electronics9040668
- Lytvyn, V., Vysotska, V., Dosyn, D., Lozynska, O., Oborska, O. (2018). Methods of building intelligent decision support systems based on adaptive ontology. In 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), pp. 145-150.