New Algorithms for Improvement of Prediction Models Using Data Parallelism

Zurab Gasitashvili
Georgian Technical University
Tbilisi, Georgia
e-mail: zur_gas@gtu.ge

Merab Pkhovelishvili
Muskhelishvili Institute of Computational Mathematics
Tbilisi, Georgia
e-mail: merab5@list.ru

Natela Archvadze
I.Javakhishvili Tbilisi State University
Tbilisi, Georgia,
e-mail: natela.archvadze@tsu.ge

***Abstract*—The paper reviews the issues of improvement of existing prediction models, for which parallel data is used. These are the data, which simultaneously impact prediction of some event. The following stages of using this method are described in the article: Delimitation of prediction accuracy by presenting n-dimensional predictions in the n-dimensional space, including the algorithm for computation of error in this space, and the improved algorithm for using parallel data in the dynamic prediction tasks.**

***Keywords*—Prediction, Parallel Data, Heterogeneous Matrices, Dinamic Prediction.**

I. INTRODUCTION

Despite the great efforts performed for improvement of prediction models, there are still not discovered those universal models, by means of which, using the existing models, it would be possible to build models providing much better results during prediction. Hybrid models are mainly proposed, which do not significantly improve the accuracy of prediction. Our aim is to propose such algorithm, which would have much better accuracy compared to the existing best prediction model.

First, the algorithms for model building were reviewed for the task of static prediction, in particular, for prediction of technogenic disasters, more specifically, for prediction of earthquakes [1].

In the problems of earthquake prediction, where models are built through existing predecessors, the probability of success is very low, no more than 5% [2]. Using the method proposed by us, it is possible to select such pairs, which, in total, ensure much better accuracy of prediction.

The new approach for solving complex mathematical problems is reviewed in [3], and one of the examples is big data stores for prediction models, which are processes using parallel data. The term “parallel data” is used in scientific articles [4, 5, 6], however, it needs explanation.

Parallel data means introduction of new type of dependence between the data, which is called parallelism between the data [7]. Parallel data are various data, dependent on or impacting one event, which act (or are dependent on) for some time period (parallel by time) or in some location (parallel by location) on the event to be occurred and/or parallel by additional data. In the practice, using parallel data is possible for prediction of earthquakes or other disasters, economical (business, macro economy), political events (elections, positions of political forces), for effective solving of prediction tasks in the sphere of medicine and other fields.

For storage and processing of parallel data, the new type of matrix is used, called heterogeneous matrix [8]. In such matrices, columns may be of various types (numerical, textual, symbols, graphic images: geometric figure or stereometric figure, textual and video files). These data of various types are used for building of a matrix of heterogeneous data, and for its processing, it is necessary to broaden mathematical operations. These definitions are described in [9].

Heterogeneous matrix combines files numerical, textual files, and files of graphic images (geometric or stereometric figures), audio recordings or video files. The paper explains operations on heterogeneous matrixes, and a rationale is provided for necessity of defining these operations for prediction tasks. The operations are defined as on the elements of the same type in the heterogeneous matrix, as well on the elements of different types.

The main difficulties are in their realization: how do heterogeneous data operate and integrate the? Since the Het-matrix contains different types of elements, in [9] let’s define two types of operations: a) Operations on the same type elements (data) of the matrix; and b) operations on different types of matrix elements (data). Operations involve the following operations: addition, subtraction, multiplication, division, and comparison operations: equal, not equal, more than or equal, less, less or equal.

Operations on the same type elements of the Het-matrix. Suppose we have two Het-matrices which have the same type of data. Operations will be determined in accordance with the types and, as far as possible, we will assert the need to determine this operation for the purpose of forecasting for example, for the Addition operation (+).

- Addition of numbers are determined according to the definitions in mathematics.
- Addition of texts is defined as the action of "union", the same "merge" or the concatenation. For example: "new" + "string" = "newstring".

The necessary condition for addition of geometric figures is that these figures should be presented in their coordinate systems. The addition of the two figures is defined as

following: both figures and then their union should be placed in the same coordinate system.

The feasibility of the definition of this operation is derived from the earthquake prediction problem that includes location definition too where earthquake may occur. The other event function defines the other location and for the united forecasting results it is necessary to add them.

- An addition of two sound recordings is defined as the sound recordings are placed on each other. For instance, if one record is only music, and the other recording-singer voice, the recording of the two recordings is a song recording where the singer sings with the music in the background.

While earthquake forecasting, two sound recordings are made at some depth. Such sounds may not be detected by the human ear. When different voices are recorded, it is necessary to find out what they have common (by the tone, wave). If geophysicists define precursor that a certain level of noise leads to an earthquake, it may be possible to make prediction by the sound level.

Operations of sound type can be defined as the other way too: By addition of two sound elements an element can be obtained in which the second sound starts after the first ends; by subtraction – the second will be deleted from the first one, etc.

- Addition of two video files can be defined as video clips are placed on each other. Placement means to rewrite the second video clip on the first one. For example, in the movie, one actor plays the roles of two twin brothers. At first one brother will be filmed, then the other one, and by placement such movies to each other both brothers can be seen at the same time.

The need to work on the video information is well illustrated in the earthquake prediction problems. In this case, filming takes place where there is a need to analyze the changes of images. For example, the length or the width or the depth of the rupture changed. Another example is the recording of radiation in the ionosphere. If the rising rate is observed, the expected earthquake area can be determined. At this time it is necessary to determine the movement rate and localization of this area.

- Addition of the two vector type objects can be defined as the combination of both elements.
  For example,
  (“asd“,”kjh”,”aas”)+(“sd”,”jh”,”KK”)=(“asd“,”kjh”,”aas” ,“sd”,”jh”,”KK”).
- Operations on Het-matrix type values are defined as gathering all relevant elements. At this time, if the elements are not of the same type, then the result of the addition is the zero element.

Other actions (Subtraction operation (-), Multiplication operation (*) and Operations on different types of Het-matrix elements), are interpreted in the same way in [9].

[10] discusses the algorithm of calculation of probability for prediction accuracy using “parallel data” - the “parallel probability”. This algorithm allows to select those prediction pairs (or triples, fours, etc.), “joint” probability of prediction accuracy of which ensures much better result than each of them separately.

Prediction models are used in many fields: Economy, business, macroeconomics, weather, elections, results of sport competitions, finances, crime, etc.

All these fields have the following common basic principles [11, 12]: a) complex models bases on statistics, are not always more precise than simple models; b) combining models or computer forecasting obtained through various models, averagely improves the accuracy of prediction; c) by increasing prediction horizon, the accuracy of prediction is reduced. Therefore, use of standard methods is not beneficial.

The following may be considered as business prediction objectives: demand, intermittent demand, temporal and spatial hierarchies, shares, macroeconomic indicators, commodity groups, new products and so on. [13,14] review two objectives of business prediction: Prediction of demand and intermittent demand by using the paradigm of “parallel data”.

Prediction model based on using of heterogeneous matrices is reviewed in [15]: the paper deals with the prediction model based on heterogeneous matrices, and prediction process is reduced to heterogeneous matrix processing to identify prediction precursors the joint use of which increases the probability of the occurrence of the event. This paper proposes the idea of usage of special splits only for certain precursors, which makes it possible to use cloud services capabilities for prediction tasks using the heterogeneous matrices as well. Big data systems also include prediction tasks so Azure IoT Hub is used as a managing service, hosted in the cloud, that acts as the central message hub for bi-directional communication between IoT application and the devices it manages.

II. “COINCIDENCE“ DURING PREDICTION

In article [6], development of prediction models was given by means of parallel data.

The main idea was that to select such pairs, triples, etc. from several models, which give much better result than a single better model from them or two models separately. This was done in the following way: Such models were found, the number of unsuccessful identical predictions of which was as low as possible, although both had successful predictions simultaneously, because this number of coincidence of unsuccessful predictions was much less; of course, the probability was higher that they would simultaneously ensure better prediction (with higher probability). [Fig. 1] Data of two models L_i and L_j compared to actual L_{real} data.

In this paper, the term “coincidence” was mainly used, which, of course, needs a more specific definition, because, if hours, minutes, and seconds coincide, this prediction time will not be precise. Difference also will be in relation to prediction location, where the given event has to occur. Of course, prediction cannot be made with accuracy of 1 km or 1 m or less. Also, regarding the additional parameters, such as earthquake power by Richter magnitude scale, further deviation of earthquake, direction, in which impact force will spread or which type it will be, vertical or horizontal impact, etc. The term “prediction coincidence” must be defined more specifically, depending on what they are related to.

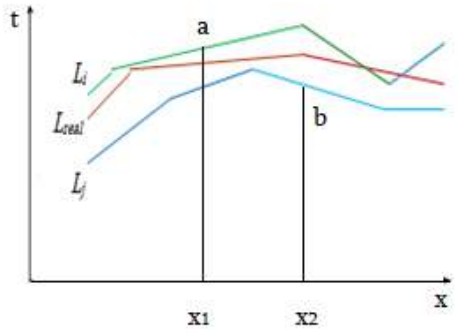

Fig.1. Graphical representation of the parallel probabilities of the GEL exchange rate.

Of course, the "unit of coincidence accuracy" must be defined by experts. For example, in case of earthquake we can take that coincidence to the epicenter of the quake is with radius of 50 km. Take coincidence in time as 24 hours. Of course, only coincidence with indication of day may not be sufficient, because one prediction may indicate 00:30 and another - 23:55 and this may be considered coincidence if we take only data of days, or it may be 23:50 and the second prediction is 00:20 and difference is only 30 minutes, but because they occurred in different days, it may be considered that they do not "coincide" due to different days. Therefore, we should set a time interval. For example, 24 hours.

As for power, predictions mainly are for those earthquakes, where power is 5 or higher on the Richter magnitude scale. But for coincidence we can take difference in power of 0.5.

III. Spatial models

If we have only 3 data and build given prediction points in the relevant 3-dimensional space: in this case, x is location, t is time and v is power. Assume, that each has its own dimension. For example: Location - plain. In this case, 4, 5 of more dimensional model will be built, depending on prediction type. Fig. Prediction values are presented in 2-dimensional space.

Here, in Fig. 2 space, points a (50, 22,3.7) and b (57, 7, 4.1) are shown. X axis means distance, t is time, and V axis - earthquake power. There is a distance between them on the figure. We say that these two points, i.e. two predictions coincide with each other, if distance between them is more than one unit. For specific examples, the distance is calculated as mathematics calculate distance. In this case, we assume that these 2 points have a distance between them of 1 unit, or they almost coincide.

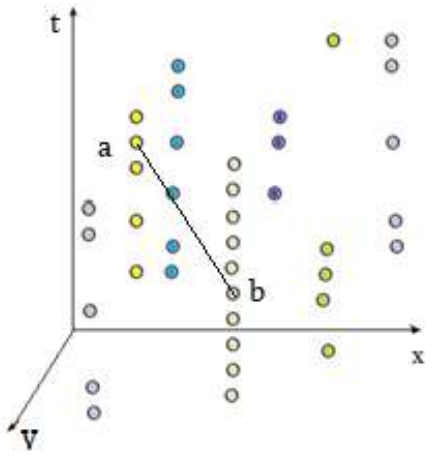

Fig. 2 Presentation of prediction data in space.

For statistical events, this common coincidence in n-dimensional space between $p(p_1,p_2,\ldots,p_n)$ and ween $q(q_1,q_2,\ldots,q_n)$ points are calculated by a simple formula:

$$d_{pq} = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

Dynamic prediction requires slightly different approach. In this case, coincidence is defined considering the fact that predictions are given regularly: daily, hourly, or even more frequently, for example, exchange rate predictions [10, 11]. Here such models were selected, the prediction data of which were close to actual data in the result of acting on actual data with some function (on these pairs, triples, etc.). Marked values are those data, which satisfy the event function, which can be a function-predicate (with true or false values), or unclear values. It is possible that the values of event function would be statistical data (Fig. 3):

Normally, the prediction is chosen, the prediction value of which is closest to actual data (higher or lower). In this case, 3.15, 3.16 were close to prediction and not to actual value. We should take 3.15 in pair with 3.13, because their arithmetic mean is 3.14, i.e. we do not take separately 3.15 and 3.13, but we take their pair. This inaccuracy is measured almost every day, thus, if we take arithmetic mean of these points, it will be much closer to the actual data, than each of them separately. Other functions also may be used - geometric mean, etc.

Dynamic prediction requires slightly different approach. In this case, coincidence is defined considering the fact that predictions are given regularly: daily, hourly, or even more frequently, for example, exchange rate predictions [10, 11]. Here such models were selected, the prediction data of which were close to actual data in the result of acting on actual data with some function (on these pairs, triples, etc.).

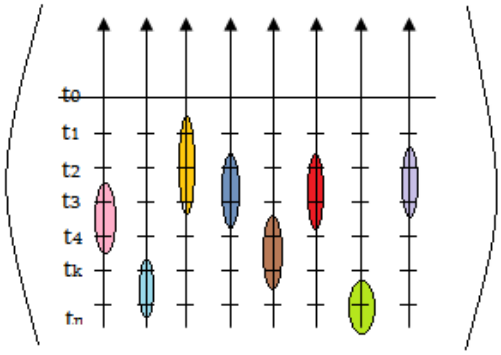

Fig.3 The expandable matrix with event function

Normally, the prediction is chosen, the prediction value of which is closest to actual data (higher or lower). In this case, 3.15, 3.16 were close to prediction and not to actual value. We should take 3.15 in pair with 3.13, because their arithmetic mean is 3.14, i.e. we do not take separately 3.15 and 3.13, but we take their pair. This inaccuracy is measured almost every day, thus, if we take the arithmetic mean of these points, it will be much closer to the actual data, than each of them separately. Other functions also may be used - geometric mean, etc.

IV. Prediction of More than One Data

Consider the case, where prediction should be done with 2 or 3 data, for example, the model must provide exchange rate of GEL to dollar, and price of gold in dollars. Then we return to the similar 2 or 3-dimensional model, depending on what occurs.In this case N-dimensional space is also built. Not a distance is calculated between the points, but a pair of actual data is indicated. In the given moment of time, the distance in the space is calculated in the points, which are close to actual data, and they will be added. The best pair will the one, which in the sum will give a close value to the actual data.

Two prediction data is in one prediction, for example, gold price and exchange rate of GEL to dollar, models are also different. We have a_1 model, which provides every day, how much the gold price and GEL exchange rate to dollar should be today, tomorrow, and day after tomorrow. We have another a_2 model, which gives gold price and GEL exchange rate to dollar. Similarly, during the earthquake, we draw relevant points in the space. Distance of spatial lines between each other is more than their distance separately to the main actual data (for example, one might be at the top, and another at the bottom).

V. Conclusion

The method proposed by us can be used not only for prediction of earthquakes, but also for other events, which are hard to predict and which use the set of precursors of arious types (often with low probabilities). Using a parallel data in these methods allows us not only to perform prediction with more probability,but also divide it into the sequental stages, which makes it controlable to monitor the expected event and take appropriate measures. Also, by using the method of parallel data, it is possible to detect new regularities, which are not yet detected because of limitation of sequential algorithms.

We showed that in the dynamic models too, building spatial models may be necessary (not with one data, but with several data), and added dynamic prediction models with two or three arguments.

Reference

[1] N.Archvadze, M.Pkhovelishvili. "Prediction of Events by Means of Data Parallelism". *Proceedings of International Conference on Matematics, Informatics and Informtional Technologies (MITI2018).* pp.120-121, 2018.

[2] Завьялов А.Д. Прогноз землетрясений: состояние проблемы и пути решения, в журнале Земля и вселенная, № 5, с. 66-79, 2018.

[3] Z. Gasitashvili, M.Pkhovelishvili, N.Archvadze. "Prediction of events means of data parallelism". *Proceedings - Mathematics and Computers in Science and Engineering, MACISE-2019*, pp.32–35, 2019. https://ieeexplore.ieee.org/abstract/document/8944725

[4] Y Chen, Y Lv, FY Wang. "Traffic flow imputation using parallel data and generative adversarial networks" - *IEEE Transactions on Intelligent,* 2019.

[5] J Bhimani, N Mi, M Leeser, Z Yang. New performance modeling methods for parallel data processing applications - *ACM Transactions on Modeling*, 2019.

[6] D Skillicorn. "Strategies for parallel data mining". *IEEE concurrency*, 1999.

[7] Z.Gasitashvili, M.Pkhovelishvili, N.Archvadze. Usage on Different Types of Data to Solve Complex Mathematical Problems. *WSEAS Transactions on Computers*, vol. 18, Art. #7, pp. 62-69, 2019.

[8] M.Phkhovelishvili, N.Jorjiashvili, N.Archvadze. "Usage of heterogeneous data and other parallel data for prediction problems". PRIP'2019. *Pattern Recognition and Information Processing (Proceedings of 14th International Conference (21-23 May, Minsk, Belarus). pp.178-181. Minsk,"Bestprint",* 2019.

[9] M.Phkhovelishvili, N.Jorjiashvili, N.Archvadze. "Using Different Types Data Operations for Solving Complex Mathematical Tasks". *Computer Science and Information Technologies. Proceedings of the conference,* Yerevan, Armenia. pp.187-190, 2019.

[10] S.Makridakis, E.Spiliotis, V.Assimakopoulos. The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting. vol.36, Issue 1, pp. 54-74*, 2020.

[11] Z.Gasitashvili, M.Phkhovelishvili, N.Archvadze, N.Jorjiashvili. "An Algorithm of Improved Prediction from Existing Risk Predictions. *Published by AIJR Publisher in "Abstracts of The Second Eurasian RISK-2020 Conference and Symposium", pp. 31,* 2020.

[12] Makridakis, S.,Hib on,M., The M3-Comp etition: results, conclusions and implications. *International Journal of Forecasting* 16, pp.451-476, 2000.

[13] M. Pkhovelisvili, M.Giorgobiani, N. Archvadze, G. Pkhovelishvili. "Modern Forecasting Models in Economy". *Proceedings of Materials of International Scientific Conference „Modern Tendencies of Development of Economy and Economic Science".* pp. 219-224, 2018.

[14] N.Archvadze, M.Pkhovelisvili. "Modern Forecasting Models in Economy". *X International Conference of the Georgian Mathematical Union. BOOK OF ABSTRACTS*. pp.55, 2019.

[15] G. Chogovadze, G. Surguladze, N. Topuria, N. Archvadze, Implementation of a prediction model with cloud services. *Bulletin of the Georgian National Academy of Sciences*, 14(3), pp. 29–35, 2020.