New Algorithm for Building Effective Model from Prediction Models Using Parallel Data

Zurab Gasitashvili
Deputy Rector for Science
of Georgian Technical
University, Professor
Tbilisi, Georgia
zur_gas@gtu.ge

Merab Phkhovelishvili
N. Muskhelishvili Computation
Mathematic Institute of
Georgian Technical University
Tbilisi, Georgia
merab5@list.ru

Natela Archvadze
Dept. of Computer Sciences
Ivane Javakhishvili Tbilisi State
University
Tbilisi, Georgia
natela.archvadze@tsu.ge

***Abstract.* Building much more effective new hybrid models from prediction models, using parallel data is discussed. The algorithm for selection of model pairs and its advantage over any best prediction model is provided. The advantage of prediction models with higher number of pairs over lower number of pairs is shown and the algorithm of taking into consideration the "approximate coincidence" of predictions is discussed when selecting pairs.**

***Keywords:* parallel data, prediction models, approximate accuracy, probablity of prediction success**

I. Introduction

Some prediction models are based on using of "parallel data" [1-4], although it must be noted that the term "parallel data" is differently explained in each of them.

In practice, parallel data is used during prediction of various events, including natural disasters: earthquake, landslide, tsunami, mudflow, etc., for prediction 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.

The definition of parallel data is based on introduction of new type of dependence between the data, which is called "parallelism between the data" [1, 5-8]. Parallelism between the data is mutual dependence between those data, which are used for prediction of the same event. Various data affecting the same event may exist in different periods (parallel by time) or locations (parallel by location) and/or provide other additional information on prediction of the same event [9, 10].

The main idea of algorithms for building of prediction models is reviewed by us through parallel data and is the following: Let us assume that there are several models of prediction. From them, it is necessary to select such pairs, triplets, etc. from several models, which give much better result than a single best model from them or two models separately.

This algorithm was the following: such models were found, for which the number of coincidences of unsuccessful predictions for some given event was as low as possible, but successful predictions were necessary for them.

In this paper we first review static prediction models for natural disasters, when a result(s) of prediction should be guessed, for example, when, where and with which specifications occurred the event of interest.

Unlike static predictions, a prediction is dynamic, when for each time interval it is necessary to forecast an event of certain value. Such is, for example, a daily forecast of exchange rate, forecast of oil price, monthly subsistence level, annual income, human health condition, scope of coronavirus spread, etc.

The distinctive sign, by which the static prediction is different from the dynamic one, is its dependence on the time of prediction event. Actually it means that we should distinguish, how a result, i.e. prediction values, are declared. If it occurs continuously, with some predefined time interval, then this is dynamic prediction, but if time is one of prediction elements, then it is static prediction. For example, earthquake prediction implies declaring that date as one of the results, when earthquake is expected, therefore, it belongs to static prediction, and currency exchange rate is forecasted daily, therefore, it is dynamic prediction.

In this article we will establish 4 lemmas for the task of static prediction and show, how the accuracy of such models is increased through our algorithm. Specific data are taken for earthquake prediction task. Each prediction model is build based on certain predecessors. For earthquake the predecessor is geophysical phenomenon (mainly), which precedes the actual earthquake. For their part, geophysical precursors are divided into the following categories: seismic, hydro geodynamic, deformation, geochemical, thermal, gravitational, electromagnetic and, precursors obtained via remote monitoring by means of satellite technologies developed recently [11].

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk :
UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

Despite the fact that quite high number of predecessors exist, not any of them ensures high-accuracy prediction for time, place and magnitude of future earthquake. The probability of successful prediction of each predecessor (ratio of number of successful predictions to the number of all given predictions) does not exceed 0.5% [12]. One of the ways for overcoming this situation is to use several prediction predecessors simultaneously, although for each of them it is necessary to perform observation for a long time and process vast amount of data, which is not done in many models till now. “The practice of recent years show that their simultaneous use would improve the reliability and efficiency of prediction assessment, at least in medium-term (first years) prediction”.

II. “SUCCESS PROBABILITY” OF PREDICTION

Assume that we have several prediction models, which provide some predictions through their predecessors (for example, for earthquakes - when it would occur, at which location and with which magnitude). These predecessors should be “necessary predecessors” that means that if earthquake occurs, they will inevitably provide the prediction. If some predecessors do not provide prediction on actually occurred earthquake, it will be no longer considered.

We study history, let's assume that there is plenty of data and it is necessary to calculate, based on predecessors, how many times the prediction of earthquake occurrence was given and how many times actual earthquake occurred. Assume that we consider the necessary predecessors and the models created for them: $A_1, A_2, ..., A_n$, where *n* is the number of considered predecessors. *t* denotes time, during which we perform analysis and the number of actually occurred earthquakes is *m*. We calculated the number of earthquakes predicted by each predecessor: $p_1, p_2, ..., p_n$. For example, A_i model, which was based on *i* predecessor, predicted earthquake occurrence p_i - times.

For each p_i let’s calculate quotients of *m*, the number of actually occurred earthquakes, write it in % and designate with K_i:

$$K_i = \frac{m}{p_i}100\%.$$

For example, if earthquake actually occurred 4 times, and we calculate $K_i = \frac{4}{20}100\% = 20\%$ then the probability of A_i success will be 20%.

Put the sequence of model success in descending order and this sequence denote as: $k_1, k_2, ...k_n$ sequence. k_i is a model created for i -th predecessors. We get that k_1 highest value, which was determined by the prediction, the value of k_2 is less than that of k_1 and so on.

It is necessary to consider a combination of models (two, three, etc.) and assessment of the probability of their combined success. The assessment and selection of combinations is done according to the parallel probabilities [13].

Lemma 1 - If such pairs of models are selected, for which the number of coincidences of unsuccessful predictions for some given event was as low as possible, but successful predictions were necessary for them, then the success probability calculated for combination of any pair so selected is always higher than or equal to the success probability of best from them.

If we take best k_i and pair it with any, even the worst value k_j, then their combined result is not worse than k_i. Proof is based on fact that for the pair k_i and k_j (ki< k_j), then the intersection of their successes, of course, is less than or equal to k_i. For example, if k_j gives conclusion that earthquake occurs 5 times, even if others give values 10 or 7, the intersection of their successes cannot exceed 5. Therefore, whether pairs, triplets, quads are selected, their combination always give better result than best of them.

Lemma 2 - The higher number of intersections of prediction models, the better prediction we would get.

For example, the best triplet - combination of three predictions would give better result than the best pair of prediction (deuce), the best quad gives better result than the best triplet and so on.

This follows from the fact that intersection of any pairs with third is lower number than each pair.

III. SELECTION OF BEST PAIRS

Lemma 3 - The best pair is one that does not have intersection between each, except actual, occurred predictions.

For example, we calculated the pair or k_i and k_j predictions and let's calculate, by their combined prediction, which number of coincidences we have with the actual situation (for example, coincided 10 times). For example, the event occurred actually 2 times, if we calculate success %, we get $\frac{2}{10}100\% = 20\%$, but if it

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

turned out that k_i and k_j jointly only two times had prediction success for the event, then it means that k_i and k_j is the best pair.

It is an interesting metamorphose - it turns out that if we consider all those models jointly, which have many errors, their combination may give the best result.

IV. Modification of prediction model

Geophysical characteristics of environment constantly change, for example, an average temperature, erosion of ocean shores, etc. The question is, how to plan the change of selected scheme of prediction, from which time pairs, triplets, etc. should be selected and success probabilities recalculated, what we should do, recalculate everything that occurred till today?

Lemma 4 - Modification of prediction model, i.e. recalculation should be done from the day of last earthquake occurrence.

For consideration of relevant pairs of models, it is necessary to take corresponding figures from such moment, when we have the aggregate of all input data. Of course, it is possible that new prediction models may be introduced with the data of relevant predecessors, and additional regulations are required for consideration, because of search for the relevant pair. It is possible that after each actually occurred event, the selected pairs of prediction models would be changed and other pairs would become better for prediction. Therefore, selection of each new pair should be done after occurrence of each event during static prediction, and for dynamic prediction, the process of determination of such pairs should be regulated within certain time periods. For example, if we have daily prediction data, new pairs should be selected at least once a week.

V. "Approximate coincidences" of predictions

In accordance with Lemma 2, the prediction pairs are selected. When selecting them, we determined the number of "accurate" coincidences of predictions. Now this is not sufficient and a "coincidence accuracy" should be determined. Of course, prediction data should not directly coincide with each other, but coincide within certain intervals of time, place or other characteristic.

Lemma 5 - When selecting the pairs of prediction models, an "approximate coincidence" should be taken into account.

An interval of "approximate coincidence" is determined with the help of experts. It may be prediction of such time period, when occurrence of given event is expected, or determination of certain radius from the epicenter. This task is faced in cases of earthquake, virus origin, beginning of military conflict, etc. Of course, time interval has great importance. For earthquakes, a short-term forecast - where (with radius of 50 km), when (with interval of 24 hours), and with which magnitude (with difference of 0.5) the earthquake is expected.

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 how much parameters are in prediction. Prediction data are presented in 3-dimensional space on Fig. 1. Here the distance between two points (predictions) is the error between their predictions. The value of "approximate coincidence" (i.e. this distance) is determined by an expert.

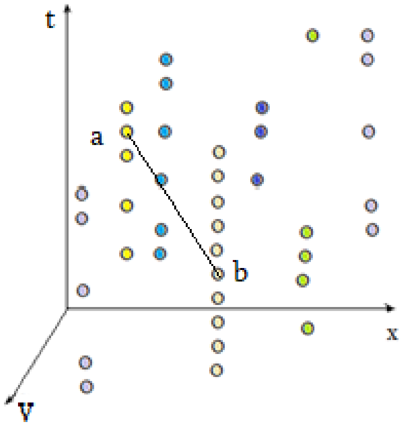

Fig. 1. Presentation of prediction data in space

Introduction o "approximate coincidences" sharply increases the volume of calculations needed for selection of pairs of prediction models. It is necessary to utilize higher computational capacities, technical capabilities of supercomputers and use the algorithms of parallel computations and relevant programs. It should be noted that, in our opinion, this can be realized by using powerful parallel and recursive computations of programming language F#, as we prove it in [14].

VI. Conclusion

We considered the possibility to build much more effective new hybrid models from prediction models, using parallel data, and by means of lemmas we state that: The advantage of selection of model pairs over any best model of prediction, it was shown that the more is the number of model pairs, the more is advantage over the lesser numbers of pairs.

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk :
UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

The best pair is one that does not have intersection between each, except actual, occurred predictions. When to reselect the pairs of prediction models and that it is necessary to consider “approximate coincidences” of predictions in this case.

References

[1] 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, 8944725, 2019. https://ieeexplore.ieee.org/abstract/document/8944725.

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

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

[4] D Skillicorn. Strategies for parallel data mining. - IEEE concurrency, 1999 .

[5] 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.

[6] Z. Gasitashvili, M. Pkhovelishvili, N. Archvadze. Usage on Different Types of Data to Solve Complex Mathematical Problems. WSEAS Transactions on Computers, vo. 18, Art. no. 7, pp. 62-69, 2019.

[7] 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.

[8] 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 (September 23-27, 2019), Yerevan, Armenia, pp. 187–190, 2019.

[9] 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“. Ivane Javakhishvili Tbilisi State University Paata Gugushvili Institute Of Economics. pp. 219-224, 2018.

[10] N. Archvadze, M. Pkhovelisvili. Modern Forecasting Models in Economy X International Conference of the Georgian Mathematical Union. Book of abstracts, pp. 55, 2019.

[11] A. D. Zav'yalov Prognoz zemletryaseniy: sostoyaniye problemy i puti resheniya, v zhurnale Zemlya i vselennaya, № 5, pp. 66–79, 2018 (in Russian).

[12] A. D. Zav'yalov Srednesrochnyy prognoz zemletryaseniy: osnovy, metodologiya, realizatsiya, Izdatel'stvo Nauka; 2006 (in Russian).

[13] 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” April 12-19, 2020, Tbilisi, Georgia, pp. 31, 2020.

[14] N. Archvadze**,** M. Pkhovelishvili. Reforming the Trees – C# and F# comporation. International Conference on “Problems of Cybernetics and Informatics" (PCI'2012). pp. 93-96, 2012.

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.