Full name: Anastasia MOTRENKO
Birth date: May 19, 1992
email: anastasia.motrenko@gmail.com

Education

- 2012 — 2014 Moscow Institute of Physics and Technology (State University), department of Applied Mathematics and Control/Computing Centre of the Russian Academy of Sciences (Master, supervised by Vadim Strijov)

- 2008 — 2012 Moscow Institute of Physics and Technology (State University), department of Applied Mathematics and Control (Bachelor, supervised byVadim Strijov)

Research areas and associated publications

- Motrenko A., Strijov V. Obtaining an aggregated forecast of railway freight transportation using Kullback-Leibler distance // *Informatics and applications*. 2014. V. 8, N. 1. Pp. 86-97. [In Russian]

  This study addresses the problem of obtaining an aggregated forecast of railway freight transportation. To improve the quality of aggregated forecast, we solve a time series clusterization problem, such that the time series in each cluster belong to the seme distribution. Solving the clusterization problem, we need to estimate the distance between empirical distributions of the time series. We introduce a two-sample test based on the Kullback-Leibler distance between histograms of the time series. We provide theoretical and experimental research of the suggested test. Also, as a demonstration, the clusterization of a set of railway time series based on the Kullback-Leibler distance between time series is obtained.

- Valkov A.S., Kozhanov E.M., Motrenko A.P., Husainov F.I. Constructing a cross-correlation model to forecast the utilization of a railway junction station // *Machine Learning and Data Analysis*. 2013. V. 1, N. 5. Pp. 505-518. [In Russian]

  The problem of detecting causal relationships between time series is studied. The authors propose a forecasting model that considers detected relationships. The model is aimed to forecast the utilization of a railway junction station. The model relies on the history of a junction station utilization as well as on the time series for the main financial instruments and regulations. Expert's assessments are used to construct the model. A method that evaluates plausibility of the expert's assessments is proposed. The method is illustrated with the Russian Railways data.

- Motrenko A., Strijov V., Weber G.-W. Sample size determination for logistic regression // *Journal of Computational and Applied Mathematics*. 255(2014), 743-752. [In English]
  Motrenko A.P. Bayesian sample size estimation for logistic regression // *Journal of Machine Learning and Data Analysis*. 2012. V. 1, N. 3. Pp. 354-366. [In Rissian]

  The problem of sample size estimation is important in the medical applications, especially in the cases of expensive measurements of immune biomarkers. The paper describes the problem of logistic regression analysis including model feature selection and includes the sample size determination algorithms, namely methods of univariate statistics, logistic regression, cross-validation and Bayesian inference. The authors, treating the regression model parameters as the multivariate variable, propose to estimate sample size using the distance between parameter distribution functions on cross-validated data sets.

- Motrenko A.P. Joint probability density estimation // *Journal of Machine Learning and Data Analysis*. 2012. V. 1, N. 4. Pp. 428-436. [In Rissian]

  When solving a classification problem one often has to deal with both discrete and continuous variables.

for example, in the logistic regression independent variables are distributed continuously, while a target variable follows Bernoulli distribution. In this paper a method is resented that allows to estimate joint probability distribution which include discrete and continuous variables. A case when no probabilistic assumptions can be made is considered. The methods of nonparametric regression are used. Also a comparison to the classic methods of probability theory is presented. The experiment is conducted on the real and synthetic data.

- Motrenko A.P., Strijov V. V. Multiclass classification of cardio-vascular disease patients // *Journal of Machine Learning and Data Analysis*. 2012. V. 1, N. 2. Pp. 225-235. [In Rissian]

International conferences

- Bayesian Sample Size Estimation for Patient Classification Survey // IFORS 2014, Barcelona, July 2014.

  We seek to increase the quality of classification of Cardio-Vascular Disease patients. As a part of research, arises the problem of determining the minimum sample size necessary for statistical significance of classification. Previously, we proposed a method of sample size determination that involved comparing empirical distributions, evaluated on different subsets of a sample. To measure similarity, the Kullback-Leibler distance was used. We now investigate further the features of this distance and provide some theoretical background for the method.

- Small CVD sample set classification: generative versus discriminative // XXVI EURO conference, Rome. July 2013.

  The challenge of the Cardio-Vascular Disease patients classification problem is the small sample size. To make a classification model we combine generative and discriminative classifiers (known from the supervised and non-supervised approaches to Machine Learning). Our goal is to obtain the maximum generalizing ability of the classifier. The quality function is a linear combination of generative and discriminative likelihoods. It includes evaluation of both discrete and continuous joint distribution of random variables. We study the dependence of the combination structure on the sample size.

- Multiclass classification of cardio-vascular disease patients with sample size estimation // XXV EURO conference, Vilnius. July 2012.

  We discuss an algorithm that classifies four groups of patients, divided by their health condition. Concentrations of proteins in blood cells are used as features. Our first objective is to select a set of features that will classify the patients making minimum amount of errors. This selection is implemented by means of exhaustive search. Two classification strategies are investigated, "one versus all"and "all versus all". The second objective is the sample size. Amount of data is small, so we evaluate minimum sample size, necessary for statistical significance of classification.

Participation in academic projects and financial support

- In project "Algorithms for generalized linear classification model selection for low-volume samples supported by RFBR (Russian Foundation for Basic Research) in 2012-2013 (12-07-31095) as the project leader.

  Model selection in classification and forecasting problems is studied. A hypothesis is accepted that the distribution of the predicted variable is in the class of exponential distributions. The aim of the investigation is to select a model of an optimal complexity. The study targets developing and proving methods for classification models selection, as well as objects sampling and objects' class forecasting. The

main issue is that the sample size is insufficient for accepting or rejecting the data generation hypothesis reliably. The problem of model selection for small-volume samples is set for the first time. The problem is based on both Russian and international prior research. Joining generative and discriminative model selection approaches is suggested for solving the classification task.

As a member:

- "Developing and analysis of classification models for sample sets of small size supported by RFBR (Russian Foundation for Basic Research) in 2014-2016 (14-07-31045).
- "Development of theory of multivariate time series forecasting supported by RFBR in 2014-2016 (14-07-31046).
- "Development of the Theory on Hierarchical Model Selection for the Problems of Structure Learning supported by RFBR in 2013-2015 (13-07-00709).
- "Methods of mutual influence analysis for passenger and freight traffic at he Russian Railways supported by RFBR together with Russian Railways in 2013-2014 (13-07-13139).

Extra Curricular Activities

- (since 2013) assistant editor in the Machine Learning and Data Analysis journal.
- assistant chair in streams Decision Making in Inventory Systems, Mathematical and Data Models in Decision Making, Process Development and Decision Making, Methods of Multi-criteria Decision Analysis (XXVI EURO conference, Rome. July 2013)
- assistant chair in streams Data Mining: Web and Social-Oriented Applications, Applications of the Neural Networks (XXVI EURO conference, Rome. July 2013)