Algorithm for Processing the Results of Cloud Convection Simulation Using the Methods of Machine Learning

E. N. Stankova[1(✉)], E. T. Ismailova[1], and I. A. Grechko[2]

[1] Saint-Petersburg State University,
7-9, Universitetskaya nab., St.Petersburg 199034, Russia
e.stankova@spbu.ru, elaismaylova@gmail.com
[2] Saint-Petersburg Electrotechnical University "LETI", (SPbETU),
ul.Professora Popova 5, St.Petersburg 197376, Russia
grechko.irinka@gmail.com

Abstract. Data preprocessing is an important stage in machine learning. The use of qualitatively prepared data increases the accuracy of predictions, even with simple models. The algorithm has been developed and implemented in the program code for converting the output data of a numerical model to a format suitable for subsequent processing. Detailed algorithm is presented for data pre-processing for selecting the most representative cloud parameters (features). As a result, six optimal parameters: vertical component of speed; temperature deviation from ambient temperature; relative humidity (above the water surface); the mixing ratio of water vapour; total droplet mixing ratio; vertical height of the cloud has been chosen as indicators for forecasting of dangerous convective phenomena (thunderstorm, heavy rain, hail). Feature selection has been provided by using recursive feature elimination algorithm with automatic tuning of the number of features selected with cross-validation. Cloud parameters have been fixed at mature stage of cloud development. Future work will be connected with identification of the influence of the nature of the evolution of the cloud parameters from initial stage to dissipation stage on the probability of a dangerous phenomenon.

Keywords: Machine learning · Numerical model of convective cloud
Weather forecasting · Thunderstorm · Data preprocessing · Feature selection

1 Introduction

Global warming produced by permanent anthropogenic influence on the atmosphere leads to an increase in the intensity of convective processes. The increase in temperature and the increase in air humidity are two facts that together lead to an intensification of active convection in the atmosphere, which in turn entails an increase in the number of heavy rains, an increase in thunderstorm activity, an increase in the number of tornadoes and an increase of other dangerous convective phenomena that have a tremendous destructive effect. Therefore, the problem of operational forecast of dangerous convective phenomena (thunderstorm, heavy rain, hail) is one of the most relevant and practically significant.

O. Gervasi et al. (Eds.): ICCSA 2018, LNCS 10963, pp. 149–159, 2018.
https://doi.org/10.1007/978-3-319-95171-3_13

Variable specificity of convective clouds, caused by large vertical velocities within the cloud and its environment, and also the impossibility of carrying out control experiments lead to the fact that the greatest success can be achieved by computer research, which allows, without resorting to costly field experiments, to carry out an analysis of the development of the cloud. Forecast of dangerous convective phenomena is based upon the results of such an analysis.

Computer researches are based on numerical modeling. The construction of a numerical model consists of two stages: the first is the creation of a qualitative model, the second is the creation of a quantitative model. Creation of a qualitative cloud model implies formalization of the physical processes taking place in it and allows to reveal significant properties. As a result of the construction of a quantitative model, measurement scales and standards are established for each of these properties, which makes it possible to characterize the properties numerically.

Computer simulation provides a set of the output data that must be analyzed in order to build a forecast of dangerous convective phenomenon caused by the development of the cloud with such properties.

Methods of machine learning allow to automate the process of forecasting. The application of machine learning methods consists in carrying out a series of computational experiments, with the purpose of analyzing, interpreting and comparing the simulation results with the actual behavior of the object under study and, if necessary, the subsequent refinement of the input parameters.

Methods of machine learning implement the concept of data mining. This concept consists in processing large amounts of data and identifying on their basis various relationships and patterns. However, the data may be inaccurate, heterogeneous, inconsistent, contain omissions, which leads to incorrect forecasting. Therefore, an important step is feature selection, that is identification the most significant features among the data obtained.

The present paper is concerned mainly with the description of this important step of machine learning in case of preprocessing data for analyses of the results of numerical modeling of convective cloud.

2 Data Formation for the Research

As it is well known, the tasks of machine learning are reduced to the problem of finding an unknown relationship between a known set of objects and a set of answers [1, 2]. So, it is necessary to construct a function that would approximate sufficiently accurately the values of the set of responses at the points of the set of objects and on the rest of the space.

Everything can be considered as objects: web pages, countries, people, products, businesses, that is everything that carries any information (has a set of features). Features are understood as methods for measuring the characteristics of objects in the space under study.

Depending on the answers (values of the target variable), the tasks of machine learning are divided into types. The main types of machine learning tasks are:

- classification tasks;
- regression problems;
- ranking tasks.

Our task relates to the problems of regression.

The model of relationship between a known set of objects and a set of answers is called model of algorithms. The problem of finding the dependency model is reduced to constructing an algorithm that would equally accurately approximate the unknown target dependence, both on the sample elements and on the entire object space. This task was called the training with the teacher (supervised learning).

At the training stage, a training sample is used to identify the dependency, and optimization of the parameters is performed using it.

In our case training sample represents a set of radiosonde soundings obtained at a place and on a time when the dangerous convective phenomena take place. Radiosonde soundings were used as input data for one and a half convective cloud model [3–7]. Our training set consists of numerical parameters simulated by a numerical cloud model for each sounding, and is manually marked, that is, for each sounding from our set we know whether any dangerous convective phenomenon has been observed or not.

So the fact of dangerous phenomenon occurrence can be considered as an answer, and the results of numerical modeling, using as an input the corresponding radiosonde sounding, can be considered as an object.

The numerical parameters of the simulated clouds were chosen as an object features. There is a problem that should be discussed concerned with the time and height when and where the features are to be fixed. In the previous our works [8, 9] it was decided to fix the numerical parameters at the moment of maximum cloud development and at the height, where the maximum ratio of water droplets was observed. These time moment and height correspond to the mature stage of cloud development. But there are three stages in cloud evolution: stage of development, mature stage and dissipation stage. And it would be interesting to identify the influence of the nature of the evolution of the cloud parameters from stage to stage on the probability of a dangerous phenomenon with the help of machine learning methods. But at present there are no appropriate algorithms. So the only way out is to fix the cloud parameters not only at mature stage, but at the stages of development and dissipation also. Data preprocessing and subsequent analyses should be provided for the three sets of features and the best set from the point of the most accurate forecast should be chosen.

Training sample represents a set of radiosonde soundings obtained at a place and on a time when the dangerous convective phenomena take place were obtained with the help of integrated information system [10–15], which allow to integrate information about the dates and types of different convective phenomena and about vertical distributions of temperature and relative humidity observed on these dates and places.

3 Data Preprocessing

Data preprocessing is an important stage in machine learning. The use of qualitatively prepared data increases the accuracy of predictions, even with simple models.

At the first stage of preparation, it is necessary to transform data specific for the subject domain into understandable vectors for the model. For these purposes, an algorithm was developed and implemented in the program code for converting the output data of a numerical model to a format suitable for subsequent processing (the columns correspond to the characteristics, each line to a sounding uniquely determined by the values of the signs of time and height).

The next stage of data preprocessing is the adaptation of the data set to the requirements of the algorithm. The data has been subjected to normalization in view of the fact that most of the gradient methods that underlie almost all the algorithms of machine learning are highly sensitive to data scaling.

As a result of preprocessing, the data were brought to a form convenient for further work with machine learning methods.

The main statistical characteristics of the numerical data (the number of unallocated values, mean, standard deviation, range, median, 0.25 and 0.75 quartiles) are shown in Fig. 1. Analyzing these data, we can conclude that we have a complete set of data (the number of records is the same for each column, which indicates the absence of omissions in the data, their completeness).

The mean values of numerical features in the data with and without the phenomenon are presented in Table 1.

	velocity	velocityU	temperature	deltaT	relativeH	vapor	pressure	density	aerosol
count	6.150000e+02	615.000000	615.000000	615.000000	615.000000	6.150000e+02	615.000000	615.000000	6.150000e+02
mean	9.015934e+00	0.712901	263.406597	2.463016	0.714923	3.172778e-03	58969.729092	0.780044	2.729818e-08
std	1.012207e+01	7.603501	9.732534	3.371144	0.321774	2.817081e-03	2553.681444	0.032678	4.986370e-08
min	-2.607130e+00	-28.274580	183.397840	-2.372689	0.007335	1.550653e-08	4920.701500	0.093471	0.000000e+00
25%	-7.412999e-07	-1.061997	257.285400	-0.158336	0.397372	6.999553e-04	58412.269000	0.769386	0.000000e+00
50%	1.059566e+00	-0.000169	263.817430	0.914849	0.902727	2.333101e-03	59073.761000	0.780448	0.000000e+00
75%	1.854551e+01	0.001944	270.550375	4.552562	0.998569	5.167988e-03	59853.462500	0.791151	5.672878e-08
max	3.053718e+01	93.666761	282.485530	17.292512	1.017904	1.189431e-02	74666.560000	0.960829	1.052166e-06

Fig. 1. Main statistical characteristics

From the Table 1 we can conclude that such features as temperature and pressure play a weak role in predicting the phenomenon.

Grouping of data depending on the target variable and output of statistical data allows displaying the number of unset values, average value, standard deviation, range and median separately for the sets of soundings with and without phenomena. A fragment of the statistical data grouped by the value of the target variable is presented in Table 2.

Table 1. Mean values of numerical features

Without phenomenon		With phenomenon	
velocity	-1.042726e-02	velocity	1.701783e+01
velocityU	-3.687637e-02	velocityU	1.377581e+00
temperature	2.570544e+02	temperature	2.690378e+02
deltaT	-2.149329e-01	deltaT	4.837027e+00
relativeH	4.131247e-01	relativeH	9.824677e-01
vapor	8.592840e-04	vapor	5.223697e-03
pressure	5.826526e+04	pressure	5.959424e+04
density	7.891568e-01	density	7.719661e-01
aerosol	5.676505e-08	aerosol	1.175712e-09
drop	1.086072e-05	drop	5.998802e-03
ice	6.266864e-06	ice	6.216738e-05
hailAndGrits	2.029110e-05	hailAndGrits	2.728569e-04
targetV	0.000000e+00	targetV	1.000000e+00
dtype: float64		dtype: float64	

Table 2. A fragment of the statistical data grouped by the value of the target variable

		velocity	velocityU	temperature	deltaT	relativeH	vapor	pressure	density	aerosol
targetV										
0.0	count	2.890000e+02	289.000000	289.000000	289.000000	289.000000	2.890000e+02	289.000000	289.000000	2.890000e+02
	mean	-1.042726e-02	-0.036876	257.054432	-0.214933	0.413125	8.592840e-04	58265.261853	0.789157	5.676505e-08
	std	4.497304e-01	0.981230	7.834125	0.637566	0.205231	6.454670e-04	3304.279900	0.042477	5.995266e-08
	min	-2.412241e+00	-7.118333	183.397840	-2.372689	0.007335	1.550653e-08	4920.701500	0.093471	0.000000e+00
	50%	-8.715962e-07	-0.000002	257.616430	-0.174670	0.373010	6.995521e-04	58554.729000	0.790320	5.693065e-08
	max	6.425881e+00	9.186365	270.851100	5.369964	0.996721	4.672748e-03	61204.791000	0.834002	1.052166e-06
1.0	count	3.260000e+02	326.000000	326.000000	326.000000	326.000000	3.260000e+02	326.000000	326.000000	3.260000e+02
	mean	1.701783e+01	1.377581	269.037812	4.837027	0.982468	5.223697e-03	59594.241460	0.771966	1.175712e-09
	std	7.530381e+00	10.364664	7.541692	3.013280	0.074093	2.375928e-03	1346.268411	0.016710	7.459265e-09
	min	-2.607130e+00	-28.274580	235.463740	-0.427600	0.213483	1.940047e-04	54716.773000	0.740654	0.000000e+00
	50%	1.810377e+01	-0.958654	269.804885	4.262443	0.998146	4.973325e-03	59677.786500	0.770532	0.000000e+00
	max	3.053718e+01	93.666761	282.485530	17.292512	1.017904	1.189431e-02	74666.560000	0.960829	5.622077e-08

Figure 2 shows the graphs of the dependencies of various characteristics from each other (boxplot), the histograms of the distribution are on the diagonal.

A more detailed display of the relationship between two features: the deviation of temperature from the ambient temperature and the vertical component of wind speed is shown in the Fig. 3.

The matrix of correlation of numerical features is the form of the data matrix, which includes correlation coefficients for all pairs of analyzed variables. The correlation matrix is the basis for factor analysis, canonical correlation, and other statistical

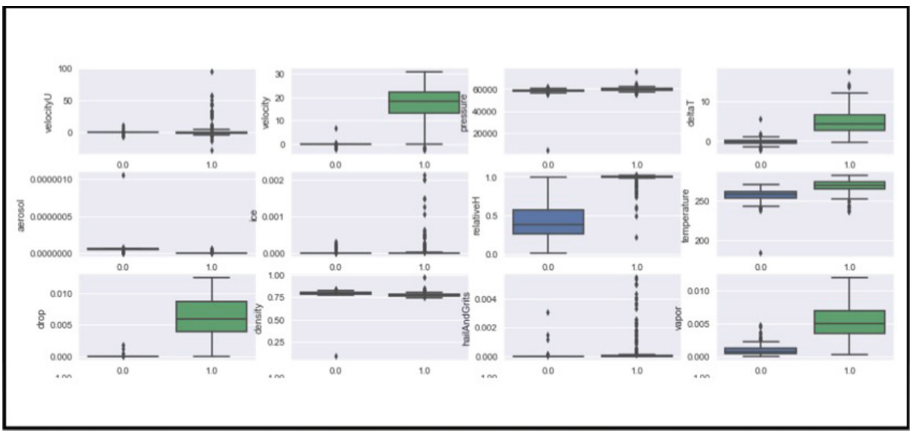

Fig. 2. Diagram of the range of feature values (boxplot)

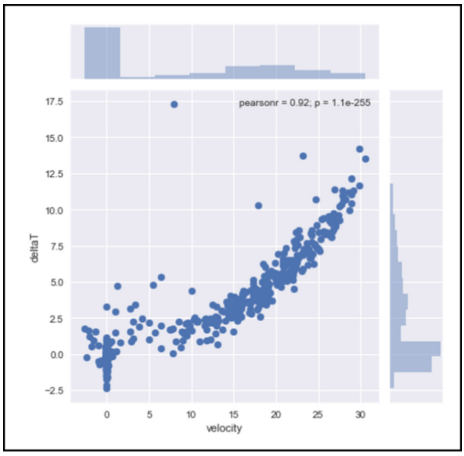

Fig. 3. The relationship between the feature of temperature excess from ambient temperature and the vertical component of velocity

techniques that reproduce the structure of the relationship between variables. A visual display of the correlation matrix of the characteristics used for the prediction of dangerous convective phenomena is given in Fig. 4.

When studying the features in a large group, different values of this characteristic are observed and occur unevenly a number of times: some more often, others less often. The distribution of the features from the minimum to the maximum is carried out, ordered when broken down into classes, that is, a variation series is constructed. The variation series are a double series of numbers consisting of the designation of classes and the corresponding frequencies.

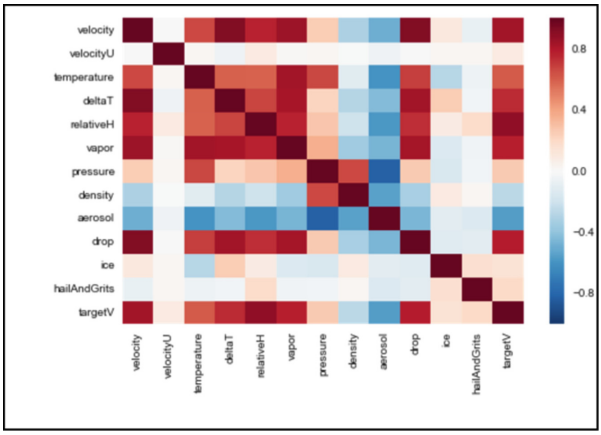

Fig. 4. Matrix of correlation of numerical features

The variation series includes all the primary material for the measurement of the feature in all representatives of the group. This material in the variation series is brought in a certain order, which makes it possible to characterize the sign, both at the average level of development, and on various details of diversity with an approximation that is quite sufficient for the first acquaintance with the feature.

For more detailed familiarization with the alignment of the characteristic, a variation curve is plotted graphically in the form of a curve whose ordinates are proportional to the frequencies of the variation series. The distribution of features can serve to identify a certain pattern. The norm of mass random manifestation of features, according to this, is called the normal distribution, which is usually hidden under the random form of its manifestation. The distributions of the characteristics are presented in the Fig. 2. A diagram of the scale (boxplot) is widely used to display the connection of characteristics with the target variable.

The boxplot is a limited area in the form of a rectangle (box), lines and points. The area bounded by the rectangle shows the inter quantile range of the distribution, that is, respectively 25% (Q1) and 75% (Q3) percentiles. The bar inside the rectangle indicates the median of the distribution. The lines that extend from the rectangle represent the entire scatter of points except the ejections, that is, the minimum and maximum values that fall within the gap.

$$(\mathrm{Q1} - 1.5 * \mathrm{IQR},\ \mathrm{Q3} + 1.5 * \mathrm{IQR}), \tag{3.1}$$

where IQR = Q3–Q1 is an inter quantile range.

Points on the graph indicate emissions, that represent those values that do not fit into the range of values specified by the lines of the graph. An example of a boxplot is shown in Fig. 5.

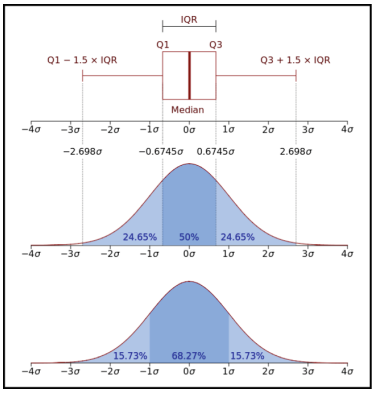

Fig. 5. Example of a boxplot

The result of plotting the boxplot for the features used in this paper is given in the Fig. 2. This kind of diagram in a convenient form shows the median, the lower and upper quartiles, the minimum and maximum value of the sample, and the outliers. Several such diagrams, constructed on the same plane, allow one to visually compare one distribution with another. Distances between different parts of the box allow you to determine the degree of dispersion (dispersion) and data asymmetry and to identify emissions.

Based on the results obtained, it can be concluded that the most interesting parameters are the vertical component of velocity, pressure, temperature deviation from ambient temperature, relative humidity.

One of the most important stages in the preparation of data is the selection of the most significant features. The reduction in the number of features (the rejection of features that are weakly correlated with the target variable) not only increases the accuracy of the prediction, but also lowers the requirements for the computing resources used.

There are various methods for feature selection, they can be divided into three groups:

- methods of filtration;
- methods for selecting the best subset;
- built-in methods.

The filtration methods are based on a statistical approach and consider the effect of each feature on the prediction error independently.

The Information gain method is one of the filtering methods. The IG (Information gain) parameter indicates the degree of correlation between the characteristic and the target variable. Thus, the method allows you to rank the characteristics by significance, degree of correlation with the target variable.

The degree of correlation of features with the target variable was represented using the matrix in the Fig. 4. According to this matrix, we can conclude that the most significant features are:

- vertical component of speed;
- temperature deviation from ambient temperature;
- relative humidity (above the water surface);
- the mixing ratio of water vapour;
- total droplet mixing ratio.

Filtering methods have low computational costs and work reliably on training sets, where the number of features exceeds the number of examples - these characteristics are advantages of this group of methods. However, an essential drawback is to work with each feature independently, because such an approach does not allow us to determine a subset on which the prediction accuracy will be the highest.

Methods for determining the best subset of characteristics consist in starting the classifier on different subsets and selecting a subset with the best parameters on the training sample. In turn, the methods of this group can be divided into inclusion methods and methods of exclusion. In the first case, the method starts with an empty subset, then at each step the optimal attribute is selected, in the second case, the initial subset is equal to the original set of characteristics. The work of the method consists in excluding the feature at each step with the reclassification of the classifier.

Recursive Feature Elimination method from the library scikit-learn is an example of methods for the gradual elimination of features [16]. To use this method, the support vector method was chosen as the classifier. As a result, the following six parameters appeared to be optimal for using as forecasting indicators:

- vertical component of speed;
- temperature deviation from ambient temperature;
- relative humidity (above the water surface);
- the mixing ratio of water vapour;
- total droplet mixing ratio;
- vertical height of the cloud.

4 Conclusions

Detailed algorithm is presented for data pre-processing for selecting the most representative cloud parameters (features). As a result, six optimal parameters: vertical component of speed; temperature deviation from ambient temperature; relative humidity (above the water surface); the mixing ratio of water vapour; total droplet mixing ratio; vertical height of the cloud has been chosen as indicators for forecasting of dangerous convective phenomena (thunderstorm, heavy rain, hail). Feature selection has been provided by using recursive feature elimination algorithm with automatic tuning of the number of features selected with cross-validation. Cloud parameters have been fixed at mature stage of cloud development.

Future work will be connected with identification of the influence of the nature of the evolution of the cloud parameters from initial stage to dissipation stage on the probability of a dangerous phenomenon. All the collected data should be integrated to the previously developed integrated information system [10–15]. In future the system

should become a consistent part of the Virtual private supercomputer [17, 18] and will be organized similar to the systems presented in [19, 20], that will enable users to provide forecasts of the dangerous convective phenomena by themselves.

Acknowledgment. This research was sponsored by the Russian Foundation for Basic Research under the projects: № 16-07-01113.

References

1. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. SSS. Springer, New York (2009). https://doi.org/10.1007/978-0-387-84858-7. http://statweb.stanford.edu/~tibs/ElemStatLearn/
2. Mitchell, T.: Machine Learning. Springer, Berlin (2009)
3. Raba, N., Stankova, E., Ampilova, N.: On investigation of parallelization effectiveness with the help of multi-core processors. Procedia Comput. Sci. **1**(1), 2757–2762 (2010)
4. Raba, N., Stankova, E.: On the possibilities of multi-core processor use for real-time forecast of dangerous convective phenomena. In: Taniar, D., Gervasi, O., Murgante, B., Pardede, E., Apduhan, Bernady O. (eds.) ICCSA 2010. LNCS, vol. 6017, pp. 130–138. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-12165-4_11
5. Raba, N.O., Stankova, E.N.: On the problem of numerical modeling of dangerous convective phenomena: possibilities of real-time forecast with the help of multi-core processors. In: Murgante, B., Gervasi, O., Iglesias, A., Taniar, D., Apduhan, B.O. (eds.) ICCSA 2011. LNCS, vol. 6786, pp. 633–642. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21934-4_51
6. Raba, N.O., Stankova, E.N.: On the effectiveness of using the GPU for numerical solution of stochastic collection equation. In: Murgante, B., Misra, S., Carlini, M., Torre, C.M., Nguyen, H.-Q., et al. (eds.) ICCSA 2013. LNCS, vol. 7975, pp. 248–258. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39640-3_18
7. Raba, N., Stankova, E.: Research of influence of compensating descending flow on cloud's life cycle by means of 1.5-dimensional model with 2 cylinders. In: Proceedings of MGO, vol. 559, pp. 192–209 (2009). (in Russian)
8. Stankova, E.N., Grechko, I.A., Kachalkina, Y.N., Khvatkov, E.V.: Hybrid approach combining model-based method with the technology of machine learning for forecasting of dangerous weather phenomena. In: Gervasi, O., Murgante, B., Misra, S., Borruso, G., Torre, C.M., et al. (eds.) ICCSA 2017. LNCS, vol. 10408, pp. 495–504. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-62404-4_37
9. Stankova, E.N., Balakshiy, A.V., Petrov, D.A., Shorov, A.V., Korkhov, V.V.: Using technologies of OLAP and machine learning for validation of the numerical models of convective clouds. In: Gervasi, O., Murgante, B., Misra, S., Rocha, A.M.A.C., Torre, C., et al. (eds.) ICCSA 2016. LNCS, vol. 9788, pp. 463–472. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-42111-7_36
10. Petrov, D.A., Stankova, E.N.: Use of consolidation technology for meteorological data processing. In: Murgante, B., Misra, S., Rocha, A.A.C., Torre, C., Rocha, J.G., et al. (eds.) ICCSA 2014. LNCS, vol. 8579, pp. 440–451. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-09144-0_30

11. Petrov, D.A., Stankova, E.N.: Integrated information system for verification of the models of convective clouds. In: Gervasi, O., Murgante, B., Misra, S., Gavrilova, M.L., Rocha, A.M.A.C., et al. (eds.) ICCSA 2015. LNCS, vol. 9158, pp. 321–330. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-21410-8_25
12. Stankova, E.N., Petrov, D.A.: Complex information system for organization of the input data of models of convective clouds. Vestnik of Saint-Petersburg University. Series 10. Applied Mathematics. Computer Science. Control Processes. Issue 3, pp. 83–95 (2015). (in Russian)
13. Petrov, D.A., Stankova, E.N.: Use of consolidation technology for meteorological data processing. In: Murgante, B., Misra, S., Rocha, A.A.C., Torre, C., Rocha, J.G., et al. (eds.) ICCSA 2014. LNCS, vol. 8579, pp. 440–451. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-09144-0_30
14. Petrov, D.A., Stankova, E.N.: Integrated information system for verification of the models of convective clouds. In: Gervasi, O., Murgante, B., Misra, S., Gavrilova, M.L., Rocha, A.M.A.C., et al. (eds.) ICCSA 2015. LNCS, vol. 9158, pp. 321–330. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-21410-8_25
15. Petrov, D., Stankova, E.: Complex information system for organization of the input data of models of convective clouds Vestnik of Saint-Petersburg University. Series 10. Applied Mathematics. Computer Science. Control Processes. Issue 3. pp. 83–95 (2015). (in Russian)
16. Scikit-learn. Machine Learning in Python. http://scikit-learn.org/
17. Bogdanov, A., Degtyarev, A., Korkhov, V., Gaiduchok, V., Gankevich, I.: Virtual Supercomputer as basis of Scientific Computing, in series: Horizons in Computer Science Research. In: Clary, T.S. (eds.), vol. 11, pp. 159–198. Nova Science Publishers (2015). ISBN: 978-1-63482-499-6
18. Korkhov, V., Krefting, D., Kukla, T., Terstyanszky, G.Z., Caan, M., Olabarriaga, S.D.: Exploring workflow interoperability tools for neuroimaging data analysis. In: WORKS 2011 - Proceedings of the 6th Workshop on Workflows in Support of Large-Scale Science, Co-located with SC 2011, pp. 87–96 (2011). https://doi.org/10.1145/2110497.2110508
19. Kulabukhova, N., Bogdanov, A., Degtyarev, A.: Problem-solving environment for beam dynamics analysis in particle accelerators. In: Gervasi, O., Murgante, B., Misra, S., Borruso, G., Torre, C.M., et al. (eds.) ICCSA 2017. LNCS, vol. 10408, pp. 473–482. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-62404-4_35
20. Kulabukhova, N., Andrianov, S.N., Bogdanov, A., Degtyarev, A.: Simulation of space charge dynamics in high intensive beams on hybrid systems. In: Gervasi, O. et al. (eds.) Computational Science and its Applications – ICCSA 2016, vol. 9786, pp. 284–295. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-42085-1_22