UDC 004.85 DOI: 10.30929/1995-0519.2019.1.33-40

USE OF METHODS OF MACHINE LEARNING FOR PREDICTION OF THE DANGEROUS CONVECTIVE PHENOMENA BY MEANS OF A NUMERICAL MODEL OF A CONVECTIVE CLOUD

O. Prykhodko
Lanzhou Jiaotong University
vul. West Road, 88, Lanzhou city, Gansu province 730070, People republic of China. E-mail: shidji@ukr.net

Purpose. On the basis of methods of machine learning to develop and implement an algorithm in a program code with use of model of a covenant cloud. Forecasting of the dangerous covenant phenomena will be the purpose of an algorithm. **Methodology**. Computational modeling is the cornerstone of a computer research. Creation of computer model consists of two stages: the first – creation of qualitative model, the second – creation of quantitative model. As a result of computer simulation at the exit there is a data set which needs to be analyzed for the purpose of creation of the forecast of temperature, rainfall, dynamics of air masses and the other atmospheric phenomena. Methods of machine learning allow automating process of creation of the forecast. Application of methods of machine learning consists in carrying out a series of computing experiments, for the purpose of the analysis, interpretation, and comparison of results of modeling to real behaviour of the studied object and, if necessary, the subsequent specification of input parameters. Methods of machine learning carry out the concept of intelligent data analysis. This concept consists of work with large volumes of data and identification on their basis of different interrelations and patterns – recovery of dependences according to empirical data. However, data can be inaccurate, diverse, contradictory, contain admissions that leads to incorrect forecasting. Therefore an important stage is an identification among these most significant signs. An intellectual component of methods of machine learning is the ability to study on "precedents" (on the basis of test selection), that is to draw conclusions on the basis of a set of private observations. Implementation of an algorithm was made on the basis of knowledge gained after processing of literature on the use of machine learning, literature and articles on the Internet on use of the Python programming language, documentation a free software machine learning library scikit-learn, information on a numerical model of a convective cloud, literature on the meteorological phenomena, their characteristics and methods of observation. **Results and Originality**. The algorithm based on use of operational data of radio sounding of the atmosphere is developed for prediction of the dangerous convective phenomena by means of methods of machine learning. The forecast is built on the basis of numerical signs which are created with use of a numerical model of a convective cloud, as a result of radio sounding data handling. **Practical value.** Thanks to development of such direction as business meteorology results of development can be used not only at scientific institutes and meteocenters, but also in case of market researches, in particular in case of SWOT analysis of the new directions of sale of goods and services, planning of commercial and production risks.

Key words: machine learning, SWOT analysis, algorithm, convective cloud, supervised learning, training sample.

ВИКОРИСТАННЯ МЕТОДІВ МАШИННОГО НАВЧАННЯ ДЛЯ ПРОГНОЗУВАННЯ НЕБЕЗПЕЧНИХ КОНВЕКТИВНИХ ЯВИЩ ЗА ДОПОМОГОЮ ЧИСЕЛЬНОЇ МОДЕЛІ КОНВЕКТИВНОЇ ХМАРИ

О. А. Приходько
Транспортний Університет Ланьчжоу
Аннін Си Лу 88, Ланьчжоу, Провінція Ганьсу 730070, Китайська народна республіка. E-mail: shidji@ukr.net

Глобальне потепління як результат постійного антропогенного впливу на атмосферу, призводить до зростання інтенсивності конвективних процесів. Збільшення температури і збільшення вологості повітря - два факти, які в сукупності призводять до інтенсифікації активної конвекції в атмосфері, що в свою чергу, тягне за собою збільшення числа сильних злив, зростання грозовий, зростання числа повітряних смерчів і збільшення ймовірності виникнення інших небезпечних конвективних явищ, які мають величезний руйнівний вплив. Тому проблема своєчасного передбачення небезпечних конвективних явищ - одне з найбільш актуальних і практично значущих напрямків досліджень. Передбачення метеорологічних процесів є складним завданням протягом всієї історії людства. Проблеми точності передбачення пов'язані з тим, що система атмосфери Землі дуже складна і динамічна. Прогнози погоди розраховуються на основі метеорологічних даних, зібраних мережею метеостанцій, радіозонда, радіолокаторами і супутниками по всьому світу. Дані надходять в метеорологічні центри, де їх вносять в прогностичні моделі і проводять розрахунок стану атмосфери. Прогностичні моделі засновані на фізичних законах. Кожна прогностична модель працює по вкрай складним алгоритмам, в яких закладені відомі закони поведінки атмосфери. В результаті комп'ютерного моделювання на виході є набір даних, який необхідно проаналізувати з метою побудови прогнозу температури, опадів, динаміки повітряних мас і інших атмосферних явищ. Методи машинного навчання здійснюють концепцію інтелектуального аналізу даних. Дана концепція полягає в роботі з великими обсягами даних і виявленні на їх основі різних взаємозв'язків і закономірностей - відновлення залежностей за емпіричними даними. Однак дані можуть бути неточними, різнорідними, суперечливими, містити пропуски, що призводить до невірного прогнозування. Тому важливим етапом є виявлення серед даних найбільш значущих ознак. Інтелектуальної складової методів машинного навчання є здатність навчатися за «прецедентів» (на основі тестової вибірки), тобто робити висновки на основі набору приватних спостережень.

Ключові слова: машинне навчання, SWOT аналіз, алгоритм, конвективна хмара, навчання з вчителем, навчальна вибірка.

PROBLEM STATEMENT. Clouds and associated precipitation play a decisive role in weather formation. By changing the thermal and radiation regime of the atmosphere, clouds have a huge impact on many aspects of human activity, on the flora and fauna of the Earth. First of all, a great influence can be traced in the sphere of agricultural production. In addition, the dependence of different modes of transport on clouds, fogs and precipitation is great. Transport-one of the most weather-dependent sectors of the economy. The impact on air and water transport is particularly evident, and the most complete, detailed information on both the actual observed weather and the forecast is required to ensure its normal operation.

Weather conditions have a significant impact on the economic performance of vehicles, on traffic safety. People's lives and health often depend on meteorological conditions and timely and qualitative information about them. Routes of air, sea vessels and road freight have a length measured by many hundreds and thousands of kilometers, so the specific requirements of transport to meteorological information is the scale of weather information.

According to the international civil aviation organization – ICAO (ICAO – International Civil Aviation Organization), the cause of 6 to 20% of accidents officially recognized adverse weather conditions; in addition, even more (one and a half times) the number of cases they are indirect or concomitant cause of such accidents. Thus, about a third of all cases unfavourable completion of flights, weather conditions play a direct or indirect role [1].

In addition, according to ICAO, violations of the flight schedule due to the weather, depending on the time of year and the climate of the area occur in an average of 1-5% of cases. More than half of these violations are flight cancellations due to adverse weather conditions at the airports of departure or destination.

In order to meet the needs of transport for meteorological information, it was necessary not only to create special meteorological services (aviation and maritime — everywhere, and in some countries also rail, road), but also to develop new branches of applied meteorology: aviation and marine meteorology.

Many atmospheric phenomena are dangerous for air and sea transport, and some meteorological values for the safety of modern aircraft and the navigation of modern ships must be measured with extreme precision.

The impact of transport needs on the development of meteorology in recent years has been decisive, it has led to the technical re-equipment of meteorological stations, and the use in meteorology of advances in radio engineering, electronics, telemechanics, as well as the improvement of weather forecasting methods, the introduction of tools and methods for pre-calculation of the future state of meteorological quantities (atmospheric pressure, wind, air temperature) and the calculation of the movement and evolution of synoptic objects.

Methods of observation of meteorological phenomena.

There are different ways to observe the weather: first there were networks of ground weather stations, observations in the waters of the oceans it was conducted on vessels and by means of a network of special buoys, from air data receive by means of radiosondes and satellites. Weather radars are now widely used, and their reflected signal can be used to restore clouds, precipitation, and wind [2].

Further, the results of measurements are received in the data centers. Then they come to the model through the data assimilation system, where the state of the atmosphere at each point of the planet is calculated.

To date, only 10 centers in the world have their own independent technology of global medium-term forecast.

The basis of numerical models of meteorological phenomena – equations of hydrodynamics. The equations are solved numerically using finite-difference schemes. There is a parameterization of processes (when observing the weather, these can be processes of exchange with the surface, radiation flows, chemical processes, processes of cloud formation and precipitation). Since parametrization is characterized by the assumption of some error in the calculations, with the increase of computer power for a more accurate approximation of the grid step is constantly decreasing, increasing the ability to solve processes directly, without parametrization.

The large amount of data obtained leads to the problem of finding methods that can work with so many dimensions. For example, the Russian hydrometeorological center processes "raw" data of more than 20 million numbers per day. Table 1 shows the amount of data on different types of observations that are input to the analysis scheme after selection, quality control, thinning and aggregation [3].

Table 1 – Number of Data, entering on an entrance of the scheme of the analysis

Type of observation	Number of data, numbers/day
Ground observations (SINOP, SHIP, BUOY)	140000
Aerological observations (TEMP)	60000
Plane observations about temperature and wind (AIREP, AMDAR)	800000
Scatterometric observations about drive wind (ASCAT)	15000
Observations of atmospheric occultation (air refractive index profiles, COSMIC, GRAS)	90000
Radiation observations AMSU-A (sensitive, mainly to the temperature in the thick layers of the atmosphere)	180000
Radiation observations MHS (humidity sensible)	80000

MATERIAL AND RESULTS. *A numerical model of a convective cloud.* Clouds are one of the most significant factors of formation of weather. They significantly influence distribution of energy resources, the lower inflow of radiation to the land surface, reduce

heat waste at the expense of radiation. Besides, clouds are a source of rainfall, thunderstorms and other adverse phenomena.

Classifications of numerical models.

One of the most efficient instruments of studying of convective clouds is numerical model operation. There is a set of models of development of the convective clouds differing both in dimension, and extent of specification of the description of microphysical processes.

Classification of models by dimension of space.

All set of numerical models of convective clouds on dimension of space is subdivided on:

- one-dimensional;
- one and a half dimensional;
- two-dimensional;
- three-dimensional.

There are ranges of application of numerical models of a cloud which have no huge computing resources and need carrying out operational forecasts. I.e. to them deep interrelations are not important, only reliable procreation of characteristics of a cloud which can become predictors for the existing methods of the forecast of the dangerous convective phenomena is required. One of ways of simplification is use of models with smaller dimension of space that allows to carry out calculations on ordinary personal computers. In [4] it is shown that by means of representations of small dimension it is possible to reveal a number of regularities of development. Such representations allow to receive existential distribution of dynamic and microphysical properties.

Optimal variant from the point of view of computing capacities and specification of procreation of cloudy processes are one and a half measuring model. In it the area of a cloud has the cylinder form. Unlike the one-dimensional models having only a vertical component of speed, this representation follow-up considers the radial component sent horizontally in the cylinder or to the external environment and also interaction with the external environment through a lateral area of the cylinder.

One and a half dimensional models are the good tool for authentic forecasts for the near-term period. However, at some data of radio intubation of the atmosphere (actual and model) such one and a half dimensional representation cannot reproduce adequately all stages of life of a cloud, in particular, a stage of a dissipation (scattering) [5].

Description of microphysical processes.

In the numerical model used in work the following processes are considered:

- condensation (P1);
- autoconversion (P2auto) and coagulation (P2coll), (P2 = P2auto + P2coll);
- freezing (heterogeneous freezing) (P3);
- sublimation (P4);
- melting (P5);
- evaporation of cloudy drops (P6);
- evaporation of rain drops (P7);
- evaporation of ice crystals (P8);
- evaporation of the thawing ice crystals (P9);
- grain formation (P10).

P1..., P10 — speeds of the corresponding processes.

In the fig. 1 the schematic description of the processes corresponding to a microphysical part of model is submitted.

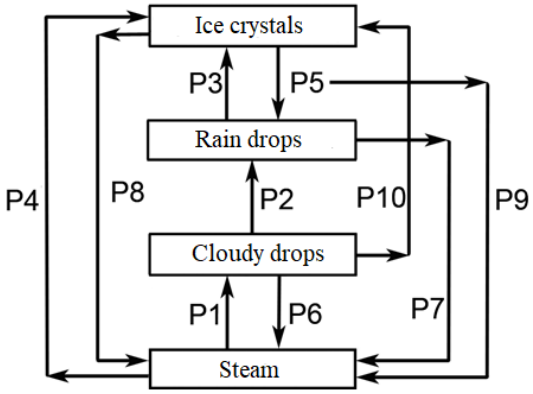

Figure 1 – Microphysical process of a numerical model of a convective cloud

The microphysical part of the model, taking into account the processes and their interrelations presented in Figure 1, is determined by the system by the following set of equations.

$$F_T = (L_v/c_p)(P_1 - P_6 - P_7 - P_9) + (L_s/c_p)(P_4 - P_8) + (L_f/c_p)(P_3 + P_{10} - P_5), \quad (1)$$

$$F_v = -P_1 + P_6 + P_7 + P_8 - P_4 + P_9, \quad (2)$$

$$F_c = P_1 - P_2 - P_6 - P_{10}, \quad (3)$$

$$F_r = P_2 - P_3 + P_5 - P_7, \quad (4)$$

$$F_i = P_3 + P_4 - P_5 - P_8 - P_9 + P_{10}, \quad (5)$$

where c_p – specific heat of air with a constant pressure; L_v, L_s and L_f – a latent heat of vaporization, sublimations and melting's respectively.

As a result, it is possible to draw conclusions:

1. The model with two cylinders gives more realistic picture of life cycle of a cloud where there are all reference stages of its development: (development, a maturity and dissipations) while the model with one cylinder reproduces only the two first stages of life, and further it is stabilized [6].

2. Generally, on the basis of these numerical calculations it is possible to tell that the radius of the external cylinder is less, the upper bound of a cloud is lower, the cloud lifetime, smaller values of the maximal vertical component of speed is less. It is caused by the fact that at the smaller radius of the external cylinder the downflow has high speed and stronger interferes with development of a cloud.

3. The coefficient defining a ratio of radiuses of external and internal cylinders can be used as the adjusting parameter for reduction of calculation data's in compliance with results of natural experiments.

Application of methods of machine learning.

Researches by means of methods of machine learning are based on data therefore for obtaining the most precise decisions, it is necessary to use reliable sources of information. Also, important role is played by forma-

tion of correct structure of data for after-treatment by means of methods of machine learning.

Selection of signs.

One of the most important stages of data preparation – selection of the most significant signs. Reduction of number of signs (a throw of the signs which are poorly correlating from a target variable), not only increases prediction accuracy, but also lowers requirements to the used computing resources.

There are different methods of selection of signs, they can be subdivided into three groups:

- filtering methods;
- methods of the choice of the best subset;
- built-in methods.

Methods of filtering are based on statistical approach and consider influence of each sign on a forecasting error independently.

The Information gain method (gain of information) – one of filtering methods. The IG parameter (Information gain) shows correlation degree between sign and a target variable. Thus, the method allows to range signs on the importance, correlation degrees from a target variable [7].

Degrees of correlation of signs from a target variable were presented by means of a matrix in last chapter. On this matrix it is possible to draw a conclusion that signs have the greatest importance:

- vertical component of speed;
- temperature deviation from ambient temperature;
- a relative humidity (over a water surface);
- relation of mix of steam;
- general relation of drops.

Variance Threshold – the approach consisting in an exception of signs which dispersion does not satisfy to some set threshold. By default, the algorithm deletes signs with zero dispersion, i.e. those which have identical values on all data set [8]. Results of execution of this method are reduced in table 2.

Table 2 – Results of selection of signs by means of the Variance Threshold method

Threshold value	Amount of the selected signs	The selected signs
0	12	Full range of signs
5	5	• vertical component of speed; • horizontal component of speed; • temperature; • temperature deviation from ambient temperature; • pressure
8	5	• vertical component of speed; • horizontal component of speed; • temperature; • temperature deviation from ambient temperature; • pressure

The signs having the most expressed interrelation with a target variable can be selected by means of statistical criterions. The scikit-learn library contains the class SelectKBest implementing one-dimensional selection of signs (univariate feature selection). This class can be applied together with different statistical criterions to selection of the set number of signs.

Results of selection of signs by means of this method are given in table 3.

Table 3 – Application of the SelectKBest method

Criteria for evaluation of quality	f_classif	Mutual_info_classif
Signs	• vertical making speeds; • temperature deviation from temperature environment; • relative humidity; • content is couple; • maintenance of cloudy drops.	• vertical making speeds; • temperature deviation from temperature environment; • relative humidity; • mix of aerosols; • maintenance of cloudy drops.

Methods of definition of the best subset of signs consist in start of the qualifier on different subsets and the choice of a subset with the best parameters on the training selection. In turn, methods of this group can be subdivided into methods of inclusion and methods of an exception. In the first case the method begins work with an empty subset, then on each step the optimum sign, in the second – an initial subset to equally initial feature set is selected. Work of a method consists in a sign exception on each step with recalculation of the qualifier.

Example of methods of a gradual exception of signs – Recursive Feature Elimination from scikit-learn library. For use of this method as the qualifier the support vector machine was selected [9] (tabl. 4).

Table 4 – Application of the Recursive Feature Elimination method

Algorithm	RFE	RFECV
Signs	• vertical making speeds; • temperature deviation from temperature environment; • relative humidity; • content is couple; • maintenance of cloudy drops.	• vertical making speeds; • temperature deviation from temperature environment; • relative humidity; • mix of aerosols; • maintenance of cloudy drops.

Proceeding from data retrieveds, the algorithm of REFS was started for selection of 3, 4 and 6 signs.

As a result of the choice of three signs:

- temperature deviation from ambient temperature;
- relative humidity;

• density.

As a result of the choice of four signs:

• temperature deviation from ambient temperature;
• relative humidity;
• density;
• relation of mix of steam.

As a result of the choice of six signs:

• vertical component of speed;
• temperature deviation from ambient temperature;
• relative humidity;
• density;
• relation of mix of steam;
• general relation of drops.

Formation of data for a research.

To use machine learning methods to predict dangerous convective phenomena, a sample of probing with and without phenomenon was formed. These selections were received by means of simulation of a convective cloud. Basic data for simulation were obtained by means of an information system of integration of meteorological information [10].

Probing, the formations of input data received by means of an end-to-end information system, come to model where on their basis simulation of a convective cloud by means of a one and a half dimensional nonstationary numerical model is made. At the exit of model data in the CSV format (Comma-Separated Values) for each probe in the following format are obtained: time, height, name of parameter, parameter value. The following parameters were removed:

• vertical component of speed;
• horizontal component of speed;
• temperature;
• the relative humidity (over a water surface);
• pressure;
• density;
• temperature deviation from ambient temperature;
• general relation of a compound of aerosols (thaw, ice particles, grain, hailstones);
• vertical power of a cloud.

The example of a fragment of the output file of model is given in a figure 2.

Time	Height	Name of parameter	Parameter value
1160	4050	velocity	13.067016
1160	4050	velocityU	9.343779
1160	4050	temperature	272.22715
1160	4050	relativeHumidity	1.0090644
1160	4050	vapor	0.0058138833
1160	4050	pressure	61652.788
1160	4050	density	0.7889759
1160	4050	aerosol	0.134232
1160	4050	drop	0.002514255
1160	4050	iceHailAndGrits	2.2671957E-0006
1160	4200	velocity	11.32452
1160	4200	velocityU	39.735083

Figure 2 – Application of methods of machine learning

Applying machine learning techniques.

For prediction of the dangerous convective phenomena the following methods of machine learning were used:

• the linear regression;
• ridge regression;
• ridge regression with the sliding monitoring;
• Lasso method;
• the Lasso method with the sliding monitoring;
• method of stochastic graded-index descent;
• linear support vector machines.

The linear regression.

The linear regression – a method of restoration of the linear dependence. Regression analysis implies statistical investigation of influence of one or several independent variables (predictors, signs) on a dependent (target) variable.

The target variable y in this case is represented as a linear combination of predictors $(x_1, x_2, \dots, x_n)$.

$$y(w, x) = w_0 + w_1 x_1 + \cdots + w_p x_p, \qquad (6)$$

where w_p – regression coefficients.

Linear regression adjusts the linear model by using regression coefficients to minimize the residual sum of squares between the dataset responses and the responses predicted by the linear approximation.

$$min_v \|x_w - y\|^2. \qquad (7)$$

The least squares estimation is very sensitive to random errors, so it is important that the input features are independent (have a weak correlation).

Fig. 3 shows the dependence of the target variable Y (abscissa axis) on the predictors X (ordinate axis). In the figure, the points indicate the values of the sample, the line reflects the regression dependence.

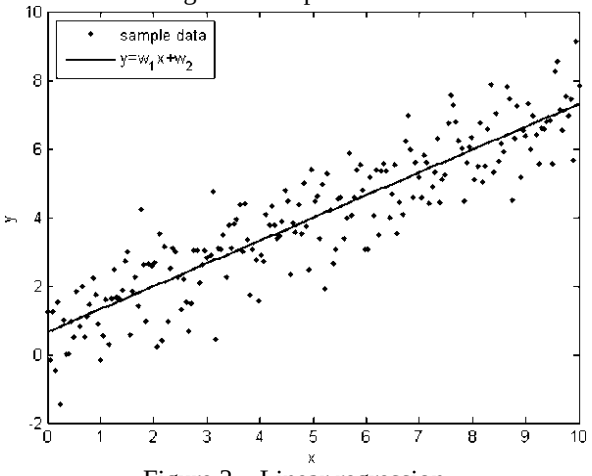

Figure 3 – Linear regression

The results of linear regression are presented in table 5.

Table 5 – Results of linear regression

Sampling	Prediction accuracy, %
Test (full)	90,08
Control (full)	89,03
Test (selected characteristics)	87,56
Control (with selected features)	87,34

Decision function is described by the equation:

$$f(\bar{x}) = 0.0243521840x_1 - 0.0163660081x_2 + 0.651072304x_3 - 4.46247383x_4 + 85.2249114x_5 + 0.00083849x_6 - 0.208302049813, \quad (8)$$

where $x = (x_1;\ x_2;\ x_3;\ x_4;\ x_5;\ x_6)$ the vector of signs consisting of the normalized values of parameters of a cloud. Here x_1 is equal to value of a vertical component of speed, x_2 – temperature deviation from ambient temperature, x_3 – relative humidity, x_4 – the relation of mix of steam, x_5 – the general relation of drops, x_6 – the vertical power of a cloud.

The rule of classification is defined as $G(x) = sign(x)$, by that is in case by $f(x) \geq 0$ – the dangerous convective phenomenon is observed, and at $f(x) < 0$ – will not be.

Ridge regression.

Ridge regression is one of methods of lowering of dimension. This method is improvement of a method of a linear regression, solving its main problem – instability of estimates of coefficients with reredundancy in data when independent variables correlate with each other (i.e. the multicollinearity takes place) (tabl. 6). This problem is solved with the help of imposing of the penalty depending on the size of coefficients that is the problem is solved:

$$min_w \|x_w - y\|^2 + a\|w\|^2, \quad (9)$$

where a – the complex parameter controlling penalty value (than more a , especially steady against collinearity are coefficients).

Table 6 – Results of ridge regression

Sampling	Prediction accuracy, %
Test (full)	84,94
Control (full)	83,91
Test (selected characteristics)	86,48
Control (with selected features)	86,07

Decision function is described by the equation:

$$f(\bar{x}) = 0.018369x_1 - 0.013384x_2 + 0.705794x_3 - 0.013036x_4 + 0.025917x_5 - 0.00107x_6 - 0.22714, \quad (10)$$

where $x = (x_1;\ x_2;\ x_3;\ x_4;\ x_5;\ x_6)$ the vector of signs consisting of the normalized values of parameters of a cloud. The feature set and the rule of classification are similar to a linear regression.

Lasso.

Lasso (Least Absolute Shrinkage and Selection Operator) – it is a method of lowering of dimension. The method consists in imposition of the restriction on norm of a vector of coefficients of model that leads to the address to 0 some coefficients of model. The method leads to increase in stability of model in case of a large number of conditionality of a matrix of signs of X, allows to receive the interpreted models – the signs having the greatest impact on a vector of answers are selected (tabl. 7).

$$min_w \frac{1}{2n_{samples}} \|x_w - y\|_2^2 + a\|w\|_1, \quad (11)$$

where a – the complex parameter controlling penalty value (than more a , especially steady against collinearity are coefficients); $\|w\| - l_1$- norm of a vector of parameters.

Table 7 – Results of the Lasso method

Sampling	Prediction accuracy, %
Test (full)	73,11
Control (full)	69,13
Test (selected characteristics)	77,97
Control (with selected features)	76,39

Decision function is described by the equation:

$$f(\bar{x}) = 0.022395x_1 + 0.002004x_6 - 0.099006, \quad (12)$$

where $x = (x_1;\ x_2;\ x_3;\ x_4;\ x_5;\ x_6)$ the vector of signs consisting of the normalized values of parameters of a cloud. The feature set and the rule of classification are similar to a linear regression.

Logistic regression.

Logistic regression, despite its name, is linear model of classification, but not regression. In this model, the probabilities describing possible results of one testing are modelled with use of logistic function. Logistic regression – an algorithm which belongs to group of statistical qualifiers. The difference of logistic regression is that prediction is based not on the basis of original values of signs, and on function value of probability of whether this original value is a member of a target class.

The main idea of logistic regression is that the space of original values can be divided by linear border (i.e. a straight line) into two areas corresponding to classes. In case of two measurements – it is just direct line without bends. In case of three – the plane, and so on. This border is set depending on the available basic data and the training algorithm (tabl. 8).

Table 8 – Results of logistic regression

Sampling	Prediction accuracy, %
Test (full)	94,54
Control (full)	94,05
Test (selected characteristics)	98,84
Control (with selected features)	97,84

Decision function is described by the equation:

$$f(\bar{x}) = 0.697294x_1 + 0.637496x_2 - 1.013573x_3 - 0.018674x_4 + 0.003274x_5 + 0.00314x_6 - 3.081611, \quad (13)$$

where $x = (x_1;\ x_2;\ x_3;\ x_4;\ x_5;\ x_6)$ the vector of signs consisting of the normalized values of parameters of a cloud. The feature set is similar to a linear regression. Probability that there will be a phenomenon at the set vector of signs $P_r\left(y = \frac{1}{x}\right) = f(z) = \frac{1}{1+e^{-z}}$.

Respectively, than it is more, especially emergence of the dangerous convective phenomenon is probable. Probability of an opposite event, that is the fact that the dangerous convective phenomenon will equally not be

observed$1 - f(z)$, respectively if $f(z) \geq 0{:}5$, then the phenomenon will be observed and at $f(z) < 0{:}5$ - will not be.

Stochastic gradient descent.

Stochastic gradient descent is a simple but very effective approach to linear models. The gradient approach is used as a method for selecting the weight vector w in the linear classifier. The stochastic approach implies the choice of only one object at each iteration of the algorithm. Those. The weight vector is adjusted to a new object each time.

The method of stochastic gradient descent is to change the weights vector at each iteration of the algorithm in the direction of the greatest decrease of the functional (tabl. 9).

Table 9 – Results of stochastic gradient descent

Sampling	Prediction accuracy, %
Test (full)	47,44
Control (full)	45,94
Test (selected characteristics)	97,67
Control (with selected features)	96,22

Decision function is described by the equation:

$$f(\bar{x}) = 1583.45x_1 + 496.86x_2 - 198.27x_3 - 0.29x_4 - 0.01x_5 - 2.14x_6 - 984.89, \quad (14)$$

where $x = (x_1;\ x_2;\ x_3;\ x_4;\ x_5;\ x_6)$ is a feature vector consisting of normalized values of cloud parameters. The feature set is similar to linear regression. The probability that there will be a phenomenon for a given feature vector corresponds to logistic regression.

Linear Support Vector Machines.

Linear Support Vector Machines – the controlled method of training solving problems of classification and regression.

The support vector machine belongs to family of linear classifiers and is one of the most popular methods of machine learning. The key idea of a method consists in translation of initial vectors in space of higher dimension and search of the dividing hyperplane with the maximum gap in this space. We will call reference vectors a vector the next to an opposite class (that is lying on class border). Two parallel hyperplanes are under construction through reference a vector. The hyperplane maximizing distance to two parallel hyperplanes will be the dividing hyperplane. The algorithm works in the assumption that the difference or distance between these parallel hyperplanes is more, the average error of the qualifier will be that less (tabl. 10).

Table 10 – Results of the support vector machine

Sampling	Prediction accuracy, %
Test (full)	47,44
Control (full)	45,95
Test (selected characteristics)	98,84
Control (with selected features)	98,38

Decision function is described by the equation:

$$f(\bar{x}) = 0.152699x_1 + 0.004444x_2 - 0.13214x_3 - 0.000293x_4 + 0.000192x_5 + 0.00128x_6 - 0.965603, \quad (15)$$

where $x = (x_1;\ x_2;\ x_3;\ x_4;\ x_5;\ x_6)$ feature vector consisting of normalized values of cloud parameters. The feature set is similar to linear regression. The classification rule is defined as $G(x) = sign(x)$, that is, in the case when $f(x) \geq 0$, a dangerous convective phenomenon will be observed, and with $f(x) < 0$, there will be no.

The most accurate predictions were shown by the methods:

- linear support vector machines.;
- stochastic gradient descent;
- logistic regression.

CONCLUSIONS**.** As a result of application of methods of machine learning series of calculations for definition of set of optimum signs were carried out. Methods were for this purpose used: Variance Threshold, SelectKBest(f_classif), SelectKBest(mutual_info), Recursive Feature Selection, Tree Based Feature Selection. It was succeeded to define optimum amount of signs for creation of forecasts by these methods and to select signs. In total the received results, optimum it appeared to use set of six signs: a vertical component of speed, temperature deviation from ambient temperature, a relative humidity (over a surface of the water), the relation of mix of steam, the general relation of drops, vertical power of a cloud.

The created feature set served as input data of methods of machine learning. In this research methods were used:

- linear regression;
- ridge regression;
- ridge regression with a sliding control;
- lasso method;
- method the Lasso with a sliding control;
- logistic regression method;
- stochastic gradient descent method;
- support vector machine.

The methods showed the greatest accuracy:

- logistic regression method;
- stochastic gradient descent method;
- support vector machine.

Combining all the results of the work done, we can give an algorithm for the practical application of the numerical model to predict the dangerous convective phenomenon.

1. Modified probes are used as input for the numerical cloud model.
2. The cloud is modeled and the necessary numerical parameters are selected at the right time and at the right height.
3. The selected parameters are used as features to calculate the value of the decision function.
4. The conclusion is drawn from the results of the forecast of one or more decisive functions. For example, it can be concluded that there will be a dangerous convective phenomenon if at least two of the three decisive functions have determined that it will be observed.

In the future, it is planned to implement a double classification: first to determine whether there will be a phenomenon or not, and then if there will be a phenomenon, then what kind.

REFERENCES

1. Astapenko, P. I. Voprosy o pogode [Questions about the weather], [Online], available at: https://unotices.com/book.php (Accessed: 17.03.2017).

2. Tunegolovecz, V. P. (2002), Lekcii po navigacionnoj gidrometeorologii, [Lectures on navigational hydrometeorology], [Online], available at: https://www.twirpx.com/file/1410694/ (Accessed: 10.06.2017).

3. Vil'fand, R. M. (2014), Tekhnologii meteorologicheskogo prognozirovaniya v Rossijskoj Federacii: sostoyanie i perspektivy [Technologies of meteorological forecasting in Russian Federation: status and prospects].

4. Asai, T., Kasahara, A. (1997), "A theoretical study of the compensating downward motions associated with cumulus clouds", *J. Atm. Sci*., V. 24, p. 487–497.

5. Raba, N. O., Stankova, E. N. (2009), "Issledovanie vliyaniya kompensiruyushchego niskhodyashchego potoka, soputstvuyushchego konvektivnym techeniyam, na zhiznennyj cikl oblaka s pomoshch'yu polutoromernoj modeli s dvumya cilindrami", [Investigation of the effect of compensating downward flow on the life cycle of a convective cloud using a numerical one-and-a-half-dimensional model with two cylinders], *Proceedings of the Main Geophysical Observatory. A. I. Voeikov*, no. 559, pp. 192-209.

6. Stankova, E. N., Raba, N. O., Ampilova, N. B. (2008), Chislennoe modelirovanie mikrofizicheskih processov v smeshannyh konvektivnyh oblakah. Sravnenie rezul'tatov raschètov s dannymi naturnyh ehksperimentov, [Numerical simulation of microphysical processes in mixed convective clouds. Comparison of calculation results with the data of field experiments], *Scientific conference of Roshydromet institutes "Theoretical and experimental studies of convective clouds*", pp. 90 – 91.

7. Shai, Shalev-Shwartz, Shai, Ben-David (2014), *Understanding Machine Learning*. p. 48 – 65.

8. Goodfellow, I., Bengio, Y., Courville, A. (2016), *Deep Learning*. p. 120 – 138.

9. Scikit-learn Machine Learning in Python. [Online], available at: http://scikit-learn.org. (Accessed: 15.09.2017).

10. Arhiv pogody [Weather archive], [Online], available at: http://meteocenter.net/ussr_fact.htm. (Accessed: 24.07.2017).

ИСПОЛЬЗОВАНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ ОПАСНЫХ КОНВЕКТИВНЫХ ЯВЛЕНИЙ С ПОМОЩЬЮ ЧИСЛЕННОЙ МОДЕЛИ КОНВЕКТИВНОГО ОБЛАКА

О. А. Приходько

Транспортный университет Ланьчжоу

Аннин Си Лу 88, Ланьчжоу, провинция Ганьсу 730070, Китайская народная республика.

E-mail: shidji@ukr.net

Работа содержит обзор методов машинного обучения и сфер их применения, в частности приводится обзор существующих решений, использующих технологии машинного обучения в метеорологии. Так же работа включает в себя описание численной модели конвективного облака. Разработанный алгоритм основан на использовании оперативных данных радиозондирования атмосферы для прогноза опасных конвективных явлений с помощью методов машинного обучения. Прогноз строится на основе многочисленных признаков, которые формируются с использованием численной модели конвективной облака, в результате обработки данных радиозондирования. Благодаря развитию такого направления как бизнес-метеорология результаты разработки могут быть использованы не только в научных институтах и метеоцентрах, но и при маркетинговых исследованиях, в частности при SWOT-анализе новых направлений сбыта товаров и услуг, планировании коммерческих и производственных рисков.

Ключевые слова: машинное обучение, SWOT анализ, алгоритм, конвективное облако, обучение с учителем, учебная выборка.

Стаття надійшла 01.02.2019.