The European Proceedings of
Social & Behavioural Sciences
EpSBS

Future Academy

ISSN: 2357-1330

https://dx.doi.org/10.15405/epsbs.2018.11.02.21

ICPE 2018
International Conference on Psychology and Education

INTELLECTUAL CLASSIFICATION OF THE IT PROJECTS DURATION IN THE EDUCATIONAL SPHERE

Liliya Demidova (a), Irina Klyueva (b)*, Roman Tishkin (c)
*Corresponding author

(a) Moscow, Russia, Moscow Technological Institute, Ryazan, Russia, Ryazan State Radio Engineering University, demidova.liliya@gmail.com
(b) Ryazan, Russia, Ryazan State Radio Engineering University, i.klyueva-job@yandex.ru
(c) Moscow, Russia, Limited Liability Company "Areal-98", rvt@areal98.ru

Abstract

This article discusses the aspects of the IT technologies application in the educational sphere. The implementation of the IT project in the education, as in any other sphere of activity, is a complex science-intensive process that requires the certain time costs for implementation. In solving the problems of the duration estimating of the new IT projects under uncertainty, there is often a need to use the methods of statistical data analysis. Therefore, we propose to use the hybrid intelligent classification technology to the duration assessment of the new IT projects based on the hybridization of the SVM algorithm with the intelligent classification algorithms such as the random forest algorithm and the *k* nearest neighbor algorithm. The hybridization of the SVM classifier with the intelligent classification algorithms allows to obtain the sufficiently high values of the classification quality indicators with the less time expenditures than in the alternative approaches to application of the SVM classifier based on the search of its parameters values using, for example, the grid search algorithms or the evolutionary algorithms. The offered hybrid intelligent classification technology was tested using the different datasets and can be recommended for classification of the new objects, in particular, to predict the implementation duration of the new IT projects in the educational sphere. The enlarged scheme of the process of developing of the intelligent data analysis system and predicting of the new projects duration has been presented.

Keywords: Intelligent classification algorithm, duration assessment of the IT project, SVM algorithm, *k* nearest neighbor algorithm, random forest algorithm.

https://dx.doi.org/10.15405/epsbs.2018.11.02.21
Corresponding Author: Irina Klyueva
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

1. Introduction

Currently, the process of integration of information technologies (IT) in the educational sphere is actively developing.

In the learning process, various educational technologies, including the remote and electronic technologies, should be used. This is a step forward in providing equal opportunities for children studying in urban and rural schools. It allows to preserve the small-scale schools, allows to study according to the individual in-depth program to the gifted children, allows to master the school curriculum and even get a professional education to the children with disabilities.

The use of IT allows to improve the teaching methods. It is possible to use the methods of active and distance learning, to integrate the webinars of companies (developers and vendors) in the educational process, to organize various types of scientific events with the involvement of experts, including remotely.

At the world level, the educational platforms such as Coursera and TED, publishing the educational materials in the Internet in the form of a set of online courses and video lessons, have long been popular.

The project "Lectoryum" was launched by the teachers from St. Petersburg on the basis of the idea of the American service “Academicearth”. First, the "Lecterium" was the media library of the best lectures of the Moscow and St. Petersburg universities. Then, the "Lectoryum" went on to produce and record lectures of its choice and began to attract not only universities, but also museums and corporations.

The educational institutions are very interested and are always ready to use cloud computing, social networks, mobile technologies offered by IT companies, but on a free basis.

Therefore, it is necessary to support educational projects with grants.

Despite the existing budgetary constraints (with the allocation of grants), the universities are already using the advanced informational technologies. For example, the Moscow Energy Institute (MEI) uses the cloud technologies such as the MathCAD server, mobile devices and applications, in particular the Wi-Fi access from any device to the real data from the MEI TPP. Also, the Linux-based center has been created to use the open source software, and there are many MEI groups in the social networks.

There are examples of the social networks created on the initiative of the Russian universities. The Novgorod State University named by Yaroslav Wise completed the project to introduce the social network of the university, uniting fifteen thousand students and three thousand teachers. The introduction of new functions of the social services - communities, forums, tasks, bookmarks – allow to receive the access to knowledge bases, effectively interact with colleagues and conduct joint educational and research projects to the students and university staff.

In the Pokrov branch of the Moscow State Mining University, the cloud technologies are used to create the social educational networks, for example, on the Bitrix24 platform, as well as to store and organize the data access, teaching to project management technologies with the help of the Megaplan SaaS system, and working with documents.

The implementation of the IT project in education, as in any other field of activity, is a complex science-intensive interaction that requires time-consuming implementation costs and is limited by many factors, including the choice of the platform, the resource support of the team, the limited budget, etc.

For example, the practical implementation of the project on the use of electronic and distance technologies to expand the access to the educational environment requires new computing power,

https://dx.doi.org/10.15405/epsbs.2018.11.02.21
Corresponding Author: Irina Klyueva
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

communication channels and applications, the purchase and provision of which can also affect the project implementation time.

IT project is a temporary enterprise aimed at developing a unique product related with information technologies (in particular, software development, information system). IT project is characterized by a certain period of implementation, cost, restrictions on resources. It has its own quality criteria and the notion of a successful completion.

A large number of the versatile requirements are shown to the IT projects quality. The basic from such requirements is need of implementation of the IT projects within minimum possible terms (duration) at achievement of high level of efficiency that is a hard-hitting task. Determination of the exact duration of the IT project is almost impracticable task. The reason of this is the need of introduction of the unplanned changes for the purpose of receiving of the results in the course of the IT project life cycle.

2. Problem Statement

In practice, there are often cases when the IT projects do not fit within the assigned time limits, that leads to exceeding the project budget, and, as a consequence, to the failure to fulfill the main stated requirements for the project. According to the Standish Group (USA) studies (CHAOS Report. 2014)], more than 30% of projects from more than 8,000 projects surveyed were failed, the total costs for them exceeded about $ 80 billion. A total of 16% of the projects were completed on time within the budget, for the rest projects the over expenditure was about 189%. The results of these studies show that the projects planning is often unsatisfactory, that leads either to a breakdown in the timing of the projects or to their complete cessation.

The effectiveness assessment of IT projects is the comprehensive assessment of the degree to which the IT projects meet the objectives, taking into account the costs and risks of the IT projects.

In view of the limited budget, it is necessary to make an informed decision on the implementation of the particular IT project.

In the traditional understanding of the effectiveness of the IT project, the cost-benefit ratio of the project is implied. The cost of the IT projects, in particular when implementing a new information system (IS), is understood as the total costs:

- the costs of acquiring licenses, installing, configuring and supporting software providing for the deployment of the new IS;
- the costs associated with the acquisition and support of the required IT equipment and other technical means;
- the labor costs of the project team (in the case if the project is implemented by third-party specialists), staff training, etc.

The results are understood as the effect that is achieved when implementing and further operating the software.

There are different approaches to assessing the effectiveness of the project.

The posteriori approach which combines the methods of the direct evaluation of the implementation results of information systems at the operation stage. These methods take into account various key factors

https://dx.doi.org/10.15405/epsbs.2018.11.02.21
Corresponding Author: Irina Klyueva
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

before and after the implementation of information system and the comparison of the result with the efforts spent on the implementation of the information system implementation project.

The a priori approach which combines the methods of estimating and forecasting of the implementation results of information systems s at the stage of choosing decisions and agreeing on the volume of investments.

These methods use the forecast values of the key factors, which are determined on the basis of the developed models. A herewith, various types of risks that affect both the efficiency and costs of the information system implementation project, as well as various implicit opportunities can be taken into account.

The most preferred methods for assessing the implementation effectiveness of the IT projects are methods that allow to evaluate the desired effectiveness of the IT project before its implementation, i.e. at the stage of the feasibility study.

These methods include the methods in the framework of the a priori approach: IRR (Internal Rate of Return), ROI (Return on Investment), TEI (Total Economic Impact), NPV (Net Present Value), BSC (Balanced Scorecard), EVA (Economic Value Added) and others (CHAOS Report, 2014). The key dignity of these methods is the assessment of the predicted effect of the IT project implementing.

Evaluating within aprioristic approach before the implementation of the project, will allow to answer a question of expediency of investment into this IT project.

Assessment of expediency of implementation of the project is carried out on the basis of comparison of expenses and effect (results) of its implementation.

One of important indicators of the investment attractiveness of the project is the duration of its implementation. Many characteristics of the project depend on this indicator, and in the calculation of characteristics. The duration indicator of the project can act as the main dependent variable. A herewith, the expenses, in turn, are a function of the duration of work. The duration of the project is interlinked with the risks of the project, the implementation of which may entail the termination or delay of the project.

The formula of calculation of the general expenses on the project can be presented in the following form:

$$Z = T(t) + S(t) + A(t) + O(t) + L(t) + K(t) + M(t) + P\,, \quad (1)$$

where Z is the total project costs; $T(t)$ is the costs for the salary of the project team; $S(t)$ is the contributions to social funds; $A(t)$ is the depreciation charges; $O(t)$ is the costs for maintenance and servicing of technical facilities (computers and peripherals); $L(t)$ is the energy costs; $K(t)$ is the costs for leasing or maintaining communication channels; $M(t)$ is the costs for consumables; P is the other expenses.

In practice of implementation of the IT projects, the initial calculation of the project duration is required at the stage of the requirements formation for introduction of the information systems. The initial evaluation of the IT project duration is carried out in the absence of complete information on the scope of the project, its functionality and complexity, since a full survey is to be performed at a later stage. Thus, it is required to assess the feasibility of the IT project and to decide on the continuation of the work based on almost zero information about this project.

https://dx.doi.org/10.15405/epsbs.2018.11.02.21
Corresponding Author: Irina Klyueva
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

At the stage that follows the initial assessment stage, the direct determination of volumes of the project on the volume of terms, requirements and project cost begins. Only at the design stage, the information that allows to reliably calculate the duration of the project, appears, and the project plan is formed.

Therefore, when solving the problems of the duration assessment of the new project, there is a need to use the statistical analysis methods.

3. Research Questions

The main research questions in this work are the following:

- the study of the problem of the IT projects efficiency in the educational sphere;
- the study of the approaches to the IT projects duration assessment;
- the development of the hybrid data classification technology based on the hybridization of the SVM algorithm with other intelligent classification algorithms such us the random forest algorithm and the k nearest neighbor algorithm. The results of the hybridization should ensure high values of the classification quality in the problem of the IT projects duration assessing;
- the development of the intellectual data classification system, which provides the process of developing of the intelligent data classifier and predicting of the new projects duration.

4. Purpose of the Study

The research objective is the study of the problem of the IT projects efficiency in the educational sphere and the development of the hybrid data classification technology to assess the duration of the new IT projects.

5. Research Methods

Let it is necessary to solve the problem of the duration assessment of the IT projects. At the duration assessment the new project is assigned to one of the classes formed on the basis of the statistical data on the previously completed projects. A herewith, the duration value (or the duration interval) is known for each class. Due to accumulation of statistical information on the actual duration of the executed projects the required accuracy in assessment of duration of new projects will be reached.

It is suggested to perform training based on the statistical information used to form the training dataset, and then predict the duration of new IT projects.

Let the dataset be a set in the form of $\{(z_1, y_1),...,(z_s, y_s)\}$, in which each object $z_i \in Z$ is assigned to a number $y_i \in Y = \{+1;\ -1\}$ having a value of +1 or −1 depending on the class of the object z_i ($i = \overline{1,s}$; s is the number of objects). It is assumed that the i-th object is mapped to q-dimensional vector $z_i = (z_i^1,...,z_i^q)$ of the numerical values of the characteristics, where z_i^l is the numeric value of the l-th characteristic for the i-th object ($i = \overline{1,s}$; $l = \overline{1,q}$).

https://dx.doi.org/10.15405/epsbs.2018.11.02.21
Corresponding Author: Irina Klyueva
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

A herewith, each object is understood as the completed IT project, characterized by a vector of characteristics $z_i = (z_i^1,...,z_i^q)$.

The belonging of the IT completed projects to the class is determined depending on the actual duration of the projects.

At the training stage, the completed IT projects are analyzed and the general properties of the projects having the similar duration values are studied, and also the distinctive characteristics of the projects with the different duration are determined.

As a result the characteristics influencing on the duration of the IT projects are chosen.

The characteristics whose values are available at the early stages of design are selected from the characteristics that affect the duration of the IT projects.

Then, the characteristics values of the IT projects are normalized and the development of classifier is carried out. This classifier can be use for classification of the new IT projects.

As the statistical data accumulates at the subsequent iterations of the classifier functioning, the classifier will go into the relearning mode.

It is proposed to divide the IT projects into two classes in the form: short-term projects – up to 1 year; long-term projects – lasting from 1 year.

In general, the characteristics of the IT projects can be divided into some groups, on the basis of which the classification of the IT projects is carried out.

- The characteristics of the IT project team: the level of the project manager, the the presence of the project executors (analysts, experts, programmers, etc.), the team coherence, the experience in the applied field, the general qualifications of the project executors.
- The characteristics of the customer: the branch area (public sector, medicine, education, banking sector, transport and communications, industry, defense industry, etc.), the experience of employees responsible for the product being introduced; the number of the qualified IT staff.
- The characteristics of the IT project: the cost, the complexity of the product, the type of the product (information system, software product, hardware, software and hardware, services), the volume of the automated business processes (main and auxiliary, technological and office, management, analytical, transaction, data transmission , storage organization, processing of media content, etc.), the degree of the required documentation, the similarity to the previously implemented projects, the required reliability and level of the security requirements, the level and the experience of the intended user, the number of the end users.

It is possible to include to the IT projects characteristics in the educational sphere the following: the coverage and distribution of the end users, the team skills and teamwork, the presence of the project executors, the complexity and type of the project product (training platform, electronic library, social network, etc.), the presence of the realized analogues-projects (world experience), the support of the state authorities, etc.

At the stage of decision-making at the start of investing in the project, in the absence of sufficient information about the characteristics of this project, the methods available for this stage to estimate the duration of the project (for example, expert assessments) give low accuracy. The offered intelligent data classification technology allows to evaluate the duration of development of new IT projects in the early

https://dx.doi.org/10.15405/epsbs.2018.11.02.21
Corresponding Author: Irina Klyueva
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

stages of their implementation on the basis of the hybrid of the SVM classifier (Vapnik, 1998; Demidova et al., 2017a; Demidova et al., 2017b) and the *k*NN classifier (*k* nearest neighbor classifier) (Demidova et al., 2017b) or the RF classifier (random forest classifier) (Breiman, 2001; Hastie, Tibshirani & Friedman, 2009; Pal, 2015). A herewith, when developing the proposed approach, the problem of imbalance in the initial datasets obtained with the help of expert assessments was investigated, and the approach to solve the problem of data imbalance by sampling was developed (Demidova et al., 2017a).

It is necessary to use the kernel function $\kappa(z_i, z_\tau)$ to develop SVM classifier $F: \ Z \to Y$, which assigns the class (the number from the set $Y = \{+1; \ -1\}$) to the object z_i from the set Z. The separating hyperplane for the objects from the training set can be represented by equation $< w, z > + b = 0$, where w is a vector-perpendicular to the separating hyperplane; b is a bias; $< w, z >$ is a scalar product of vectors w and z. The condition $-1 < < w, z > + b < 1$ specifies a strip that separates the classes. The wider the strip, the more confidently we can classify objects. The objects closest to the separating hyperplane, are exactly on the bounders of the strip.

Finding the separating hyperplane is basically the dual problem of searching a saddle point of the Lagrange function, which reduces to the problem of quadratic programming, containing only dual variables (Demidova et al., 2017b):

$$\begin{cases} -L(\lambda) = -\sum\limits_{i=1}^{S} \lambda_i + \\ \quad + \frac{1}{2} \cdot \sum\limits_{i=1}^{S} \sum\limits_{\tau=1}^{S} \lambda_i \cdot \lambda_\tau \cdot y_i \cdot y_\tau \cdot \kappa(z_i, z_\tau) \to \min\limits_{\lambda}, \\ \sum\limits_{i=1}^{S} \lambda_i \cdot y_i = 0, \\ 0 \le \lambda_i \le C, \ i = \overline{1, S}, \end{cases} \tag{2}$$

where λ_i is a dual variable; z_i is the object of the training set; y_i is a number (+1 or −1), which characterize the class of the object z_i from the experimental data set; $\kappa(z_i, z_\tau)$ is a kernel function; C is a regularization parameter ($C > 0$); S is the number of objects in the dataset; $i = \overline{1, S}$.

In training of the SVM classifier it is necessary to determine the kernel function type $\kappa(z_i, z_\tau)$, values of the kernel parameters and value of the regularization parameter C, which allows finding the compromise between maximizing of the gap separating the classes and minimizing of the total error. The linear, polynomial, radial basis or sigmoid function can be used as the kernel function $\kappa(z_i, z_\tau)$. In this research the radial basis function (Gaussian function) is used as the kernel function type: $\kappa(z_i, z_\tau) = exp(- < z_i - z_\tau, z_i - z_\tau > / (2 \cdot \sigma^2))$, where $< z_i - z_\tau, z_i - z_\tau >$ is a scalar product; σ [$\sigma > 0$ (by default $\sigma^2 = 1$)].

As a result of the training, the classification function is determined in the following form:

$$f(z) = \sum_{i=1}^{S} \lambda_i \cdot y_i \cdot \kappa(z_i, z_\tau) + b. \tag{3}$$

The classification decision, associating the object z to the class −1 or +1, is adopted in accordance with the rule:

https://dx.doi.org/10.15405/epsbs.2018.11.02.21
Corresponding Author: Irina Klyueva
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

$$F(z) = sign(f(z)) = sign(\sum_{i=1}^{S} \lambda_i \cdot y_i \cdot \kappa(z_i, z_\tau) + b) . \quad (4)$$

6. Findings

The enlarged scheme of the process of developing of the intelligent data analysis system and predicting of the new projects duration can be can be described as follows (Figure 1).

The input of the intelligent data analysis system receives the information about the characteristics of the new IT projects. Let it be required to fulfill the forecast of the implementation duration of these IT projects.

It is assumed that there already exists a certain database containing the statistical information on the already implemented IT projects of the similar orientation, therefore it is possible to form the dataset that will be used to develop the intelligent data classifier based on the SVM algorithm (Vapnik, 1998; Demidova et al., 2017a; Demidova et al., 2017b). A herewith, the characteristics values in the formed dataset should be normalized.

Previously, before the development of the intelligent data classifier, the relevance of the imbalance problem of the used dataset is checked (Demidova et al., 2017a). If this problem is actual, then the dataset is balanced, for example, using various modifications of the SMOTE algorithm (Demidova et al., 2017a).

The dataset formed on the basis of information on the previously implemented IT projects, and possibly balancing, is used to develop the SVM classifier. There are various variants for developing the intelligent classifier.

When applying one of the most easy to implement variants for developing the intelligent classifier, one or another algorithm for finding the optimal parameters values of the SVM classifier, for example, the particle swarm algorithm or the grid search algorithm, can be used (Demidova et al., 2017b). In this case, the development of the SVM classifier is accompanied by significant time costs.

Another variant for developing the intelligent classifier is to hybridize the SVM classifier, developed with the default parameters values, with some additional classifier, for example, with the RF classifier (Breiman, 2001; Hastie, Tibshirani & Friedman, 2009; Pal, 2015) or the *k*NN classifier (Demidova et al., 2017b). A herewith, the time to develop the SVM classifier is minimal, and the time costs for developing the additional classifier based on the RF or *k*NN algorithms are not so great, since the additional classifier is developed based not on the entire original dataset, but on the truncated dataset containing the information about the objects outside the separating strip. This truncated dataset is used to specify the classification of the objects within the separating strip, generated using the SVM classifier.

In any case, if the values of the classification quality indicators calculated for the developed intellectual classifier are acceptable, then it can be used to predict the duration of the new IT projects.

https://dx.doi.org/10.15405/epsbs.2018.11.02.21
Corresponding Author: Irina Klyueva
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

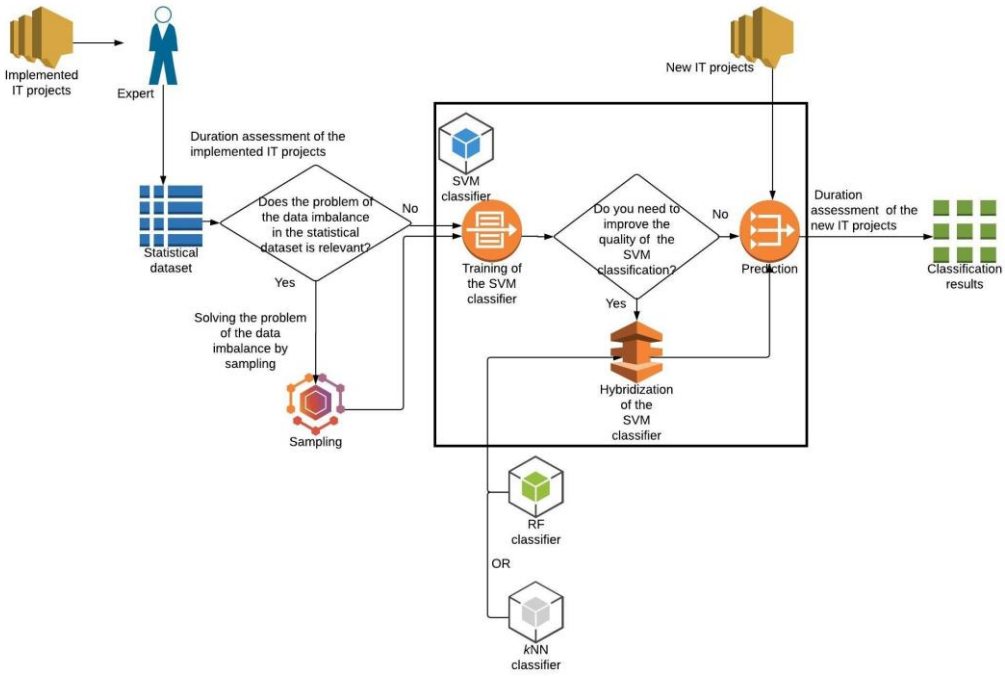

Figure 01. The intelligent data classification technology to the duration assessment of the new IT projects

To assess the quality of the developed classifiers the following classification quality indicators can be used: accuracy (*Acc*, also called the Overall Success Rate (*OSR*)); sensitivity (*Se*); specificity (*Sp*); and the balanced *F*-measure (*F*1) (Demidova et al., 2017a, Demidova et al., 2017b).

The hybridization of the SVM classifier with the RF classifier allows to obtain the sufficiently high values of the classification quality indicators with the less time expenditures than in the alternative approaches to application of the SVM classifier based on the search of its parameters values using, for example, the particle swarm algorithm or the grid search algorithm (Demidova et al., 2017b).

The RF algorithm compared to the kNN algorithm, works somewhat more slowly, but the RF classifier provides the better values of the classification quality indicators.

The testing of the developed intelligent data analysis system was carried out in the Python 3.5 software environment (Rossum & Drake, 2011) using some model datasets selected from the open sources (Demidova et al., 2017b). When using the hybridization of the SVM classifier with the RF classifier, almost in all cases, the values of the mentioned above classification quality indicators reached 100%. A herewith, when using the hybridization of the SVM classifier with the *k*NN classifier ensured the lower values of the classification quality indicators (from 98% to 100%).

The offered intelligent data analysis system provides the high values of the classification quality indicators and can be recommended for classification of the new objects, in particular, to predict the implementation duration of the new IT projects in the educational sphere.

7. Conclusion

The project duration is one of the most important indicators of its investment attractiveness. In view of the existing budget constraint in the funds allocation for the implementation of IT projects in education, the duration assessment of the of the IT new projects is an urgent problem.

https://dx.doi.org/10.15405/epsbs.2018.11.02.21
Corresponding Author: Irina Klyueva
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

In this paper, we propose the intelligent data classification technology to classification of the IT projects in the educational sphere which provides the assessment of the IT project duration. This intelligent data classification technology can be applied at the initial stages of the IT projects implementation in the absence of the complete information on their characteristics.

References

Breiman L. (2001). Random forests, *Machine Learning, 45*(1): 5–32.

CHAOS Report (2014). The Standish Group International, Inc..

Demidova L., Klyueva I. (2017a). SVM Classification: Optimization with the SMOTE Algorithm for the Class Imbalance Problem. In *6-th Mediterranean Conference on Embedded Computing (MECO)*, (p. 1-4).

Demidova L., Klyueva I., Sokolova Y., Stepanov N., Tyart N. (2017b). Intellectual approaches to improvement of the classification decisions quality on the base of the SVM Classifier, *Procedia Computer Science, 103*, 222-230.

Hastie T., Tibshirani R., Friedman J. (2009) Random Forests. Chapter 15 In *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag.

Pal M. (2015) Random forest classifier for remote sensing classification. *International Journal of Remote Sensing, 26*(1), 217–222.

Rossum G., Drake F. L. (2011). *The Python Language Reference Manual.* Network Theory Ltd.

Vapnik V. (1998). *Statistical Learning Theory*. New York: John Wiley & Sons..