UDC 004.85

Victoria M. Ruvinskaya[1] **,** Candidate of Technical Sciences, Professor of the Department of System Software, Institute of Computer Systems, E-mail: iolnlen@te.net.ua,
ORCID: http://orcid.org/0000-0002-7243-5535
Igor Shevchuk[1], student of the System Software Department, Institute of Computer Systems,
E-mail: rainn907@gmail.com, ORCID: http://orcid.org/0000-0002-1325-0450
Nikolai Michaluk[1], student of the System Software Department, Institute of Computer Systems,
E-mail: nmikhaluk@gmail.com, ORCID: http://orcid.org/0000-0003-3622-499X
[1]Odessa National Polytechnic University, Shevchenko Avenue, 1, Odessa, Ukraine, 65044

MODELS BASED ON CONFORMAL PREDICTORS FOR DIAGNOSTIC SYSTEMS IN MEDICINE

Abstract*: A disadvantage of many diagnostic systems is the inability to sufficiently assess the decisions reliability. While solving the problem of classification, each example may be classified with different degree of quality. So, a measure of the quality of an example classification was used (a non-conformity measure). The goal of the research is to improve evaluation of the diagnostics reliability in medicine based on conformal predictors which allow carrying out a probabilistic classification, as well as identifying abnormal cases when either the classifier is unable to determine the class for a particular object, or assigns one object to several classes at once. The paper describes the constructing and testing of various probabilistic binary classification models based on machine learning, particularly, the SVM method and conformal predictors using a non-conformity measure. For learning and testing the medicine Breast Cancer Wisconsin (Diagnostic) Data Set was used to construct linear, polynomial of different degrees and RBF models. We assessed the prediction results for every example from the test set as well as the integral characteristics of the quality of the models, taking into account both the correctness of the predictions for each class and the number of different types of anomalies. On the basis of the best selected models (linear, polynomial model of the 2nd degree and RBF), we developed an intelligent diagnostic system in medicine, which allows automating the model's construction, as well as carrying out the diagnostics and displaying the confidence of the received diagnosis or a message about the impossibility of making a diagnosis. The program also allows multiple doctors to log in to the system, adding new patients and editing information about them; every patient has their medical record with the results of the examination and the diagnoses given. The results of the research can be applied in the diagnostic systems for various diseases. This can be done by using the data with the symptoms and the corresponding diagnoses and constructing the appropriate models on this basis.*

Keywords: *data set; model; conformal predictors; machine learning; classification; significance; confidence; credibility; support vector machines*

1. Introduction

Currently, there is a tendency to increase the number of diagnostic medical information systems being developed. This is associated with an increase in the amount of biomedical information received, with the development of telemedicine, with an increase in requirements for the early diagnosis of diseases and several other factors. One of the main tasks, which are solved when designing systems for medical diagnostics, is the task of building a classifier.

The solution of the classification problem assumes that all diagnostic objects are characterized by certain features, according to which, based on some rules, belonging to a class is determined considering the goal of the problem is solved. For example, according to the results of the examination of the patient (features), the diagnosis (class) is determined in accordance with the classification rule, that is, a certain problem is diagnosed in the human body or a specific disease.

The classification rules depend on the complexity of the original features and can be set by an expert or obtained based on data in an explicit or implicit form. Currently, the most popular in the construction of classifiers are methods of machine learning.

Machine learning is a field of computer science that allows computer systems to "learn" from data without explicit programming [1; 2]. Machine learning is used to solve various problems, for example the problem of classification.

In order to "teach" a machine to determine the class of an object, it is necessary to set the decision rules for classification. For this, various methods of machine learning are used, such as Decision Tree [3], KNN (k-nearest neighbors) [4], SVM (Support Vector Machines) [5]. These methods allow determining the class of an object, but often this is not enough and it is required to determine the reliability of the confidence in the prediction results for a specific object can be estimated. Such methods do exist, it is a naive. Bayes classifier [6], logistic regression [7] and others. However it is reasonable, while using

the previously listed methods, to be able to carry out a probabilistic classification, and also, which is important, to detect anomalous cases when either the classifier is not able to define the class for a specific object, or assigns an object to several classes at once. This can be achieved with the help of conformal predictors [8].

II. Analysis of the Literature Data and Formulation of the Problem

Intelligent diagnostic systems based on the description of the subject area by an expert are widely used in connection with the development of knowledge engineering [9; 10; 11].

The history of the creation of diagnostic medical information systems begins with the MYCIN system, developed in the 70s of the last century [10]. It used an inference engine and a knowledge base of ~ 600 rules. The program asked the user (a doctor) a long series of simple "yes / no" or text questions. As a result, the system provided a list of suspected bacteria, sorted by probability, indicated a confidence interval for the probabilities of diagnoses and their justification, and also recommended a course of treatment.

Widely known the system for general functional diagnostics [12] already used not rules, but a database of diagnostic characters, which contains criterial representations of diseases, grouped by the diseases of functional systems and organs, as well as specific areas of medicine. Together they are presented as a logical tree that contains also the types of examinations that might reveal pathological changes, as well as the characteristic features (indicators) of a pathologically changed state of the patient's organs. Based on the selection of the closest criterial representations, the system gives the doctor a prompt about the variants for the likely diagnosis of the disease. The system contains also a database of the recommended methods of treatment.

The intellectual decision support system, presented in [13; 14], is used to assist during the diagnostics. Its knowledge database contains the symptoms, laboratory data and procedures linking them with a list of diagnoses. It provides support and justification for differential diagnoses and subsequent researches. Its database contains 4500 clinical manifestations that are associated with more than 2000 different nosologies.

For the above described systems, the knowledge base is entered manually, which is a labor-consuming process. Therefore, nowadays the most developed systems are those that integrate different approaches related to manual knowledge input and its construction on the basis of machine learning [15]. So an automated system for diagnosing bronchial asthma and chronic obstructive pulmonary disease is implemented on the basis of a neuro-fuzzy network and solves the problem of confirming or denial the proposed diagnosis [16]. The methods for constructing decision trees (algorithms: ID3, C4.5) allow forming decision rules based on the objects features, using machine learning [3; 17; 18]. Their main advantage is the ability to visualize the resulting patterns. In addition to giving a diagnosis, they can also be used to determine the medicine that is best suited for a given disease.

One of the approaches to diagnostics is the use of some decision support theory methods, in particular, the hierarchical structuring of complexity, risk management, paired comparisons, and others [19].

Existing diagnostic systems use modern software and hardware, for instance, the Isabel web system, a decision support system for differential diagnostics is available in various environments, particularly on mobile devices. It can be used as part of EHR / EMR or independently [20].

A disadvantage of many diagnostic systems is the inability to sufficiently assess the decisions reliability. While solving the problem of classification, each example may be classified with different degree of quality. A measure of the quality of an example classification – a non-conformity measure – was introduced by Vovk and Gammerman in [8], and later developed in [21; 22; 23]. The non-conformity measure is used to supplement the well-known classification algorithms with new ways of assessing the level of confidence in the results of their work [24]. For modern diagnostic systems, an important task is the introduction of modern methods for evaluating the reliability of decisions made and diagnostic results [25].

The goal of the research is to improve evaluation of the diagnostics reliability in medicine based on conformal predictors.

To achieve the goal the following problems should be solved:

– constructing various models for evaluation of decisions reliability using classification methods and conformal predictors based on the medical dataset and, as the result of the analysis, select the best ones for their application in the diagnostic system;

– developing an intelligent diagnostic system containing two subsystems: one for automating the work of the model developer, the second – for the doctor's use.

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

III. Constructing and Testing the Models for Diagnostics in Medicine

We chose the SVM (Support Vector Machine) [26] as the basic classification method. Its role for the binary classification is building a gap of maximum width bounded by two parallel hyperplanes, such that on one side of the gap there are objects (vectors) of one class, and on the other side – vectors of the other class; inside the gap there are no vectors. It is proved that not all vectors should be involved in the training of the model, but only those on the hyperplanes of the gap, they are called support vectors [27]. For them, the Lagrange multipliers a_i, participating in the calculations, are nonzero, for the other vectors $a_i = 0$.

The SVM classifier is described as follows [5]:

$$h(\square) = sign(u(x))$$

$$u(x) = \sum_{i=1}^{n} \alpha_i y_i K(x_i, x) - w_0,$$

where:

x – input vector for classification;

n – number of vectors in training set;

w_0 – absolute term;

α_i – Lagrange multipliers;

x_i, y_i – vectors from training set with their labels;

$K(x_i, x)$ – kernel function.

Examples of kernel functions are given below:

- linear kernel: $K(x_i, x) = x_i^T x$;
- polynomial kernel of degree k: $K(x_i, x) = (x_i^T x + c)^k$;
- RBF (Radial basis function): $K(x_i, x) = \exp\left(-\frac{||x - x_i||^2}{2\sigma^2}\right)$, where $\sigma = const$.

Support Vector Machine training comes down to finding α_i, Lagrange multipliers, and w_0 by solving optimization problem:

$$L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \rightarrow \max_{\alpha},$$

under restrictions:

$$\begin{cases} \sum_{i=1}^{n} \alpha_i y_i = 0 \\ 0 \le \alpha_i \le C; \; C = const. \end{cases}$$

However, SVM only determines if an object belongs to a class but does not determine the level of confidence in the classification results. In [8], a method was proposed for determining the measure of non-conformity for a classification using SVM based on Lagrange multipliers, because the a_i values determine how an element fits the training set. If $a_i = 0$, this means that the example fits the set very well (such vectors are uninformative and SVM ignores them when making predictions). If $a_i = C$, then such element is outlier; if $0 < a_i < C$, the large value of a_i indicates that the corresponding vector poorly fits the training set [24].

If we have a training set with labels from l objects, then when predicting a label for the element $l+1$, it is necessary to choose such a class Y that it "does not stand out" from the examples of the training set, that is, that a_{l+1} is less than as many a_i (i = 1, …l) corresponding to the training set as possible. For the most of data sets $a_{l+1} = 0$ where $Y = y_{l+1}$. Based on this rule and formula (1) below, the prediction is determined [21].

In order to determine the probability of an object belonging to a certain class, it is necessary to define the p-test of an object for each class. This value is determined by the formula (1) using the Lagrange multipliers a_i for each object from the training set, as well as the Lagrange multiplier a_{l+1} of a test object for which the class must be defined:

$$p_Y = \frac{\left|\{i = 1,\ldots,l : a_i \ge a_{l+1}\}\right|}{l+1}, \quad (1)$$

where:

p_Y is the p-test for class Y;

$l+1$ is the number of objects in the training dataset along with the test object.

p_Y, calculated by the formula (1), is the ratio of the number of a_i (i=1,2, … , l+1) greater than a_{l+1} to the total number of examples of the training set increased by 1 (the training set with the addition of one test object).

For a binary classification, it is necessary to define two p-tests: p_1 – test for the positive class and p_{-1} – for the negative class. For example, if the value is $p_{-1} < p_1$, we will be able to predict that the object belongs to class "1" with *Confidence* equal to $1 - p_{-1}$ and *Credibility* equal to p_1. And vice versa, if $p_1 < p_{-1}$, the object belongs to class "-1". Usually, for objects for which the correct prediction is obtained, *Credibility* = 1: for most datasets the percentage of support vectors is small, and thus a_{l+1}=0, and p_Y=1 when $Y = y_{l+1}$ [21].

SVM-based classification models using conformal predictors were constructed and tested for *Breast Cancer Wisconsin (Diagnostic) Data Set* from *Machine Learning Repository* [28]. Each of this dataset objects contains a patient's ID, nine fea-

tures (the results of the patients' examination, representing the characteristics of the extracted cells from the selected mass for observation) and the class. The examination results are integer numeric values: – Clump Thickness; – Uniformity of Cell Size; – Uniformity of Cell Shape; – Marginal Adhesion; – Single Epithelial Cell Size; – Bare Nuclei; – Bland Chromatin; – Normal Nucleoli; – Mitoses. The features in this dataset are the pre-calculated mathematical weightings for each input. They are obtained as a result of rating different attributes of the cells on a scale of one to ten, one being indicative of a benign mass and ten being indicative of a malignant tumor, before they definitively determined the diagnosis of the mass. Class "1" means that breast cancer was diagnosed and class "-1" means it was not. Thus, the model's input is the results of the processed patients' examination. Its output is the class which denotes whether the disease was detected or not (if the classification was successful; if not, it is indicated in a message), as well as the assessment of the diagnosis reliability.

During the formation of the training and test set duplicate objects were removed. Thus, a dataset was formed, which contained 564 objects, of which 309 were in the training set, 155 in the test set.

The experiments were conducted using the *LIBSVM* library [29] with the implementation of the *SVM* classification method. The following types of models were used: *Linear*, *Polynomial* (degrees of polynomials 2, 3, ..., 9) and *RBF (Radial basis function)* [30]. We also created a Java *Model Development* program implementing the above-described SVM-based classification using conformal predictors.

Table 1 shows a fragment of an extended table with the results of testing a polynomial model of the second degree. The test set contains 155 examples.

Table 1. A table fragment with the testing results, POLY: n = 2 – polynomial degree

ID	p_1	p_{-1}	True class	Predict class	Confidence (%)	Credibility (%)	Significance Level				
							0,2	0,15	0,1	0,05	0,01
…	…	…	…	…	…	…	…	…	…	…	…
125	1	0,03	1	1	97,09	100,00	1	1	1	1	2
126	0,1	1	-1	-1	89,64	100,00	1	1	2	2	2
127	0,14	0,07	1	1	92,56	13,92	0	0	0	0	0
128	1	0,03	1	1	97,09	100,00	1	1	1	1	2
129	1	0,03	1	1	96,76	100,00	1	1	1	1	2
130	0,14	1	-1	-1	86,08	100,00	1	1	2	2	2
131	1	0,04	1	1	96,12	100,00	1	1	1	1	2
132	1	0,03	1	1	96,76	100,00	1	1	1	1	2
133	0,07	1	1	-1	93,02	100,00	1	1	1	2	2
134	0,07	1	-1	-1	93,20	100,00	1	1	1	2	2
135	1	0,03	1	1	96,76	100,00	1	1	1	1	2
136	1	0,036	1	1	96,44	100,00	1	1	1	1	2
137	1	0,03	1	1	96,76	100,00	1	1	1	1	2
138	0,11	1	-1	-1	89,00	100,00	1	1	2	2	2
139	0,08	1	-1	-1	91,59	100,00	1	1	1	2	2
140	0,1	1	-1	-1	90,29	100,00	1	1	1	2	2
141	1	0,04	1	1	96,44	100,00	1	1	1	1	2
…	…	…	…	…	…	…	…	…	…	…	…

The values of the columns are calculated as follows:

– p_1 and p_{-1} are calculated by the formula (1);

– *True Class* is the example label in the dataset;

– *Predict class* is class "1" or "-1" depending on $max\ (p_1, p_{-1})$ (coincides with SVM prediction);

– *Confidence* (%) = $(1 - min\ (p_1, p_{-1}))*100$;

– *Credibility* (%) = $max\ (p_1, p_{-1})*100$.

These values are enough for creating and evaluating the prediction of a specific example.

For most examples, the prediction coincides with the real label (this is seen in Table 1); for them *Credibility* is 100 %; *Confidence*, i.e. prediction probability, ranges from 84 % to 97,4 %, but values above 90% prevail. This indicates that satisfactory

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

predictions were made for these examples with reasonable reliability. However, as we can see, it is not possible to make a correct prediction for example 127 due to the fact that none of *p* (neither *p1* nor *p-1*) are equal to 1 (are not large enough) and, accordingly, *Credibility* is small (equal to 13,92 %), i.e. much less than 1. This means that for this instance, the prediction (whether it corresponds to a real label or not) is impossible (Empty); and this fact can be determined on the basis of *Credibility*. In example 133, *Credibility* is 100 %, but the prediction does not coincide with the real label, which means, in particular, that either the real label was set incorrectly, and in this case, using the features of this example, it is possible to assert that the prediction for this example is true, or, perhaps, the prediction was incorrectly performed using the method with the help of which conformal predictors are built, in this case – SVM.

The work of the model was analyzed, examining the prediction results for each example of test data set. Similarly, we constructed and analyzed the results for other types of *SVM* (for each of 155 examples of the test set).

How to *evaluate the integral quality of the models* and, as a result, choose the best one for use in further diagnostics? Indeed, in the general case there are many examples in the test set, therefore it is difficult and visually impossible to make such a choice. To solve the problem, the concept of *Significance level* was used:

Significance Level = 1 – Confidence (%) / 100.

This work uses the following *Significance Level Thresholds*: 0,2; 0,15; 0,1; 0,05; 0,01, which correspond to 80 %, 85 %, 90 %, 95 %, and 99 % *Confidence* (see the names of 5 right columns in Table 1). On this basis, additional calculations are made, as described below.

We add columns to the extended testing table (see Table 1) corresponding to the taken *Significance Levels* and perform calculations, namely, calculate the number of classes that we can predict for a given example using the model for different *Significance Levels*. For the calculations, it is necessary to compare the p_1 and p_{-1} values of the object with the *Significance Level value*, so the following rules are used:

1) If *max* (p_1, p_{-1}) = 1 (this means that *Credibility* = 1) and *min* (p_1, p_{-1}) <= *Significance Level* (the *Significance Level* for the example is less than the *Threshold Significance Level*), then the value of the calculated column related to *Significance Level* is 1, since one class is precisely defined.

2) Otherwise, if *max* (p_1, p_{-1}) = 1 and *min* (p_1, p_{-1}) > *Significance Level*, the value of the calculated column corresponding to the *Significance Level* is 2, since both classes are determined as the result (Uncertain).

For example 125 *max* (p_1, p_{-1}) = 1 and p_{-1} = 0.03, i.e. p_{-1} <0,05 <0,1 <0,15 <0,2, which means that the first four of the last columns are equal to 1; but p_{-1}>0,01, so the last column is 2.

3) If *max* (p_1, p_{-1}) ≠ 1 (or not close to 1), then it is impossible to assign the example to any of the classes, therefore all additional columns for this example are equal to 0 (Empty) for all the significance levels.

For example, 127 both *p* is not equal to 1, which means that all the additional columns are equal to zero.

Furthermore, based on the performed calculations, we built integrated tables showing the models goodness of fit. Tables 2; 3 and 4 have the data for the chosen models which showed the best results on test data (polynomial models of the 1st and 2nd degree, as well as RBF).

The first column shows the taken *Significance Levels*.

The second and third columns show the number of correct and incorrect predictions from 87 examples belonging to class “1”; the fourth and fifth columns show the same for 68 examples belonging to class “-1”.

The sixth and seventh columns present the number of Empty and Uncertain (multi-valued) predictions.

The eighth and ninth columns show the overall results by which one can assess the quality of the models (the eighth column has the number of examples with the correct prediction, the ninth – the sum of incorrect predictions, as well as Empty and Uncertain ones).

Table 2. Integrated table. RBF: gamma = 1.0/9.0

Significance Level	Real: 1 Predict: 1	Real: 1 Predict: -1	Real: -1 Predict: -1	Real: -1 Predict: 1	Empty	Uncertain predictions	OK	Fail
1	2	3	4	5	6	7	8	9
0,01	0	0	0	0	0	155	0	155
0,05	0	0	0	0	0	155	0	155
0,1	64/87	7/87	68/68	0	6	10	132	23
0,15	62/87	3/87	67/68	0	23	0	129	26
0,2	59/87	2/87	66/68	0	28	0	125	30

Table 3. Integrated table. POLY: degree = 2

Significance Level	Real: 1 Predict: 1	Real: 1 Predict: -1	Real: -1 Predict: -1	Real: -1 Predict: 1	Empty	Uncertain predictions	OK	Fail
1	2	3	4	5	6	7	8	9
0,01	0	0	0	0	0	155	0	155
0,05	71/87	0	14/68	0	0	70	85	70
0,1	76/87	5/87	40/68	0	1	33	116	39
0,15	73/87	8/87	57/68	6/68	9	2	130	25
0,2	73/87	8/87	59/68	4/68	11	0	132	23

Table 4. Integrated table. POLY: degree = 1 (linear)

Significance Level	Real: 1 Predict: 1	Real: 1 Predict: -1	Real: -1 Predict: -1	Real: -1 Predict: 1	Empty	Uncertain predictions	OK	Fail
1	2	3	4	5	6	7	8	9
0,01	0	0	0	0	0	155	0	155
0,05	0	0	0	0	0	155	0	155
0,1	0	0	0	0	0	155	0	155
0,15	75/87	1/87	65/68	0	14	0	140	15
0,2	75/87	1/87	65/68	0	14	0	140	15

Analysis of the integrated tables leads to the conclusion that:

– a polynomial model of the 1st degree (linear) shows satisfactory results only at the levels of significance 0,15 and 0,2, and does it better than other models;

– a polynomial model of the 2nd degree at a significance level of 0,1 shows better results than a linear model, but for the levels of 0,15 and 0,2 the results are worse;

– the RBF model is better than the polynomial model of the 2nd degree; it shows the results at the significance levels of 0,1; 0,15 and 0,2.

In order to facilitate the process of creating classification models based on conformal predictors, that is, training and testing process, we created the *Model Development* program that allows loading ready-made as well as creating new data sets, constructing models and viewing training and prediction results with different levels of detail. Fig. 1 shows a window for entering modelling parameters and viewing the results in the form of an integrated table.

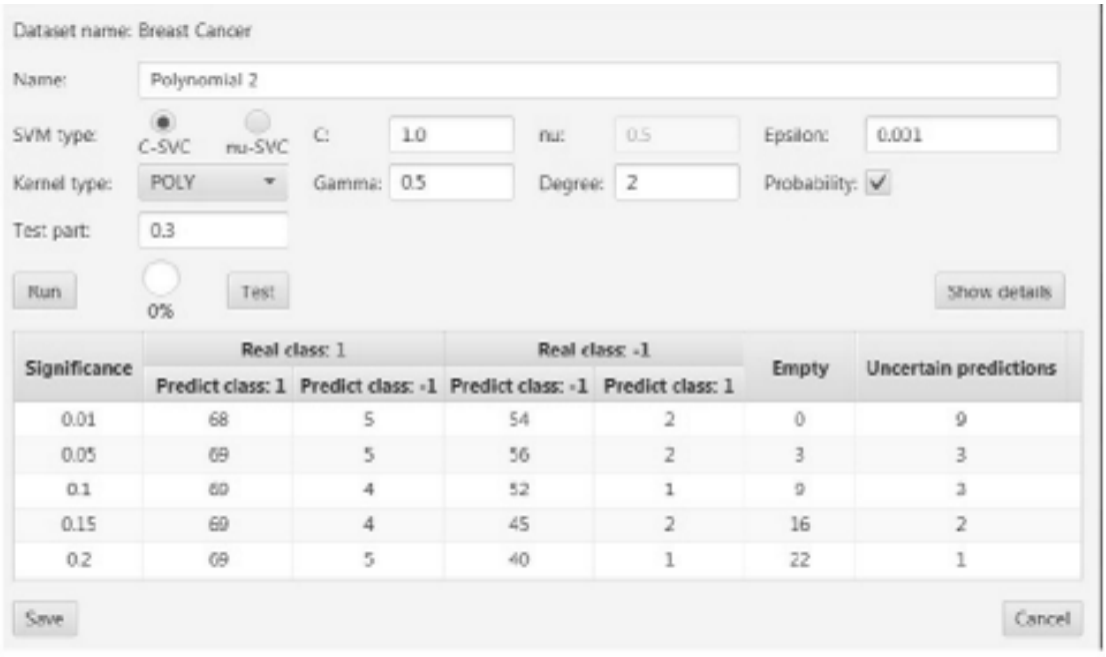

Fig. 1. The window for parameters tuning and viewing the modelling results

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

The resulting models were used to diagnose cancer based on the examination results. For this purpose, we have developed a program for the doctor which provides an opportunity for the doctor to enter a patient's examination data, select the type of disease and models required for diagnostics. Then the information is processed and the diagnostic result (positive or negative) is issued, as well as the *Confidence* of this result in percent (Fig. 2). If the system cannot diagnose, a message is shown.

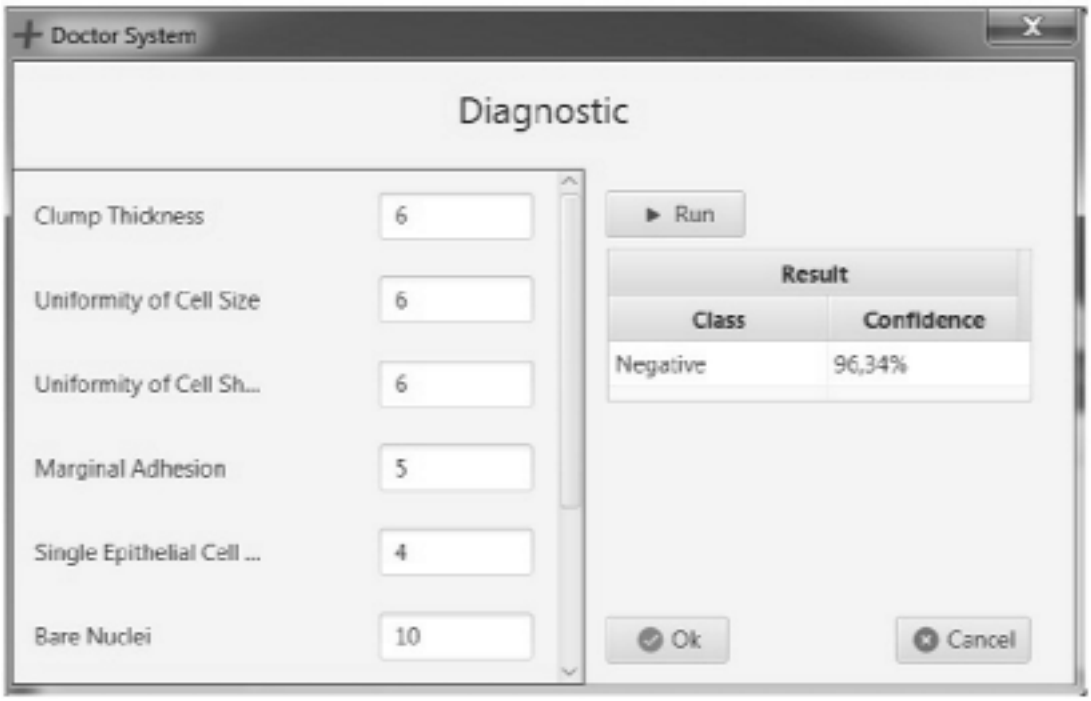

Fig. 2. The diagnostic window for the doctor

The developed program also provides windows for doctors and model developers to log in to the system. The system allows interactively adding new patients and editing information about them. Every patient has their electronic medical record with the results of the examination and the diagnoses given.

The architecture of the software is client-server. Data is stored and all the calculations are made on the server-side, while on the client-side data sets and models are created, and the classification results are displayed.

In the future, it is advisable to expand the system so that, based on the results of the patient's examination, it will be possible to receive not only one diagnosis, but a list of diagnoses ranked by probability. There are two standard ways to reduce the classification problem for the case of several classes to a binary classification: the “one-against-the-rest” and “one-against-one” procedures. In [8], it is shown how the measure of non-conformity is calculated in the case of more than two classes for each of the procedures, based on the measure of non-conformity for the binary classification. On this basis, a_i is calculated, and then p_Y for each of the classes. Such an approach is possible if the classes do not overlap, that is, if a person haven’t multiple illnesses at the same time considered for diagnosis (this is often required for the correct work of classification methods). If the classes overlap, it is advisable to carry out a binary classification for each disease separately and find the probabilities of assigning an object to each of the classes.

IV. Conclusion

The analysis of existing diagnostic medical information systems and properties of the solution to the problem of classification for diagnosis has been carried out. The necessity and importance of assessing the quality of the diagnoses is shown. It is proposed to apply conformal predictors to analyze the credibility measure and the reliability of the results based on the modern machine learning methods with calculations of the non-conformity measure for examples when solving the classification problem.

The models were formed on data from medical dataset using the SVM method based on conformal predictors and was assessed their quality. As a result, three best models were selected, allowing not only to get diagnosis and determine its probability (as it is practiced in known systems), but also to detect cases when diagnosis is impossible: firstly, the classifier is unable to determine the class for the object, secondly, the classifier relates the object to sev-

eral classes. The last is possible only with the use of conformal predictors.

A software product has been developed to automate the process of constructing models. In addition, this software allows the oncologist to diagnose using the created models and to assess the reliability of each diagnosis.

The developed solution allows, when expanding the accumulated data on symptoms and diagnoses based on them, to build appropriate classification models with an assessment of the quality of the diagnoses for a wide range of medical problems. Thus, the obtained results can be used in the systems of medical diagnostics for various diseases.

References

1. (2019). "A Course in Machine Learning", Posted by Hal Daumé III, 2015, [Electronic Resource]. – Access mode: http://ciml.info/dl/v0_9/ciml-v0_9-all.pdf. – Active link – 27.02.2019.

2. Flach P. (2015). Mashinoe obuchenie: nauka i iskustvo postroenia algoritmov, kotorie izvlekaut znania iz dannich, [Machine Learning: The Art and Science of Algorithms that Make Sense of Data], *Publ. DMK Press*, 400 p. (in Russian).

3. (2012). "Using Decision Trees in Evidence Based Medicine". Posted by Venky Rao, [Electronic Resource]. – Access mode: https://www.datasciencecentral.com/profiles/blogs/using-decision-trees-in-evidence-based-medicine. – Active link – 13.03.2019.

4. (2019). "Data Algorithms by Mahmoud Parsian", [Electronic Resource]. – Access mode: https://www.oreilly.com/library/view/data-algorithms/9781491906170/ch13.html. – Active link – 02.03.2019.

5. Statistical Learning Theory, & Vladimir N. Vapnik, (1998). "JOHN WILEY & SONS", Incorporation.

6. *Shitikov V., & Mastickiy S.* (*2017). Klasifikacia*, regressia i drugie algoritmi Data Mining s ispolsovaniem R, [Classification, Regression and other algorithms of Data Mining with the use of R], [Electronic Resource] – Access mode: https://ranalytics.github.io/data-mining/index.html (in Russian). – Active link – 15.03.2019.

7. David, W. Whosmer, & Stainly, Lemetshow. (2000). "Applied Logistic Regression", JOHN WILEY & SONS, Incorporation.

8. Vovk, V., Gammerman, A. & Shafer, G. (2005) "Algorithmic Learning in a Random World", *Publ. Springer*, New York.

9. Gavrilova, T., & Choroshevskiy, V. (2005). Basi znaniy intelektualnich system, [Knowledge bases of intellectual systems], *Publ. Izd. SPb*. Piter, Russian Federation (in Russian).

10. Joseph C. Giarratano, & Gary D. Riley. (2007). "Expert Systems: principles and programming, Thomson course technology", PeopleSoft, Incorporation.

11. Yalovec A. (2011). Predstavlenie i obrabotka znaniy s tochki zrenia matematicheskogo modelirovania. Problemi i reshenia, [Representation and processing of knowledge from the point of view of mathematical modeling. Problems and Solutions], Kiev, Ukraine, *Publ. Naukova Dumka,* NAN Ukraine, 399 p. (in Russian).

12. Strahov A., Strahov O., & Strahov E. (2019). Sposob avtomatizacii obshei funkcionalnoi dignostiki, [Automation method of general functional diagnostics], [Electronic Resource]. – Access mode: http://www.findpatent.ru/patent/220/2205447.html (in Russian). – Active link – 10.03.2019.

13. (2019). "DXplain: Patterns of Use of a Mature Expert System". Edward, P Hoffer, Mitchell, J. Feldman, Richard, J. Kim, Kathleen, T. Famiglietti, and G. Octo, Barnett, 2005, [Electronic Resource] – Access mode: https://www.ncbi.nlm.nih.gov/pmc/articles/ PMC1560464/. Active link – 20.01.2019.

14. (2019). "Using decision support to help explain clinical manifestations of disease", [Electronic Resource] – Access mode: http://www.mghlcs.org/projects/dxplain/. Active link – 06.01.2019.

15. Ruvinskaya V. & Moldavskaya A. (2018). "Methods for Automated Generation of Scripts Hierarchies from Examples and Diagnosis of Behavior, Recent Developments in Data Science and Intelligent Analysis of Information", ICDSIAI 2018, June 4-7, Kyiv, Ukraine, pp. 189-198. Part of "Advances in Intelligent Systems and Computing", *Publ. Springer*, Cham, Vol. 836.

16. Bezrukov, N., Eremin, E., & Perelman, Y. (2007). Avtomatizirovannay systema diagnostiki zabolevaniy legkich, [Automated lung disease diagnosis system], Control Sciences, 5, pp. 75-80 (in Russian).

17. (2019). "Building Classification Models: ID3 and C4.5, Temple University (US), [Electronic Resource] – Access mode: **Ошибка! Недопустимый объект гиперссылки.**. – Active link – 02.02.2019.

18. Ruvinskaya, V., Parshin, I., & Shevchuk, I. (2018). Provedenie experimentov po diagnostike v medicine na osnove metodov klassifikacii i analyz

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

resultatov, [Conducting experiments on the diagnosis in medicine and analysis of the results], *Modern Information Technologies*, pp. 27-28 (in Russian).

19. (2013). "M-health: supporting automated diagnosis", electronic health records by Efthimios, Alepis and Christos, Lambrinidis, [Electronic Resource] – Access mode: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3611032/. – Active link – 03.01.2019.

20. (2016). "Isabel Symptoms Checker for Patient Engagement", 1st European Conference, 3rd June, 2016, Rotterdam, the Netherland, [Electronic Resource]. – Access mode: http://v4.isabelhealthcare.com/home/default – Active link – 11.03.2019.

21. Alex, Gammerman, & Vladimi, Vovk. (2007). "Hedging Predictions in Machine Learning", the *Computer Journal*, 50: pp. 151-163.

22. Paolo, Toccaceli, Ilia, Nouretdinov, & Alexander, Gammerman. (2016). "Conformal Predictors for Compound Activity Prediction", Conformal and Probabilistic Prediction with Applications. 5th International Symposium, COPA 2016, Madrid, Spain, April 20-22, 2016. Proceedings, Part of "Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)", *Publ. Springer Naturem*, Vol. 9653, pp. 51-66.

23. Antonis, Lambrou & Harris, Papadopoulos. (2016). "Binary Relevance Multi-label Conformal Predictor", Conformal and Probabilistic Prediction with Applications, *5th International Symposium, COPA 2016,* Madrid, Spain, April 20–22, 2016. Proceedings, Part of "Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)", *Publ. Springer Nature,* Vol. 9653, pp. 90-104.

24. Viugin, V. (2013). Matematicheskie osnovi teorii mashinogo obuchenia i prognozirovania, [Mathematical foundations of the theory of machine learning and forecasting], Moscow, Russian Federation, 387 p. (in Russian).

25. (2019). "Machine Learning in Medicine", [Electronic Resource]. – Access mode: http://circ.ahajournals.org/content/132/20/1920.short . – Active link – 20.01.2019.

26. Donskoi, V. (2014). Algoritmicheskie modeli obuchenia i klasifikacii: obosnovanie, sravnenie, vibor, [Algorithmic training and classification models: justification, comparison, choice], Simferopol, *Publ. DIIP*, 228 p. (in Russian).

27. Hastie, T., Tibshirani R., & Friedman, J. (2009). Chapter 7.9. Vapnik–Chervonenkis Dimension, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction", 2nd ed. *Publ. Springer-Verlag,* 746 p.

28. (2019). Breast Cancer Wisconsin (Diagnostic) Data Set, [Electronic Resource] – Access mode: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29. – Active link – 21.02.2019.

29. (2019). Chih-Chung Chang and Chih-Jen Lin. LIBSVM – A Library for Support Vector Machines, [Electronic Resource]. – Access mode: https://www.csie.ntu.edu.tw/~cjlin/libsvm/. – Active link – 13.03.2019.

30. Zagoruiko, N. (1999). Prikladnie metodi analiza danich i znaniy, [Applied methods of data and knowledge analysis], Novosibirsk, Russia, *Publ. Izdatelstvo Instituta Matematiki* (in Russian).

Received 20.03.2019

УДК 004.85

[1]**Рувінська, Вікторія Михайлівна,** кандидат технічних наук, професор, професор кафедри системного програмного забезпечення інституту комп'ютерних систем, E-mail: iolnlen@te.net.ua, ORCID: http://orcid.org/0000-0002-7243-5535
[1]**Шевчук, Ігор**, студент кафедри системного програмного забезпечення інституту комп'ютерних систем, E-mail: rainn907@gmail.com, ORCID: http://orcid.org/0000-0002-1325-0450
[1]**Михалюк, Микола**, студент кафедри системного програмного забезпечення інституту комп'ютерних систем, E-mail: nmikhaluk@gmail.com, ORCID: http://orcid.org/0000-0003-3622-499X
[1]Одеський національний політехнічний університет, 1, проспект Шевченка, Одеса, Україна, 65044

МОДЕЛІ НА БАЗІ КОНФОРМНИХ ПРЕДИКТОРІВ ДЛЯ СИСТЕМ ДІАГНОСТИКИ В МЕДИЦИНІ

Анотація: *Недоліком багатьох діагностичних систем є неможливість в достатній мірі оцінити достовірність рішень. При вирішенні проблеми класифікації кожен приклад може бути класифікований з різним ступенем якості. Запропонована міра якості зразкової класифікації (міра невідповідності). Мета дослідження - поліпшити оцінку достовірності діагностики в медицині на основі конформних предикторів, які дозволяють проводити вірогідну класифікацію, а також виявляти ненормальні випадки, коли класифікатор не може визначити клас для конкретного об'єкта, або відносить один об'єкт до окремих класів одночасно. У статті описується побудова і тестування різних імовірнісних моделей двійковій класифікації на основі машинного навчання, зокрема, методу SVM і конформних предикторів, що використовують міру невідповідності. Для вивчення і тестування моделей був використаний базі набір даних Breast Cancer Wisconsin (Diagnostic) Data Set для побудови лінійних, полиномов різного ступеня і моделей RBF. Оцінені результати прогнозування для кожного прикладу з набору тестів, а також інтегральні характеристики якості моделей, з урахуванням як правильності прогнозів для кожного класу, так і кількості різних типів аномалій. На основі кращих відібраних моделей (лінійна, поліноміальна модель 2-го ступеня і RBF) розроблена інтелектуальна діагностична система для застосування в медицині, яка дозволяє автоматизувати побудову моделі, а також проводити діагностику і відображати достовірність отриманого діагнозу або повідомляти про неможливість поставити діагноз. Програма також дозволяє декільком лікарям входити в систему, додавати нових пацієнтів і редагувати інформацію про них. Кожен пацієнт має свою медичну карту з результатами обстеження і поставленими діагнозами. Результати дослідження можуть бути застосовані в системах діагностики різних захворювань. Це можна зробити, використовуючи дані з симптомами і відповідними діагнозами і створивши відповідні моделі на цій основі.*

Ключові слова: *набір даних; модель; конформні предиктори; машинне навчання; класифікація; рівень значимості; впевненість (достовірність); правдоподібність; метод опорних векторів*

УДК 004.85

[1]**Рувинская, Виктория Михайловна,** кандидат технических наук, профессор, профессор кафедры системного программного обеспечения, института компьютерных систем, E-mail: iolnlen@te.net.ua, ORCID: http://orcid.org/0000-0002-7243-5535
[1]**Шевчук, Игорь**, студент кафедры системного программного обеспечения, института компьютерных систем, E-mail: rainn907@gmail.com, ORCID: http://orcid.org/0000-0002-1325-0450
[1]**Михалюк, Николай**, студент кафедры системного программного обеспечения, института компьютерных систем, E-mail: nmikhaluk@gmail.com, ORCID: http://orcid.org/0000-0003-3622-499X
[1]Одесский национальный политехнический университет, проспект Шевченко, 1, Одесса, Украина, 65044

МОДЕЛИ НА БАЗЕ КОНФОРМНЫХ ПРЕДИКТОРОВ ДЛЯ СИСТЕМ ДИАГНОСТИКИ В МЕДИЦИНЕ

Аннотация: *Недостатком многих диагностических систем является невозможность в достаточной степени оценить достоверность решений. При решении проблемы классификации каждый пример может быть классифицирован с разной степенью качества. Предложена мера качества примерной классификации (мера несоответствия). Цель исследования - улучшить оценку достоверности диагностики в медицине на основе конформных предикторов, которые позволяют проводить вероятностную классификацию, а также выявлять ненормальные случаи, когда классификатор не*

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

может определить класс для конкретного объекта, либо относит один объект к нескольким классам одновременно. В статье описывается построение и тестирование различных вероятностных моделей двоичной классификации на основе машинного обучения, в частности, метода SVM и конформных предикторов, использующих меру несоответствия. Для изучения и тестирования моделей был использован базе набор данных Breast Cancer Wisconsin (Diagnostic) Data Set для построения линейных, полиномов различной степени и моделей RBF. Оценены результаты прогнозирования для каждого примера из набора тестов, а также интегральные характеристики качества моделей, с учетом как правильности прогнозов для каждого класса, так и количества различных типов аномалий. На основе лучших отобранных моделей (линейная, полиномиальная модель 2-й степени и RBF) разработана интеллектуальная диагностическая система для применения в медицине, которая позволяет автоматизировать построение модели, а также проводить диагностику и отображать достоверность полученного диагноза или сообщать о невозможности поставить диагноз. Программа также позволяет нескольким врачам входить в систему, добавлять новых пациентов и редактировать информацию о них. Каждый пациент имеет свою медицинскую карту с результатами обследования и поставленными диагнозами. Результаты исследования могут быть применены в системах диагностики различных заболеваний. Это можно сделать, используя данные с симптомами и соответствующими диагнозами и создав соответствующие модели на этой основе.

Ключевые словаs: *набор данных, модель; конформные предикторы; машинное обучение; классификация; уровень значимости; уверенность (достоверность); правдоподобие; метод опорных векторов*