Вісник Тернопільського національного технічного університету
https://doi.org/10.33108/visnyk_tntu
Scientific Journal of the Ternopil National Technical University
2019, № 3 (95) https://doi.org/10.33108/visnyk_tntu2019.03
ISSN 2522-4433. Web: ***visnyk.tntu.edu.ua***

INSTRUMENT-MAKING AND INFORMATION-MEASURING SYSTEMS

ПРИЛАДОБУДУВАННЯ ТА ІНФОРМАЦІЙНО-ВИМІРЮВАЛЬНІ СИСТЕМИ

UDC 004.021

ABOUT THE APPROACH OF SOLVING MACHINE LEARNING PROBLEMS INTEGRATED WITH DATA FROM OPEN SOURCE SYSTEMS OF ELECTRONIC MEDICAL RECORDS

Vasyl Martseniuk[1]; Nazar Milian[2]

[1]Belsko-Biala University, Belsko-Biala, Poland
[2]Ternopil Ivan Puluj National Technical University, Ternopil, Ukraine

Summary*. In recent decades, open source health solutions and commercial tools have been actively developed. The most common open source electronic health accounting systems are WorldVistA, OpenEMR and OpenMRS. Scientists drew attention to the prospects of open-source electronic health records software and free systems for countries with certain financial difficulties and such developing countries. Setting the task of machine learning in medical research has been carried out. The flowchart presented in the paper demonstrates the main steps for developing a machine learning model. It is noted that the task of importing training, testing and forecasting data sets from EMR systems in the machine learning environment is not so trivial for a number of reasons discussed in the study. Here are some basic approaches for accessing patient medical record data in conventional EMR systems. Some features of approaches for the two most common EMR open source systems are presented: OpenEMR, OpenMRS. Despite a long period of development and applications, even leading and widespread EMR systems (both commercial and free open source) have limited or partial support for HL7 capabilities. Despite the challenges that the implementation level is considering, there are enough arguments to adapt the use of data formats compatible with HL7 and to develop information systems that are machine learning oriented. Experimental studies are related to the prediction of fractures for middle-aged women, confirm that this is a pressing, preventive problem today. The development of the machine learning model is implemented in the free software environment R, using the mlr package. As a result, we get machine learning models based on five methods. The results of the effectiveness of the methods, using the mmce measure, show that the exact model of compliance with the assessment of prediction quality is the random forest method, worst of all is the ferms method.*

Key words: *machine learning, PCA, classification, EMR system, mlr.*

https://doi.org/10.33108/visnyk_tntu2019.03.105 *Received 06.10.2019*

Problem statement. Today, machine learning in medical research is one tool for analyzing experimental data in clinical research, just as the language by which the findings can be presented is not just the task of machine learning in medicine. The mathematical machine learning apparatus is widely used to diagnose solving classification problems and finding new representations to predict, formulate, and test new scientific hypotheses. The use of machine learning algorithms involves knowledge of the main methods and stages of data analysis: their consistency, necessity and adequacy. The proposed work does not focus on the detailed

Corresponding author: Nazar Milian; e-mail: nazar.milyan@gmail.com

presentation of formulas for the development of methods, but on their content and application rules in the case of open source electronic medical record systems.

Analysis of known research results. Public health solutions based on open source software have been actively developed in recent decades together with commercial means [3], [7], [1]. The most common open source systems of electronic medical records are WorldVistA, OpenEMR and OpenMRS. The prospects of open source electronic medical records software and free systems for developing countries and countries with financial difficulties were discussed in the works of F. Aminpour, F. Fritz, C. J. Reynolds and others [3], [7], [4]. Approaches to the implementation of open source systems of electronic medical records, especially OpenEMR, OpenMRS and OpenDental, in the health care system of Ukraine have also been studied, as well as methods of integration of these systems with other software of medical direction, developed by authors in recent years [8], [9], [6], [5].

The goal of the work. The goal of the work is to develop mathematical, software for the development of models of machine learning in medicine, based on the use of systems of electronic medical records with open source and means of machine learning.

Setting up the objectives. Mathematically, machine learning tasks in medical research are based on such data is multiple *D* containing *N* tuples. Depending on the task, certain sets of tuples can be used for training, testing, and forecasting. Each *i* tuple $(a_1^i, a_2^i, ..., a_p^i, c^i)^T$ consists of input data $(a_1^i, a_2^i, ..., a_p^i)^T$ (called attributes) and output data c^i which is an attribute of the class. Let the vector string $a_j = (a_j^1, a_j^2, ..., a_j^N)$ represent the value of its j – attribute of all *N* tuples. Attributes $a_1, ..., a_p$ can accept both numeric and categorical data. *C* class attribute takes one of the *K* discrete values $c \in \{1, ..K\}$.

The goal is to predict using some predictor, the attribute value of the class *C* based on the attribute $a_1, ..., a_p$ values. This should maximize the accuracy of the class attribute prediction, namely the probability $P\{c = c^*\}$ for the attribute $c^* \in \{1, ..., K\}$.

The first task solved in real medical research is to reduce the dimension $p \in N$. To this end, this modification of the principal component analysis (PCA) method is proposed. This modification includes the following steps:

Input $A = \{(a_1^i, a_2^i, ..., a_p^i, c^i)^T\}_{i=1}^N$:

Output: Main components, along with attributes.

1. Convert all categorical attributes by encoding them as a set of Boolean inputs, each represented by category 0 or 1. We can generate columns with categorical check boxes automatically. As a result, we get a numerical matrix $X = \{(a_1^i, x_2^i, ..., x_{p_1}^i)^T\}_{i=1}^N \in R^{p_1+1\times N}$.

2. Calculate average value for rows $\bar{x}_i = \frac{1}{N}\sum_{j=1}^N x_i^j, i = \overline{1, p_1}$.

3. Calculate variations $Var(x_i)$, $i = \overline{1, p_1}$. We assume that the variation is a complete variation $Var(X) = \sum_{i=1}^{p_1} Var(x_i)$ (the sum of the sample variations).

4. Calculate deviation matrix $X' = \{x_j^i - \overline{x_i}\}_{i=\overline{1,p_1}, j=\overline{1,N}} \in R^{p_1 \times N}$.

5. Evaluate Cavariation Matrix $C = \frac{1}{p_1} X'(X')^T \in R^{p_1}$.

6. Calculate eigenvalues of matrix $C: \lambda_1 \le \lambda_2 \le \ldots \le \lambda_{p_1}$.

7. Calculate eigenvectors *C*. Consider eigenvectors w_{p_1} and $w_{p_1-1} \in R^{p_1}$, corresponding to λ_{p_1} and λ_{p_1-1} and respectively. Take the first two components $PC1 = X^T w_{p_1}$ and $PC2 = X^T w_{p_1-1}$. We calculate variations of $Var(PC1)$ and $Var(PC2)$. Hence we have a percentage of the explained variation that corresponds to the first two components namely and respectively $ExplainedVar(PC1) := \dfrac{Var(PC1)}{Var(X)}$ and $ExplainedVar(PC2) := \dfrac{Var(PC2)}{Var(X)}$.

8. We ordering the value of eigenvectors w_{p_1} and w_{p_1-1} in the descending of their absolute values. To this we use $\pi(w_{p_1})$ and $\pi(w_{p_1-1})^5$ permutations. Then we return the names of the first $ExplainedVar(PC1)*100\%$ attributes in the permutation $\pi(w_{p_1})$ and the first $ExplainedVar(PC2)*100\%$ attributes in the permutation $\pi(w_{p_1-1})$.

As a result of the decrease in dimension, we get some numerical matrix $X^{red} = \{(x_1^i, x_2^i, \ldots, x_{p_2}^i, c^i)^T\}_{i=1}^N \in R^{p_2+1\times N}$. This data can then be used as training for a number of machine learning tasks.

Presentation of the main material of the research.

1.1. Machine Learning Model Development.

Our approach is based on the development of a machine learning model, we are referring to a common block diagram that allows the use of machine learning solver with the capabilities of accurate, sensitive, sustainable results.

According to the diagram shown in Figure 1, we start with import and preparation (feature ingeneering, gap filling, normalization), which are collected in EMR systems. The ways in which you can import datasets from EMR systems are discussed in section 1.2. Here, we notice that the choice of open source electronic medical record systems compared to commercial systems is critical, as it provides opportunities to have open access to clinical data that is processed and selected in the next machine learning steps.

In real-world applications, as shown in the study results, we are dealing with many attributes and only some of them can be essential to solving machine learning problems. So, naturally try to reduce the dimension of the attributes left with the largest variations.

Then we have to denote the task itself in terms of machine learning. It can be regression, classification, rating, etc. Then we select the appropriate solution method (Lener). The most important is the selection of parameters for the methods. This affects the accuracy of the model.

The last steps (starting with selecting a method) can be repeated to get the most efficient model. The capabilities of modern programming tools even allow to compare methods using the corresponding benchmarks, which are developed according to certain tasks.

The machine learning model presented below is consistent in its entirety with the **mlr** package that is demonstrated in the results of the study.

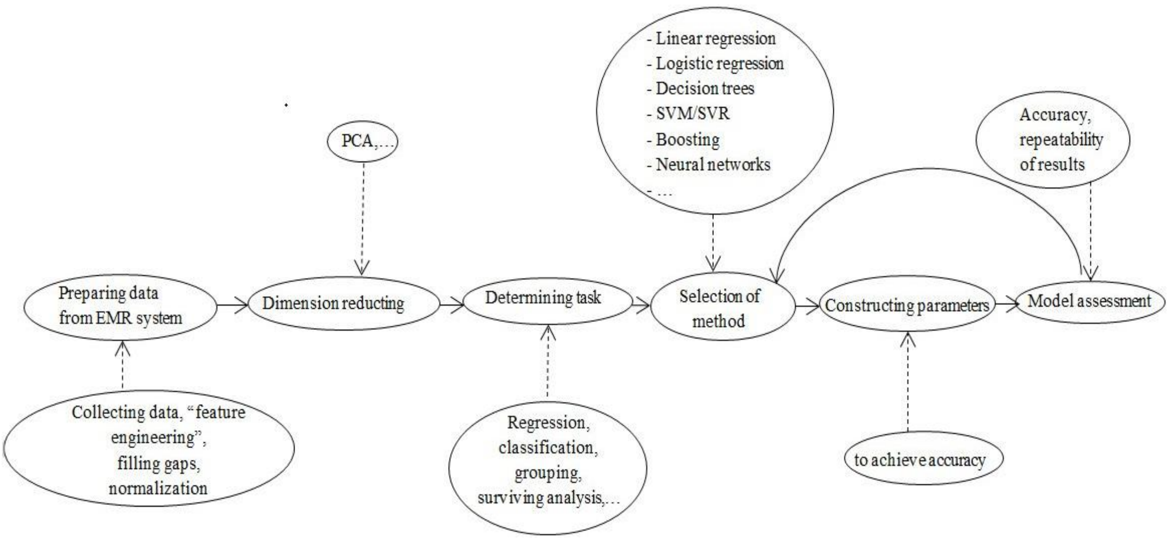


Figure 1. Development of machine learning model for medical research

1.2. Task of importing data from open source systems of electronic medical records.

The task of importing training, testing and forecasting data sets from EMR systems in a machine learning environment is not so trivial for a number of reasons: in terms of programming, ethical and legal. There are several basic approaches for accessing patient medical record data in conventional EMR systems:

- Using remote access API libraries for EMR systems (REST, XMLRPC, SOAP).
- Using standardized HL7 data formats.
- Using common standard data formats (XML/CSV).

Here are some features of applying the above approaches to the two most common EMR open source systems: OpenEMR, OpenMRS.

Looking at these native API remote access libraries for commercial EMR systems, we note the lack of (or at least open information is not available) technology tools (especially API remote access libraries). Even free open source EMR systems have little remote access using API libraries.

For example, OpenEMR offers a single native remote API library. Although it was developed in 2013, the Master Mobility App is still useful and fully functional.

OpenMRS implements remote access to internal APIs via REST using the REST module. The API OpenMRS specification is available.

HL7 provides a framework (and some standards) for sharing, integrating, disseminating, and receiving electronic medical information. The following versions of HL7 is further active: HL7 Version 3 (V3) Normative Edition (is based on the big XML set of forms) HL7 Version 3 Clinical Document Architecture (CDA) HL7 Fast Healthcare Interoperability Resources Specification (FHIR, also on the basis of XML) HL7 Version 2 Product Suite (versions 2.x (V2) of HL7 are used by connection of record of exchange of messages on the basis of text files). Please note that modern HL7 CDA documents use Continuity of Care Document (CCD) term very often.

Unfortunately, despite a long period of development and applications, even leading and widespread EMR systems (both commercial and free open source) have limited or partial support for HL7 capabilities. For example, OpenEMR has only a few built-in HL7 capabilities:

- Officially supported import of patient medical records, but only in the form of HL7 Version 2 messages.

- The CCD patient data export format is HL7 compatible with CDA is available, but only for the individual patient.
- In 2017 implementation of FHIR framework integration started, but it is still fully procured. Also promising an integration process based on refactoring HAPI FHIR libraries using the PHP programming language.

OpenMRS provides more capabilities, including bi-directional data processing HL7 using user modules:

- Socketh7Listener Module accepts and processes the import of HL7 messages (version 2.x only).
- CCD Module exports patient medical summary records in HL7 third version of the related CCD format.
- HL7Query Module24 also supports the export of patient data as 2.x HL7 messages (current ORUR01 version only) using a special module.

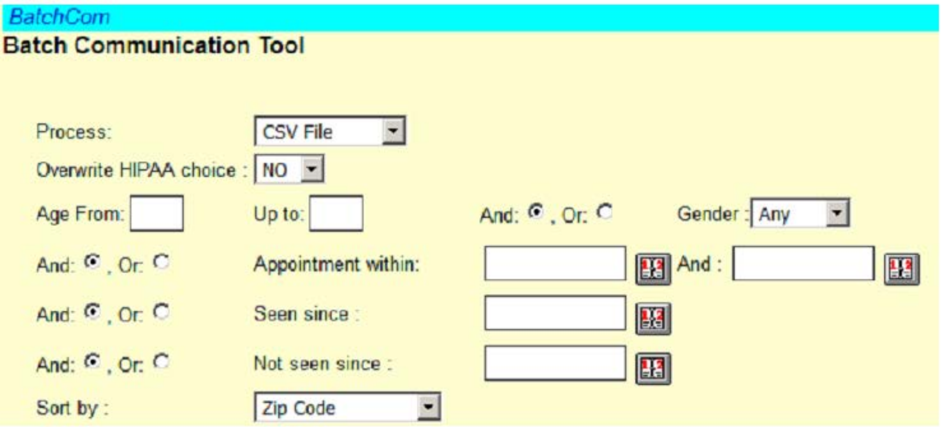
Despite the above problems, the implementation level is considered, there are sufficient arguments for adapting the use of HL7 compatible data formats and developing information systems that are machine learning oriented. The availability and active development of a large number of free open source HL7 parsers for different programming languages is one of such reasons.

Exporting data in common format is implemented in EMR systems in the following ways:

- Export of individual documents (medical forms) for selected patients. Such traits are supported by most EMR systems. Common export formats are PDF, XML, XLS, CSV.
- Export of full patient medical records. Typical implementations include the use of XML-based data, according to the Continuity of Care Record (CCR) specification. The CCR itself is an implementation of another compatibility standard, ASTM International.

Exporting patient medical records command files from EMR systems is rarely supported, especially in the case of commercial systems. Considering open source EMR systems, then the OpenEMR system has several built-in features, exporting data in user formats:

- available export of patient data record to ASTM compatible CCR data format based on XML;
- individual lab documents forms and data records can be exported in PDF format (rarely XML or CSV);
- each individual patient record regarding demographic data can be exported to an XML dataset;
- several (selected) patient demographic data records can be exported to a CSV file using the Batch Communication Tool (Fig. 2).



BatchCom

Batch Communication Tool

Process: CSV File

Overwrite HIPAA choice : NO

Age From: Up to: And: ● , Or: ○ Gender : Any

And: ● , Or: ○ Appointment within: And :

And: ● , Or: ○ Seen since :

And: ● , Or: ○ Not seen since :

Sort by : Zip Code

Figure 2. Batch export of multiple records into CSV in OpenEMR

The OpenMRS provides more advanced and flexible features, but in the form of client modules:

- i2b2export Module exports OpenMRS data, including patients, providers, concepts, sources of concepts, meetings, observations, and drug orders, to a custom XML schema.
- Cohort Builder Module29 provides export capabilities for patient records, included in the number in different formats (supported XML and CSV).

Therefore, establishing compatibility and communication between different EMR systems is an important task [2]. This leads to the development not only of the above-mentioned international standards, but also of special platforms and frameworks. Among dozens of solutions, we turn our attention to OpenEHR, which is an open domain platform for the development of flexible electronic health systems. Another way of compatibility was offered by NextGen Healthcare using the open source NextGen Connect tool (formerly Mirth Connect). For these reasons, we conclude that an efficient way to train, test, and predict data sets from the EMR system is to export-import medical data records into XML-compatible formats. Such a file is partially or even fully compatible with CCR/CCD (CDA) – depending on the level of acceptance of international standards by the specific EMR system with which the researcher works. At the same time, XML data files have been converted and developed using dozens of libraries (e.g., Python or R).

1.3. Dimension reduction.

In medical research, there are always a large number of dependencies of the investigated. However, many are not statistically significant and in the absence of clinical significance they are not so necessary to be counted and interpreted. How significant from the point of view of the subjects of the study these relationships cannot be assumed in advance. For example, only 5–25% of all possible dependencies are statistically and clinically significant.

Moreover, most indicators may still have dependencies between themselves. By using them as the starting set of indicators, we can set and link the task of building on them more complex, integrating indicators, using, for example, principal component analysis (PCA). Where the number of such complex indicators of the component will be significantly smaller than the number of baseline indicators, the result of this procedure is that the new indicators compactly provide significantly more information. As a result, it is possible to filter out random components and provide more reliable information on the structure of the original indicators of the study groups of patients.

1.4. Classification.

Three main types of task classification are used in medical research:

1. Binary classification (here there can be positive and negative values).
2. Multi faceted classification (the object carries a characteristic of one class).
3. The object belongs to many classes at the same time.

Let 's look at three examples for the following:

1. Identify if person has any disease X . 1 – person is sick, 0 – person is healthy.
2. Identification of a certain disease among diseases A , B , C , D , E .
3. Definition of several diseases among A , B , C , D , E .

Classification algorithms (like the induction of decision tree nodes) automatically divide the value of multiple attributes a_j into two intervals: $a_j \le y_j$, $a_j > y_j$ and into categorical attributes a_k are divided into two subsets: $a_k \in S_k$, $a_k \notin S_k$.

The separation of numerical attributes is generally based on measures, on entropy or Gini Index. The separation process is recursively repeated while the prediction accuracy is improved. The last step involves deleting the model retraining avoidance nodes. As a result, we have to get a set of rules, start from the root tree to each terminal node, including inequalities of numerical attributes and inclusion conditions for categorical attributes.

Results of the research.

Our experimental studies are related to the prediction of fractures for middle-aged women, is today a topical, preventive problem.

We consider the data of clinical, laboratory studies of 1,469 women who were previously stored in the OpenEMR system. The most important group of indicators is the results of bone densitometry. In total, we have $p = 182$ indicators. After you delete spaces in the dataset, there are 1,242 tuples left.

Patients were grouped into three groups according to the presence of fractures. For this purpose we use the attribute of the class «Fractions» with such categorical values: 1 – no fracture, 2 – peripheral fractures, 3 – vertebral fractures.

Applying dimension reduction using the method of PCA, we consider two main components. Namely, *PC1* has 43.4% explanatory variation, *PC2* – 12.5%. That is, these two components are able to explain 55.9% of the data variation. *PCA* results are shown in Figure 3. According to the approach proposed in the task setting, proportional to interest $ExplainedVar(PC1)$ and $ExplainedVar(PC2)$ we choose 8 attributes based on *PC1* and 2 based on *PC2*. Namely «Weight», «TOTAL_Fat», «RIGHT_TOTAL_Fat_g», «LEFT_TOTAL_Fat_g», «RIGHT_TOTAL_Total_Mass_kg», «LEFT_TOTAL_Total_Mass_kg», «TOTAL_Tissue_g», «RIGHT_TOTAL_ Tissue_g» for $ExplainedVar(PC1)$ and «Weight», «TOTAL_Fat» for $ExplainedVar(PC2)$. All attributes are numeric and after decreasing dimension, consider $p_1 = 8$ attributes above. In Figure 4, we can analyze the area $PC1 - PC2$ according to the ellipsoids that represent three groups of patients with predicted fractures. Reducing dimension we see here attributes with the largest variations together with their directions according to patient groups.

The development of the machine learning model was implemented in the R free software environment, using the **mlr** package.

In accordance with Figure 1 we define the machine learning task. In our case, this is a classification task according to the attribute of the «Fractions» class. In terms of the **mlr** packet, this can be described as follows:

```
task  = makeClassifTask(id = "fractures_classification",
            data = df1, target = "Fractures")
```

The next step is to choose a method for solving the machine learning task. We are creating a «lantern» for this purpose. The **mlr** package allows you to create a benchmark of laners

```
lrns <- makeLearners(c("lda","rpart", "C50","rFerns",
         "randomForestSRC"), type = "classif")
```

Here, the benchmark includes 5 methods that can be applied to the classification task. Namely linear discriminant analysis (Ida), rpart, C5.0, random ferns and random forest

```
comparison <- benchmark(tasks = task, learners = lrns,
                  resampling = cv5)
```

Here we define the task of machine learning, laners (methods) and redeployment strategy.

As a result, we get machine learning models based on five methods. The results of the effectiveness of the methods using the mmce measure are presented in Figure 6. We see that the most accurate model according to the estimation of the prediction quality is the random forest method, the worst is the ferms method.

Dimension-induced decision tree after dimension reduction is shown in Figure 5. Here 5 attributes were included, their use is shown in the table.

Table 1

Attribute usage for decision tree in Figure 5

Name of attribute	Usage
RIGHT_TOTAL_Tissue _g	100.00%
RIGHT_TOTAL_Fat_g	14.05%
RIGHT_TOTAL_Total_Mass_kg	14.05%
TOTAL_Fat	11.19%
TOTAL_Tissue_g	2.61%

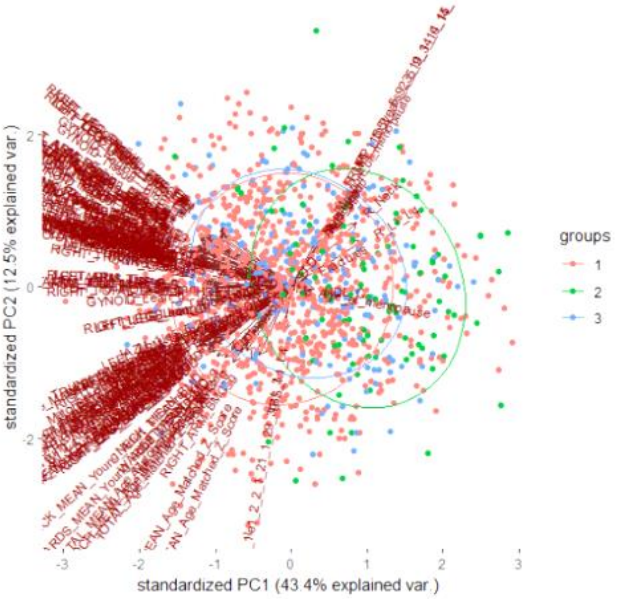


Figure 3. Data in PC1–PC2 plane. Arrows show «rotations» of attributes in $PC1-PC2$ plane

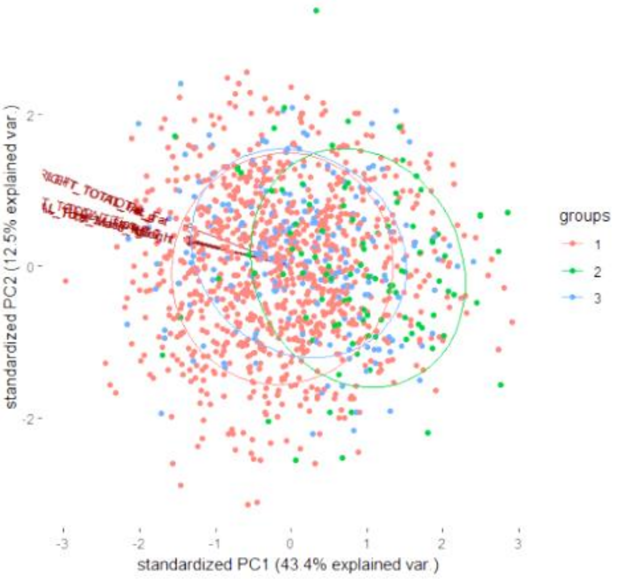


Figure 4. Data in $PC1-PC2$ plane. Arrows show the reduction of dimension and «directions» of changes in patients group

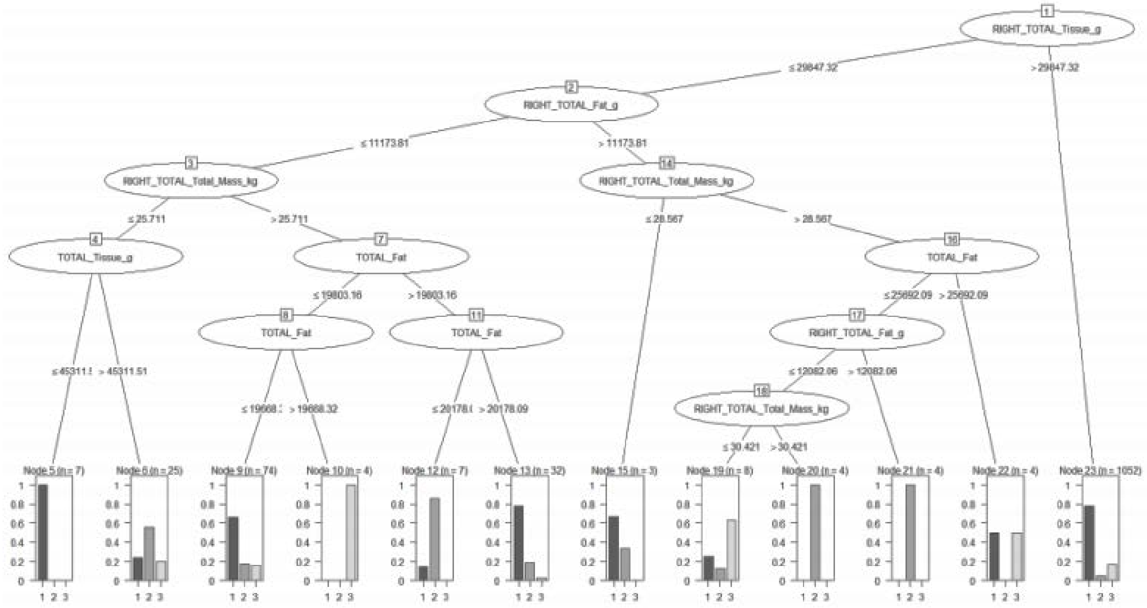


Figure 5. Decision tree, which is constructed on the basis of reduced dimension of data

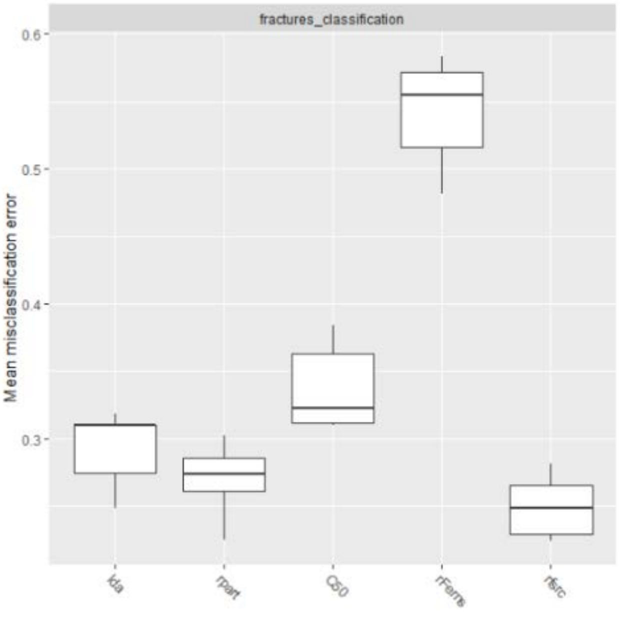


Figure 6. Comparison of performance measures for different classification methods: lda,rpart, C50, rFerns, randomForestSRC

Conclusions. So we introduced an approach to developing a machine learning model for medical research that is based on the use of open source software. The flowchart includes the basic steps of developing a machine learning model, including importing and preparing data, setting a task, choosing a method (laner), adjusting parameters, and evaluating the model. Dimensional problems that often occur in medical research are also highlighted here.

Initially, we are dealing with the export-import of EMR data. This is not a trivial task, even in the case of open source systems, we analyze this and offer solutions in the most well-known open source systems (OpenEMR and OpenMRS), which are predominantly XML based.

Special attention is given to the application of open source software in machine learning. The use of this type of software is critical in many cases, mainly related to scientific medical research for the development of prevention and treatment methods. An approach that uses data from an open source EMR system will later be used on both training and test datasets in free machine learning environments, and this is very promising. As an example, we consider the task of developing a fracture prediction classifier, where we solve all the problems that arise in the block diagram of Figure 1. With the help of a benchmark of laners we have the opportunity to compare different methods of machine learning using them in medical research.

References

1. List of open-source health software. Electronic health or medical record. URL: https://en.wikipedia.org/wiki/List_of_open-_source_health_softwareElectronic_health_or_medical_record (accessed: 2017.11.12).
2. Almeida J., Frade S., Cruz-Correia R. Exporting Data from an openEHR Repository to Standard Formats. Conference on ENTERprise Information Systems / ProjMAN 2014 – International Conference on Project MANagement / HCIST 2014 International Conference on Health and Social Care Information Systems and Technologies. 2014. URL: https://www.sciencedirect.com/science/article/pii/S2212017314003843.
3. Aminpour F., Ahamdi M. Utilization of open source electronic health record around the world: A systematic review. Journal of research in medical sciences: the official journal of Isfahan University of Medical Sciences. 2014. № 19 (1). P. 57.
4. Fritz F., Tilahun B., Dugas M. Success criteria for electronic medical record im- plementations in low-resource settings: a systematic review. Journal of the American Medical Informatics Association. 2015. № 22 (2). P. 479–488. https://doi.org/10.1093/jamia/ocu038
5. Martsenyuk V., Semenets A. On code refactoring for decision making component combined with the open-source medical information system. Advances in Soft and Hard Computing AISC 889. 2019. URL: https://doi.org/10.1007/978-3-030- 03314-9.
6. Martsenyuk V., Vakulenko D., Vakulenko L., Kos-Witkowska A. Information system of arterial oscillography for primary diagnostics of cardiovascu- lar diseases. Computer Information Systems and Industrial Management. CISIM 2018. Lecture Notes in Computer Science. 2018. P. 46–56. URL: https://doi.org/10.1007/978-3-319-99954-85. https://doi.org/10.1007/978-3-319-99954-8_5
7. Reynolds C., Wyatt J. Open source, open standards, and health care information systems. Journal of medical Internet research. 2011. № 13 (1). https://doi.org/10.2196/jmir.1521
8. Semenets A. On organizational and methodological approaches of the emr-systems implementation in public health of Ukraine. Medical Informatics and Engineering. 2013. № 3.
9. Semenets A. About experience of the patient data migration during the open source emr-system implementation. Medical Informatics and Engineering. 2015. № 1.

Список використаної літератури

1. List of open-source health software. Electronic health or medical record. URL: https://en.wikipedia.org/wiki/List_of_open-_source_health_softwareElectronic_health_or_medical_record (accessed: 2017.11.12).
2. Almeida J., Frade S., Cruz-Correia R. Exporting Data from an openEHR Repository to Standard Formats. Conference on ENTERprise Information Systems / ProjMAN 2014 – International Conference on Project MANagement / HCIST 2014 International Conference on Health and Social Care Information Systems and Technologies. 2014. URL: https://www.sciencedirect.com/science/article/pii/S2212017314003843.
3. Aminpour F., Ahamdi M. Utilization of open source electronic health record around the world: A systematic review. Journal of research in medical sciences: the official journal of Isfahan University of Medical Sciences. 2014. № 19 (1). P. 57.
4. Fritz F., Tilahun B., Dugas M. Success criteria for electronic medical record im- plementations in low-resource settings: a systematic review. Journal of the American Medical Informatics Association. 2015. № 22 (2). P. 479–488. https://doi.org/10.1093/jamia/ocu038
5. Martsenyuk V., Semenets A. On code refactoring for decision making component combined with the open-source medical information system. Advances in Soft and Hard Computing AISC 889. 2019. URL: https://doi.org/10.1007/978-3-030- 03314-9.
6. Martsenyuk V., Vakulenko D., Vakulenko L., Kos-Witkowska A. Information system of arterial oscillography for primary diagnostics of cardiovascu- lar diseases. Computer Information Systems and Industrial Management. CISIM 2018. Lecture Notes in Computer Science. 2018. P. 46–56. URL: https://doi.org/10.1007/978-3-319-99954-85. https://doi.org/10.1007/978-3-319-99954-8_5

7. Reynolds C., Wyatt J. Open source, open standards, and health care information systems. Journal of medical Internet research. 2011. № 13 (1). https://doi.org/10.2196/jmir.1521
8. Semenets A. On organizational and methodological approaches of the emr-systems implementation in public health of Ukraine. Medical Informatics and Engineering. 2013. № 3.
9. Semenets A. About experience of the patient data migration during the open source emr-system implementation. Medical Informatics and Engineering. 2015. № 1.

УДК 004.021

ПІДХІД ДО РОЗВ’ЯЗУВАННЯ ЗАДАЧ МАШИННОГО НАВЧАННЯ ІНТЕГРОВАНИХ З ДАНИМИ СИСТЕМ З ВІДКРИТИМ КОДОМ ЕЛЕКТРОННИХ МЕДИЧНИХ ЗАПИСІВ

Василь Марценюк[1]; Назар Мілян[2]

[1]Університет Бельсько-Бяла, Бельсько-Бяла, Польща
[2]Тернопільський національний технічний університет імені Івана Пулюя, Тернопіль, Україна

Резюме. *В останні десятиліття активно розробляються рішення з охорони здоров'я на основі програмного забезпечення з відкритим кодом, а також комерційні засоби. Найпоширенішими системами електронного медичного обліку з відкритим кодом є WorldVistA, OpenEMR та OpenMRS. Вчені звертали увагу на перспективи програмного забезпечення електронних медичних записів з відкритим кодом та безкоштовних систем для країн з певними фінансовими труднощами й таких, що розвиваються. Постановка задачі машинного навчання в медичних дослідженнях здійснена. На блок-схемі, представленій у роботі, продемонстровано основні кроки для розроблення моделі машинного навчання. Звернено увагу, що задача імпорту тренінгових, тестувальних і прогнозувальних наборів даних із систем EMR у середовищі машинного навчання є не такою тривіальною через ряд причин, які розглянуто в дослідженні. Наведено кілька основних підходів для доступу до даних медичних записів пацієнтів у типових системах EMR. Представлено деякі особливості застосування підходів для двох найпоширеніших систем із відкритим кодом EMR: OpenEMR, OpenMRS. Попри тривалий період розроблення й застосувань, навіть провідні й широко розповсюджені EMR системи (як комерційні, так і безкоштовні з відкритим кодом) мають обмежену або часткову підтримку можливостей HL7. Не зважаючи на проблеми, які розглядають рівні реалізації, існує достатньо аргументів для адаптації використання форматів даних сумісних з HL7 і розроблення інформаційних систем, які орієнтовані на машинне навчання. Експериментальні дослідження, пов’язані з прогнозуванням переломів для жінок середнього віку, підтверджують, що це є на сьогодні актуальною, профілактичною проблемою. Розроблення моделі машинного навчання реалізована в середовищі вільного програмного забезпечення R за допомогою пакета mlr. У результаті отримуємо моделі машинного навчання на основі п’яти методів. Результати ефективності методів, за допомогою міри mmce, показують, що найточнішою моделлю відповідно до оцінювання якості прогнозування є метод випадкового лісу (random forest), найгіршим є метод ferms.*

Ключові слова: *машинне навчання, PCA, класифікація, EMR система, mlr.*

https://doi.org/10.33108/visnyk_tntu2019.03.105 *Отримано 06.10.2019*