Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 121 (2017) 852–858

Procedia Computer Science

www.elsevier.com/locate/procedia

CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies, CENTERIS / ProjMAN / HCist 2017, 8-10 November 2017, Barcelona, Spain

Towards Identifying of Effective Personalized Antihypertensive Treatment Rules from Electronic Health Records Data Using Classification Methods: Initial Model

Anna Semakova[a,]*, Nadezhda Zvartau[a,b], Klavdiya Bochenina[a], Aleksandra Konradi[a,b]

[a]*ITMO University, 4 Birzhevaya liniya, Saint-Petersburg 199034, Russia*
[b]*Almazov Federal North-West Medical Research Center, 2 Akkuratova street, Saint-Petersburg 197341, Russia*

Abstract

Traditional clinical diagnosis and management are regulated by standards, patient management protocols with specific nosology and clinical guidelines that are limited, and their use in practice is confronted with the gap between "efficacy" and "effectiveness". Data stored patients' electronic health records (EHRs) provide previously unknown predictors that have affected the disease outcome, and allow to develop personalized treatment guidelines with the application of statistical methods and powerful machine learning techniques. This study aims to predict treatment effect of monotherapy with five main classes of antihypertensive drugs based on individual patients' profiles for a single decision time point. We transform the estimation of effective personalized antihypertensive treatment rules into a classification problem, and propose the method to adapt the CART algorithm for building a decision tree for effective personalized approach to choose monotherapy in hypertension.

Peer-review under responsibility of the scientific committee of the CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies.

* Corresponding author. Tel.: +7-911-007-8539.
E-mail address: a.a.semakova@gmail.com

1877-0509
Peer-review under responsibility of the scientific committee of the CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies.
10.1016/j.procs.2017.11.110

Keywords: Electronic health records; Personalized medicine; Classification problem; Decision trees; Model quality measures.

1. Introduction

Patient management is based on specific disease guidelines mostly incorporating results of randomized clinical trials and their meta-analysis. However, results of trials have numerous limitations and may not be generalizable to general population. That is why real-world treatment effects are often far from expected. There is the well-known gap between treatment "efficacy" (capacity to produce an effect in expert hands under ideal circumstances) and "effectiveness" (capability to produce a desired result in real life)[1]. Thus identification of patients subgroups with differential treatment risks and benefits based on data-driven approach is necessary for P4 medicine (predictive, preventive, personalized, and participatory).

Data collected through medical information systems (MIS) enable to meet the promise of personalized medicine. Data stored patients' electronic health records (EHRs) provide previously unknown predictors that have affected the disease outcome, and allow to develop personalized treatment guidelines[2,3]. Statistical methods and powerful machine learning techniques for estimating optimal medical treatment and chronic diseases management strategy according to patient-specific characteristics are currently receiving extensive attention[4,5,6].

In this study, we adapt the CART algorithm that would allow to guide selection of the effective antihypertensive treatment for particular patients based on the following baseline characteristics from patients' EHRs: blood pressure level, sex, age, body mass index (BMI), smoke, hereditary, dyslipidemia, diabetes, impaired glucose tolerance (IGT), left ventricular hypertrophy (LVH), microalbuminuria, chronic kidney disease stage (CKD), chronic heart failure (CHF), ischemic heart disease (IHD). This method is the most useful for identification of variable interactions and may be easier to use in clinical settings because it requires evaluation of simple decision rules[7]. The motivation for the hypertensive patients selection as an experimental prototype is arterial hypertension (AH) which provides life-long therapy and large bulk of treatment options without clear differences in efficacy with the available results of short-term studies on "refined" populations poorly extrapolated to routine clinical practice. In addition, hypertension is prevalent (46% of world population)[8], closely linked to excess cardiovascular morbidity and mortality, various treatments are available, however blood pressure control rate is low and rarely exceeds 37%[8].

In current study, we extracted EHRs data of 5438 patients referred for hypertension counseling to the Almazov Federal North-West Medical Research Centre for six years period (from 2010 to 2015). We extracted 1086 records for monotherapy with five main classes of antihypertensive drugs (beta-blockers, ACE inhibitors, angiotensin II receptor blockers, calcium channel blockers, diuretics) from EHRs data. By way of the initial experiment of the ongoing research we predict tailored antihypertensive monotherapy by virtue of the fact that we included only patients with an initial diagnosis of AH, which means that antihypertensive treatment is started with one of recommended drugs classes by clinical guidelines, as drop in blood pressure may lead to cardiac risk[9].

2. Problem Statement

3.1. Problem Formalization

Let $X_j(t_i) = \left\{x_j^1(t_i),...,x_j^m(t_i)\right\}, j = 1,...,k; i = 1,...,n$ is the set of possible variables' features for a subject (patient) $X_j(t_i)$ collected before the initiation of a treatment at the time t_i. $Y_j(X_j(t_i)) = \left\{y_j^1,...,y_j^l\right\}$ denotes the set of antihypertensive treatment assigned to patient $X_j(t_i)$ at the time t_i.

The goal of estimating effective personalized antihypertensive treatment rules is to create a data-driven predictive model based on individual patients' features so that in patients who follow the simulated rules blood pressure will be controlled after one month of treatment (1).

$$f : x_j^1(t_i) \times ... \times x_j^m(t_i) \rightarrow \left\{y_j^1,...,y_j^l\right\}. \quad (1)$$

We introduce a criterion of effectiveness for the assigned antihypertensive treatment (it has to be maximized): the

probability of reaching the target blood pressure level after the treatment (2). Also, we define constraints for blood pressure level and time interval between outpatient visits as (3).

$$P\left[x_j^s(t_{i+1}) < 140, x_j^d(t_{i+1}) < 90\right] \to \max. \tag{2}$$

$$\begin{cases} \left[x_j^s(t_i) \geq 140, \right. \\ \left[x_j^d(t_i) \geq 90, \right. \; where \\ t_{i+1} - t_i > 28, \end{cases} \tag{3}$$

$x_j^s(t_i)$ – patient's systolic blood pressure level at the time t_i;
$x_j^d(t_i)$ – patient's diastolic blood pressure level at the time t_i.

The criterion for the assigned antihypertensive treatment effectiveness (2) and constraints (3) are created based on clinical prerequisites, as follows:
– all subjects (patients) are hypertensive patients, so systolic blood pressure level is higher than140 mm Hg and/or diastolic blood pressure level is higher than 90 mm Hg;
– antihypertensive drugs have the "accumulation" effect. This means that initial response to antihypertensive treatment is assessed after one month of treatment (28 days);
– the clinical criterion for antihypertensive treatment effectiveness is the achievement of target blood pressure level: systolic blood pressure level is less than140 mm Hg and diastolic blood pressure level is less than 90 mm Hg.

3.2. Classification problem

We transform the estimation of effective personalized antihypertensive treatment rules into a classification problem. Let X denotes the set of possible features x^i for $i = 1,\ldots, n$. The aim of the classification task is to find a classifier $f: x^1 \times \ldots \times x^n \to \{y_1, \ldots, y_l\}$ with a loss function:

$$E(f, y, x^i) = \sum_{k=1}^{s} \left\| f_k - y_k \right\|^2 \to \min, where$$

s is the number of samples for test dataset.

Learning the classifier f is based on the training dataset $T \subset x^1 \times \ldots \times x^n \times \{y_1, \ldots, y_l\}$. The training dataset T consists of m samples:

$$t_j = (\upsilon_j, y_j) = \left(\left[\upsilon_j^1, \ldots, \upsilon_j^n \right], y_j \right), j = 1, \ldots, m, where$$

– $\upsilon_j^i \in x^i$ is a value of feature x^i for sample t_j;
– $y_j \in \{y_1, \ldots, y_l\}$ is a label of sample t_j.

A label of sample t_j of training dataset T is defined using the following expression:

$$y_j = \begin{bmatrix} 1, t_j \in \left[\upsilon_j^s < 140 \cap \upsilon_j^d < 90 \right] \\ 0, t_j \in \left[\upsilon_j^s \geq 140 \cup \upsilon_j^d \geq 90 \right] \end{bmatrix}, j = 1, \ldots, m, where$$

υ_j^s – patient's systolic blood pressure level;
υ_j^d – patient's diastolic blood pressure level.

To summarize, our approach fits the models predicted effectiveness (label is 1) or ineffectiveness (label is 0) of each class of antihypertensive drugs for patients based on their individual features for a single decision time point.

3. Methods

3.1. EHRs Description and Data Preprocessing

EHR represents the case history of a patient and contains information about all patient visits to the cardiological centre, complaints, examination results, and prescribed investigations and treatment. In our study, we included only patients with an initial diagnosis of AH. It means that high blood pressure was the main reason for the cardiological centre referral. We have applied current hypertension guidelines to select the set of precedents. Patients' features associated with hypertension stage and risk of cardiovascular complications were included as predictors. Antihypertensive treatment effectiveness criterion served as clinical outcome of treatment for a patient.

A clinical text is highly heterogeneous. It doesn't always conform to normal grammar and is rich in author-and domain-specific idiosyncrasies, abbreviations and acronyms, as well as spelling and typing errors[10]. Also EHRs may contain incomplete and inconsistent data. Thus, capturing the context of prescribed investigations and treatment is a central challenge for clinical text mining. According to text syntactic analysis we have extracted patterns for drugs assigned to patients[11]. The next step of data preprocessing was extraction of drugs and selection of drugs names with spelling or typing errors using Damerau–Levenshtein algorithm. After that, we have developed drug vocabulary based on extracted drugs. The vocabulary has hierarchical organization and provides structuring by brand-name drug, by International Nonproprietary Name (INN), and by pharmacological class.

3.2. Model dataflow architecture

In our study, we propose the method to adapt the CART algorithm for building effective personalized antihypertensive treatment rules decision tree. Principal goal of machine learning algorithms is generalizability, that means precise prediction capacity confirmed in new datasets. Consequently, in order to avoid overfitting, it's necessary to set optimal parameters of model. Our proposed method performs the following three steps.

Step 1. Use the aforementioned patients' features and treatment outcome to create training and validation set constituting 25% of the sample set.

Step 2. Compute decision tree optimal parameters using 3-fold cross validation: the CART algorithm is fitted on the training samples and the model value function, as *accuracy* which is the ratio of instances correctly classified is computed on the testing sample.

Step 3. Evaluate model predictive power on the validation set using following quality metrics: the sensitivity or true positive rate and the specificity or true negative rate.

These metrics are calculated by using the values true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) instances written in the confusion matrix. The sensitivity being the portion of actual positives predicted as positives can be calculated using the following expression:

$$Sensitivity = \frac{TP}{TP + FN}.$$

The specificity, which is the portion of actual negatives predicted as negative, can be calculated using the following expression:

$$Specificity = \frac{TN}{TN + FP}.$$

We present model dataflow architecture diagram in Figure 1.

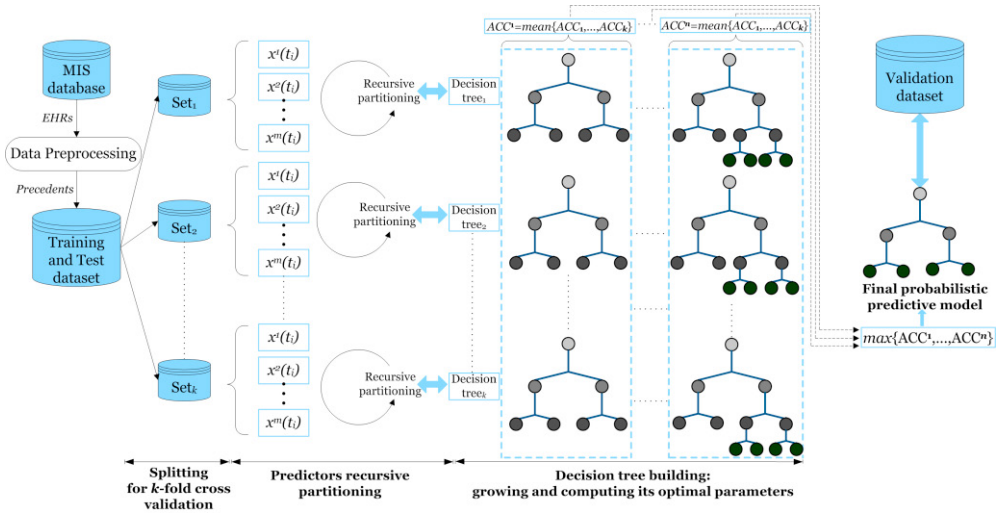

Fig. 1. Model dataflow architecture diagram.

4. Initial Results

The CART algorithm[12] was fitted using the patients' feature values for each class of antihypertensive drugs. Let Q is the set of samples at node m. For each candidate split $\theta = (j, t_m)$ consisting of a feature j and threshold t_m the data is partitioned into $Q_{left}(\theta)$ и $Q_{right}(\theta)$ subsets:

$$Q_{left}(\theta) = (x, y) \mid x_j \le t_m,$$
$$Q_{right}(\theta) = Q \setminus Q_{left}(\theta).$$

The impurity at node m is computed using an impurity function $H()$. In this work, we implement Gini impurity and entropy for the information gain.

1. ***Gini impurity*** is calculated as follows:

$$H(Q_m) = 1 - \sum_{k=1}^{l} p_{mk}^2,$$

where p_{mk} is Q_m dataset k-label probability.

2. ***Entropy*** is given by:

$$H(Q_m) = -\sum_{k=1}^{l} p_{mk} \log_2 (p_{mk}).$$

The criterion for the parameters optimal choice (θ^*) is minimizing of the objective function $G(Q, \theta)$:

$$G(Q, \theta) = \frac{n_{left}}{N_m} H(Q_{left}(\theta)) + \frac{n_{right}}{N_m} H(Q_{right}(\theta)) \to \min,$$

where N_m is number of samples at parent node; n_{left} and n_{right} are the number of samples at child nodes. Nodes are expanded until all terminal nodes (leaves) are pure or until all leaves contain less than maximum allowed decision tree depth ($N_m < min_{samples}$ or $N_m = 1$). In our approach, if decision tree accuracy is equal to local maximum among experimental values then this tree level is optimal decision tree depth. After that, the predictor is given a label of the leaf subset elements of the decision tree.

We present tree depth with a highest accuracy computed on the testing sample, obtained sensitivity and specificity

values of classifiers in Table I.

Table 1. Quality measures of constructed trees.

Tree optimal depth; quality metrics	Beta-blockers	ACE inhibitors	Angiotensin receptor blockers	Calcium channel blockers	Diuretics
Implementing predictor splitting criterion using Gini impurity					
Tree depth	4	4	4	4	5
Accuracy	0.63	0.58	0.59	0.56	0.5
Sensitivity	0.62	0.12	0.19	0.33	0.25
Specificity	0.56	0.9	0.92	0.71	0.6
Implementing predictor splitting criterion using entropy					
Tree depth	8	8	7	8	5
Accuracy	0.59	0.61	0.61	0.51	0.5
Sensitivity	0.46	0.59	0.3	0.33	0.5
Specificity	0.75	0.56	0.85	0.52	0.4

Obtained high specificity values represent a strong ability of constructed decision trees to identify treatment ineffectiveness, however, obtained low sensitivity values mean that the classifiers feebly determine treatment effectiveness.

For illustrative purposes, we present the 4-level decision tree identifying beta-blockers monotherapy effectiveness or ineffectiveness for patient based on his/her individual features (Figure 2). The most important tailoring features are systolic blood pressure level (*Gini importance* = 0.36), age (*Gini importance* = 0.29), BMI (*Gini importance* = 0.18), hereditary (*Gini importance* = 0.05), dyslipidemia (*Gini importance* = 0.08), and CHF (*Gini importance* = 0.04).

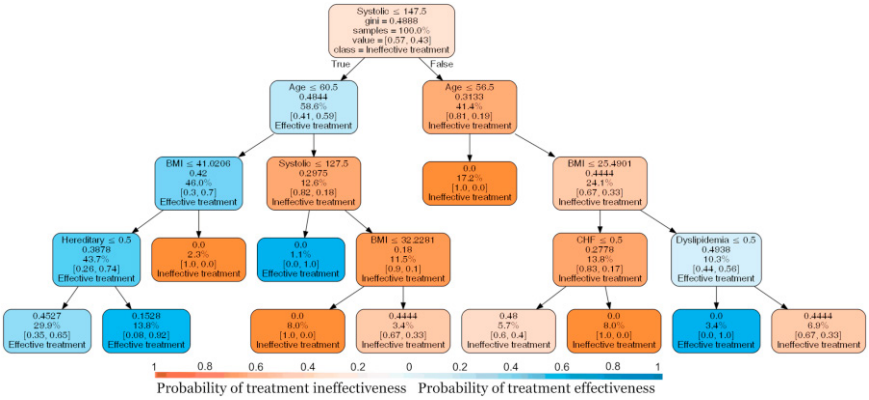

Fig. 2. Tree structure for the scenario of beta-blockers monotherapy prescription with effectiveness-covariate interactions.

5. Conclusion and Future works

In summary, results demonstrated that the constructed decision trees have a good capacity to detect the treatment ineffectiveness instances, but have a feeble capacity to determine treatment effectiveness instances. Consequently, future research will focus on calibration of models sensitivity using ensemble methods, such as bagging[13], boosting[14], random forests[15], stacking[16], for aggregating several competing models. Lastly, fuzzy sets application for constructing a fuzzy decision tree is of interest[17].

Also future probabilistic predictive model should provide a tool for simultaneous identification of personalized antihypertensive treatment rules associated with good and bed response to monotherapy and drugs combinations

treatment outcome. Semantic data integration from heterogeneous sources (24-hour blood pressure monitoring, ambulatory and home blood pressure monitoring) may be used to increase the power of precedents set (knowledge base). We will continue to extend research to multiple-stage settings where the antihypertensive treatment strategies are dynamically adapted depending on patient's time-varying features and simulate effective treatment sequences.

Acknowledgements

This research is financially supported by The Russian Scientific Foundation, Agreement #14-11-00823 (15.07.2014).

References

1. Tinetti ME, Studenski SA. Comparative effectiveness research and patients with multiple chronic conditions. *New England Journal of Medicine* 2011;**364**:2478-81.
2. Liao WL, Tsa FJ. Personalized medicine: a paradigm shift in healthcare. *BioMedicine* 2013;**3**:66-72.
3. Hamburg MA, Collins FS. The path to personalized medicine. *New England Journal of Medicine* 2011;**363(4)**:301-4.
4. Murphy SA, Oslin DW, Rush AJ, Zhu J. Methodological challenges in constructing effective treatment sequences for chronic psychiatric disorders. *Neuropsychopharmacology* 2007;**32(2)**:257-62.
5. Wang Y, Wu P, Liu Y, Weng C, Zeng D. Learning optimal individualized treatment rules from electronic health record data. *Proceedings - 2016 IEEE International Conference on Healthcare Informatics, ICHI 2016*; p. 65-71.
6. Tsai WM, Zhang H, Buta E, O'Malley S, Gueorguieva R. A modified classification tree method for personalized medicine decisions. *Statistics and its Interface* 2016;**9(2)**:239-53.
7. Zhang H, Singer B. *Recursive Partitioning and Applications*. New York: Springer; 2010.
8. Mills KT, et al. Global disparities of hypertension prevalence and control. *Circulation* 2016;**134(6)**:441-450.
9. Mancia G, et al. 2013 ESH/ESC guidelines for the management of arterial hypertension: the task force for the management of arterial hypertension of the European Society of Hypertension (ESH) and of the European Society of Cardiology (ESC). *Eur. Heart J* 2013;**34(28)**:2159-219.
10. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics* 2012;**13(6)**:395-405.
11. Friedman C. Semantic *Text Parsing for Patient Records*. Springer US, 2005. p. 423-48.
12. Chou PA. Optimal Partitioning for Classification and Regression Trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1991;**13(4)**:340-54.
13. Breiman L. Bagging predictors. *Machine Learning* 1996;**24(2)**:123-40.
14. Schapire RE. A brief introduction to boosting. *IJCAI International Joint Conference on Artificial Intelligence* 1999;**2**:1401-6.
15. Breiman L. Random Forests. *Machine Learning* 2001;**45(1)**:5-32.
16. Wolpert DH. Stacked generalization. *Neural Networks* 1992;**5(2)**:241-59.
17. Wang TC, Lee HD. Constructing a fuzzy decision tree by integrating fuzzy sets and entropy. *WSEAS Transactions on Information Science and Applications* 2006;**3(8)**:1547-52.