DOI: 10.17587/it.26.701-705

**I. Lakman**[1,2]**,** PhD, Assotiate Professor, e-mail: lackmania@mail.ru,
**R. Akhmetvaleev**[1]**,** Data Analyst, **D. Enikeev**[1]**,** Data Analyst, **R. Khaziakhmetov**[1]**,** Software Engineer,
**O. Chernenko**[3]**,** Cand. Medical Sciences, Deputy Director for Development,
[1]LLC Lexema, Ufa, 450104, Russian Federation,
[2]Institute of Economics, Finance and Business, Bashkir State University, Ufa, 450076, Russian Federation,
[3]LLC "Laboratory of hemodialysis", Ufa, 450083, Russian Federation

# Similarity Learning Algorithm Selection for Chronic Renal Failure Patients Treatment Strategy Optimization

*One of the main methods on which the personalized approach in medicine is based is finding a pair of patients who are similar in the properties of the disease. The objective of the study is to select the most effective similarity learning instrument amongst three options anaemia treatment and phosphorus-calcium balance recovery in dialysis patients, ranked according to the highest similarity to the particular patient. As soon as methods for comparing instruments will achieve the goal, the algorithm of weight tagging is used, modified by the authors by adding more weights values to important features — the cosine measure, the soft cosine measure, considering the similarity of drug alternative and their bioavailability. As a metric that evaluates the quality of algorithms, a combined metric is used that takes into account the quality of treatment classification as effective and the rank order of the greatest correspondence of therapy to a specific patient. As a result, using the opinions of nephrologists as experts, it was shown that the best measure of similarity is the soft cosine measure.*

***Keywords:*** *similarity learning, cosine measure, soft cosine measure, dialysis*

***Acknowledgments.*** The study "Development of an intelligent decision support system for prescribing personalized dialysis and drug therapy to patients with chronic renal failure using artificial intelligence algorithms" was supported by a grant No. AAAA-A20-120011490126-5 from the Foundation for the Promotion of the Development of Small Forms of Enterprises in the Scientific and Technical Sphere.

УДК 004.89:004.912 DOI: 10.17587/it.26.701-705

**И. А. Лакман**[1,2]**,** канд. техн. наук, доц., e-mail: lackmania@mail.ru,
**Р. Р. Ахметвалеев**[1]**,** специалист по анализу данных, e-mail: r_akhmetvaleev@lexema.ru,
**Д. И. Еникеев**[1]**,** специалист по анализу данных, e-mail: enikeev_di@lexema.ru,
**Р. Р. Хазиахметов**[1]**,** инженер-программист, e-mail r_haziahmetov@lexema.ru,
**О. В. Черненко**[3]**,** канд. мед. наук, заместитель директора по развитию, e-mail och@dializrb.ru,
[1]ООО "Лексема", г. Уфа,
[2]Институт экономики финансов и бизнеса, Башкирский государственный университет, г. Уфа,
[3]ООО "Лаборатория гемодиализа", г. Уфа

# Выбор алгоритма обучения подобия для оптимизации стратегии лечения пациентов с хронической почечной недостаточностью

*Один из основных методов, на которых базируется индивидуальный подход в медицине, это поиск пары пациентов, схожих по свойствам заболевания. Цель исследования — выбрать наиболее эффективный инструмент определения подобия для выбора трех вариантов лечения анемии и восстановления фосфорно-кальциевого обмена у диализных пациентов, ранжированных в соответствии с максимальным подобием с конкретным пациентом. В качестве методов сравнения вариантов лечения для достижения цели используется алгоритм весовой маркировки, модифицированный авторами путем присвоения весов более важным характеристикам, косинусная мера, мягкая косинусная мера, с учетом сходства аналогов препаратов и их биодоступности. В качестве метрики, оценивающей качество алгоритмов, используется комбинированная метрика, которая учитывает качество классификации терапии как эффективной и порядок ранжирования наибольшего соответствия терапии конкретному пациенту. В результате, используя мнения нефрологов как экспертов, было показано, что лучшей мерой сходства является мягкая косинусная мера.*

***Ключевые слова:*** *обучение подобия, косинусная мера, мягкая косинусная мера*

## Introduction

For the last ten years methods and tools of artificial intelligence have started to be widely used in medicine problems, including medical decision support systems design, medical manipulations outcome prediction, recognition of medical images of different types, etc. It is worth to pay attention to a possibility of developing the personalized medicine with the help of machine learning algorithms. One of the principal methods which personalized approach is based on is search for a pair of patients, similar to each other in a terms of disease course properties. This is essential for implementation of effective therapy, which allows to cure the patient or to relief patient's suffering from illness, to the other patient, similar to the first one. Recently, this approach became preferable as a direction in machine learning tools development for personalized medicine problem solving [1, 2].

The principal method for the patient similarity is called similarity learning. The general formulation of the tasks of the corresponding problems can be expressed as follows. Suppose, we have records from dataset in the form of feature vectors, each of those vectors corresponds to the object of the training sample. Every unique matching of two vectors from initial dataset are submitting onto income of similarity estimation model. In the case of pair was labeled as "similar", such pair is assigned to 1, if pair was labeled as "dissimilar", such pair is assigned to 0. This way, new dataset is formed with one dependable variable, and the size of newly formed dataset, provided that it consisted of unique elements, would be $2 \cdot (n - 1)$, s. t. n is the number of elements in the initial dataset. Note that determining whether objects in a pair are "similar" or "dissimilar" can be done according to the rules obtained using distance metrics based on unsupervised learning algorithms or based on expert opinion.

In the problems of personalized and predictive medicine the objects are patients. The search among the available retrospective data for the patient, according to the estimates most closest to the current allows the attending physician to correctly diagnose the disease and to prescribe the optimal treatment.

There are studies dedicated to the reasoning behind the selection of distance metrics for generating a dataset containing information about similarity of two patients. For example, in [3] a comparative review of the results of clustering using Minkowski distance, Euclidean distance, cosine measure and chi-square based distance conducted in datasets containing the values of numerical, categorial and mixed data types is given. In [4] metrics of Minkowski distance, Euclidean distance, Manhattan distance, Hamming distance are compared with each other on the results given on data of hierarchical type from THIN dataset. In [5] on the example of clusterization of binary metrics of Minkowski distance, Euclidean distance, squared Euclidian distance, Manhattan distance, Hamming distance are compared including influence of mentioned above metrics on specificity and sensitivity of classifier.

Medical images data is staying apart; for that problem, the main distance metrics usually are the Jaccard and Dice measures, as for example in [6], the methods of operation of the algorithms segmentation of medical images have been described.

Today the volume of data is growing, and this growth is corresponding raise of a need of computing powers, so, consequently the most powerful tools nowadays for a data scientist are deep learning neural networks (DLNN). As for example, in [7] the learning object is a matching matrix, that fitting on vectors, that are a result of consequent operations of convolution and pooling, performed by DLNN layers on two feature vectors of formalized anamnesis data. The anamnesis record in this research is represented as combination of date and event in patient's medical history. Very similar solution is presented in [8], the research demonstrates methods of semantic EHR data analysis.

Such variation of the similarity learning method as locally sensitive hashing, the essence of which is to select such hash functions that work on vectors in the feature space so that objects similar to each other are getting more likely to belong to the same class is used in [9] and [10].

Based on sources analysis and problem of effective therapy selection for anemia treatment and phosphorus calcium balance recovery for dialysis patients the objective of this research was determined. **The objective** of current research is to develop a method of selection of the most effective similarity learning tool for forming of the list of top 3 treatment options, ranking by similarity measure to the given patient.

## Materials and methods

In this study based on collected data on dialysis patients (social-demographic parameters, results of clinical diagnostic studies, medicines prescribed on previous stages of therapy correction, dosage and route of administration) 3 options of appropriate of anti-anemic therapy and PCBR therapy is suggested as a solution. The essence of the solution is to search for an optimal therapy from the list of all the therapies from EHR database. The search for the best is considered amongst 3 methods, which are:

Feature vectors coding with following Manhattan distance calculation considering feature weight ("weighted matching").

Vector comparison via soft cosine measure for considering an opportunity of using analogue medicines in prescribed therapies.

For feature vectors formation the following information was used:

Demographic parameters (gender, age);

Physical parameters (height, weight, body mass index — required for adequate therapy medicine dosage);

Patient blood test results (for anti-anemic therapy — hemoglobin, ferritin, transferrin, etc., for PCBR therapy — parathyroid hormone, phosphorus, calcium, etc.);

Previous and pre-previous values of blood test results;

Previous prescribed therapy (for possibility of cutting contraindications and precise search more appropriate medicine for a patient).

Supervised machine learning quality algorithm development is based on correctly labeled database. In current research physicians / nephrologists have labeled more 9000 patients' records labeled as either as "effective" or "non-effective" or "excessive". The detailed information on the dataset, used in the research is represented in [11]. Corresponding labeled records are stored in a form of database, which is also a source for "patient — prescribed treatment" pairs.

The problem of similarity learning for personalized medicine in general case could be formulated as search of likelihood function, which could measure similarity between features of any two patients. In other words, it is necessary to found function $S(x_i, x_j)$, determining similarity between vector $x_i$ of features of patient $i$ vector $x_j$ of features of patient $j$.

$$S(x_i, x_j) = x_i^{\mathrm{T}} M x_j,$$

where $x_i$ — vector of features of patient $i$; $x_j$ — vector of features of patient $j$; $M$ is the correspondence matrix.

As for first method of similarity search modified matching was used and called by the authors of this research "weighted matching". The algorithm could be represented as a sequence of steps.

On the first stage weights of features are calculated, which can then be used for two different records matching. For this logistic regression is built for all of the patient features and prescribed therapy, which could allow to estimate influence of every feature on efficiency of treatment. Having expert opinion of physicians and nephrologists is helpful for correcting of weights if it is needed.

Expert opinion of doctors allows to get groups of factors values to take into consideration nonlinear influence of the factor on difference between vectors. So, the difference between erythropoietin dosages of 10000 units and 16000 units is more significant then difference between 68000 units and 78000 units.

Next step, using formed groups the value of the feature is coded via its digit number. For example, hemoglobin value in the range in between from 100 gr per liter to 110 gr per liter is coded as "E". Using coding like that, model could generalize opinions of several experts. As the outcome of this operation, the feature vector could be coded as:

"BBADFBAEFBFHHFADBAJLFBUAAAACIINR",

s.t every position in this code stands for a certain feature.

On the next step mutual similarity between code positions of two code vectors is calculated. For this every code position is replaced by its UNICODE digit equivalent. This number multiplied on feature weight that was calculated n the first stage. Then ratio of records similarity is calculated.

After likelihood calculation all the records with similarity ratio lesser than 0.8 are deleted from result space. Amongst the rest top 3 therapies are selected from the list with contraindication medicines cutted.

For the second method number vectors, formed from initial set of features, responsible for "patient-prescribed therapy" matching. The data has to be preprocessed, since it can include outliers, and could be normalized. In order to do that the Robust Scaler method was used. It is very similar to Standard Scaler Algorithm, that transforms data such way that for every feature taken mean value will be equal to 0 and dispersion value will be equal to 1, as the result all of the features will have the same scale. However, Robust Scaler uses medians and quartiles instead of mean and dispersion by the following formula:

$$\frac{x_i - Q_1(x)}{Q_3(x) - Q_1(x)},$$

s.t. $x_i$ — element of series $x$, $Q_1(x)$ and $Q_3(x)$ first and third quartiles of initial series.

It is important to notice, that initial vector, the one search for similar records is conducted for, is also needed to be preprocessed the same way, that the rest of the data had been preprocesed. Then for transformed vectors similarity cosine measure with initial record is calculated. For the initial vector A and all of the vectors B from transformed data measure could be calculated by the following formula:

$$cosine = \frac{A \cdot B}{AB} = \frac{\sum_{i=1}^{n} A_i \cdot B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \cdot \sqrt{\sum_{i=1}^{n} (B_i)^2}},$$

where $A$ and $B$ are vectors of feature.

Amongst all of the records only those are selected, whose cosine measure is greater than 0.8, or in other words vectors similarity is greater than 80 %. After that the selection of top 3 contraindication free records has taken place.

Soft cosine measure feature is in possibility of considering the influence of medicines analogues. For this the data on medicines similarity in per cent ratio was presented by doctors. Similarity matrix $s$ was made up, consisting of similarity measures between therapy medicines $i$ and $j$ with their bio availability considering via route of administration (oral, intravenous, intramuscular, subcutaneous, infusion). After that soft cosine measure is calculated by the following formula:

$$soft\_cosine(a,b) = \frac{\sum_{i,j}^{N} s_{ij} a_i b_j}{\sqrt{\sum_{i,j}^{N} s_{ij} a_i b_j} \sqrt{\sum_{i,j}^{N} s_{ij} a_i b_j}},$$

where $A = (a_i)$ and $B = (b_j)$ are vectors of feature, $s = (s_{ij})$ — feature similarity matrix.

Next, like in the previous stage, records with the similarity below 0.8 are deleted from dataset, after that records with the patient personal contraindication are deleted from dataset.

One of the main principles of the machine learning algorithm selection is learning quality metrics correct selection. For determination, which one from considering algorithms allows to select optimal therapy for a patient, the machine learning models described above were implemented into 24 hemodialysis centers, and for the same reason, the possibility to estimate selection of therapies was given to three nephrologists. They were asked to estimate if the suggested therapy for a patient was appropriate. In the case of therapy scheme was appropriate, they were asked to rank three selected for every patient therapies using rationality of prescription criteria. To consider all of it the new metrics should be introduced. As for foundation for the new metric the common metrics of precision, recall and specificity were used. This way confusion matrix was created for every selected therapy, s.t TP — true positive rate, TN — true negative rate, FP — false positive rate, FN — false negative rate. It is important to notice, that therapy was considered truly effective in simple majority voting of the doctors for this therapy, in the opposite case therapy was considered non-appropriate for patient. Based on confusion matrix metrics could be calculated via following formulas:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN},$$

$$recall = \frac{TP}{TP + FN},$$

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

This metrics accumulating on all of selected records.

However, for every single patient the algorithm suggests three appropriate therapies, ranked by similarity measure with "patient profile". Thereby metrics was improved considering assigned ranks. For every therapy the number of inversions was calculated in the following order, so if algorithm selects the therapy as the most appropriate for a patient, and at the same time doctor expert indicates its rank as third best, then inversion for this particular record will be 2. This way, every of therapy, estimated by doctors as effective would transforms considering the number of inversions. So, the value equals to number of inversions would multiply on 0.25 and then subtracted from initial number of points of TP metrics.

## Results

For forming therapy database 6693 records of antianemia therapies and PCBR therapy were used that doctors labelled as "effective".

Testing was produced by collecting feedback from doctors after using of learning algorithms on 661 records dated from 01.10.2019 to 31.12.2019. In that case therapy recognized as effective by simple experts' opinion voting on effectivity of prescribed medicine correctness. The differences in opinions of three doctors regarding effectivity was not greater than 2 %. So, in practice for every selected therapy opinion on its effectivity for a particular patient amongst doctors were matching. Regarding ranking of selected three therapies by similarity measure, there were more of diversity between doctors' opinions was wider as well as number of possible matching pairs was higher. As a result, for every therapy the number of inversions was calculated, then the results of three doctors' opinions were averaged and multiplied on 0.25 points. This way, every therapy would have penalty point in the range from 0 to 0.5 points. Obtained values of penalty points were subtracted from measure of therapy correspondence to effective treatment. As the result effective therapies were coded from 0.5 to 1 for classification quality metrics calculation. For all of three methods of similarity learning, using in current research, the quality metrics were calculated corresponding to therapies following order in preference to particular patient. The results are demonstrated in table.

**Classification quality metrics**

| Metric | "Weighted matching" | Cosine measure | Soft cosine measure |
|---|---|---|---|
| Accuracy | 0.79 | 0.81 | 0.88 |
| Sensitivity | 0.85 | 0.86 | 0.86 |
| Recall | 0.68 | 0.72 | 0.78 |

As it could have been seen from analysis the best method, that allows to find appropriate antianemia therapy and PCBR therapy for dialysis patients is soft cosine measure.

## Discussion of the results

Obtained results on most appropriate treatment strategies selection are now in a base of "patient-therapy" similar object search algorithm development using soft cosine measure approved considering possibility of using prescribed medicines analogues. Algorithm was written on Python language and implemented in Lexema-Medicine ERP-system, serving dialysis centers (LLC "Laboratory of hemodialysis").

For Lexema-Medicine ERP system and file interaction SQL requests were used. Also, with the help of that interaction collecting of the necessary data, which then processed in Python, performed. The procedure of implementation could be described in the form of the list of following steps:

- ERP system calls SQL request with necessary attributes;
- SQL transfers collected data to Python script;
- Results transferring into SQL request;
- Transferring result into ERP system.

As the result the algorithm of search of three treatment strategies (medicine-dosage-rout of administration) regarding anti-anemia therapy and PCBR therapy, for every patient, ranked by best similarity measure, was implemented in production work of 24 dialysis centers in 3 Russia's administrative area.

### References

1. **Zhang P., Wang F., Hu J., Sorrentino R.** Towards Personalized Medicine: Leveraging Patient Similarity and Drug Similarity Analytics, *AMIA Joint Summits on Translational Science proceedings*, 2014, pp. 132—136.

2. **Wang N., Huang Y., Liu H., Fei X., Wei L., Zhao X., Chen H.** Measurement and application of patient similarity in personalized predictive modeling based on electronic medical records, *BioMedical Engineering OnLine*, 2019, vol. 18, Article number: 98.

3. **Hu L. Y., Huang M. W., Ke S. W., Tsai C. F.** The distance function effect on k-nearest neighbor classification for medical datasets, SpringerPlus 5, 2016, Article number: 1304.

4. **Hassan D., Aickelin U., Wagner C.** Comparison of Distance Metrics for Hierarchical Data in Medical Databases International Joint Conference on Neural Networks (IJCNN), 2014, pp. 3636—3643.

5. **Kumar A. D., Annie L. C.** Clustering Dichotomous Data for Health Care, *International Journal of Information Sciences and Techniques (IJIST),* 2012, vol. 2, no. 2, pp. 23—33.

6. **Bertels J., Eelbode T., Berman M., Vandermeulen D., Maes F., Bisschops R., Blaschko M.** 2019 Optimizing the Dice Score and Jaccard Index for Medical Image Segmentation: Theory & Practice, *Medical Image Computing and Computer Assisted Intervention*, MICCAI 2019, Part II, pp. 92—100.

7. **Suo Q., Ma F., Yuan Y., Huai M., Zhong W., Gao J., Zhang A.** Deep Patient Similarity Learning for Personalized Healthcare, *IEEE Transactions on NanoBioscience,* 2018, vol. 17, no. 3, pp. 219—227.

8. **Zhu Z., Yin C., Qian B., Cheng Y., Wei J., Wang F.** Measuring Patient Similarities via a Deep Architecture with Medical Concept Embedding, *IEEE 16th International Conference on Data Mining (ICDM)*, Barcelona, 2016, pp. 749—758.

9. **Amiya1 G., Anuradha J., Venkatesh B.** Classification Rule Generation for Cancer Prediction using Locality Sensitive Hashing Similarity Measure, *International Journal of Engineering & Technology*, 2018, 7 (4), pp. 5313—5317.

10. **Lashari S. A., Ibrahim R., Senan N.** Medical Data Classification Using Similarity Measure of Fuzzy Soft Set Based Distance Measure, *Journal of Telecommunication, Electronic and Computer Engineering*, 2017, vol. 9, no. 2—9, pp. 95—99.

11. **Lakman I., Padukova N., Nafikov Sh.** Methodological foundation of comprehensive support for dialysis patients based on artificial intelligence technologies, *Proceedings of the International Scientific and Practical Conference on Digital Economy (ISCDE 2019)*, 2019, pp. 934—939.