Dimensionality Reduction of Chronic Kidney Disease Data using Principal Components Analysis

Tetyana Chumachenko[a] and Kseniia Bazilevych[b]

[a] *Kharkiv National Medical University, Nauky ave., 4, Kharkiv, 61000, Ukraine*
[b] *National Aerospace University "Kharkiv Aviation Institute", Chkalow str., 17, Kharkiv, 61070, Ukraine*

Abstract

Chronic kidney disease is a long-term progressive decline in kidney function. Chronic kidney disease is widespread throughout the world. The disease is diagnosed in 10-13% of adults, 20% older than 60. Early diagnosis allows for taking timely and effective measures to reduce the risk of developing chronic kidney disease. Automated diagnostics using machine learning methods allow for making a diagnosis at an early stage with high accuracy. However, medical data requires pre-processing, and many attributes can negatively affect the model's performance. Therefore, the study of intelligent methods for reducing the dimension of medical data samples is relevant. In this article, we have developed a data dimensionality reduction model for patients with suspected chronic kidney disease based on principal component analysis. As a result, an application was implemented that made it possible to reduce the sample size from 13 to 2 principal components using the principal component analysis method.

Keywords 1

Dimensionality reduction, kidney disease, principal component analysis

1. Introduction

Chronic kidney disease is a long-term progressive decline in kidney function [1]. Any cause can cause the disease due to a significant kidney function impairment. The most common causes are diabetic nephropathy, hypertensive nephrosclerosis, and primary and secondary glomerulopathies [2]. Also, a common cause of kidney damage is metabolic syndrome, characterized by arterial hypertension and type 2 diabetes mellitus.

Chronic kidney disease is widespread throughout the world. The disease is diagnosed in 10-13% of the adult population, 20% of whom are older than 60 years [3].

Chronic kidney disease in the early stages is described as a decrease in renal reserve or renal insufficiency that may progress. The loss of renal tissue function has practically no obvious pathological manifestations because, due to the functional adaptation of the kidneys, the remaining tissue works hard.

With a moderate decrease in renal reserve, the course is often asymptomatic. Symptoms of the disease develop slowly and in later stages include [4]:

- anorexia;
- nausea;
- vomiting;
- stomatitis;
- apathy;
- chronic fatigue;
- decreased clarity of consciousness;

2nd International Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2022), December 2-4, 2022, Łódź, Poland

EMAIL: tatalchum@gmail.com (T. Chumachenko); ksenia.bazilevich@gmail.com (K. Bazivelych)
ORCID: 0000-0002-4175-2941 (T. Chumachenko); 0000-0001-5332-9545 (K. Bazilevych)

- fluid retention;
- muscle convulsions and spasms;
- peripheral neuropathies;
- epileptic seizures;
- etc.

The presence of chronic kidney disease is first suspected with an increase in serum creatine levels [5]. At the beginning of the diagnosis, it is determined whether the kidney failure is acute, chronic, or acute, which has passed into chronic. The examination includes a urinalysis with microscopy of the urinary sediment, an assessment of the level of electrolytes, urea nitrogen, creatinine, phosphates, calcium, and a complete blood count. A history of elevated creatinine or abnormal urinalysis is most helpful in the differential diagnosis. Therefore, early diagnosis is an essential tool for detecting and preventing acute and chronic kidney disease development.

In recent years, data-driven medicine and intelligent technologies for healthcare have been widely developed. Such approaches are used for automated diagnostics [6], forecasting the development of infectious morbidity [7], studying epidemic processes [8], analyzing medical data [9], studying factors affecting the dynamics of morbidity [10], developing medical decision support systems [11], etc.

Machine learning methods are the most effective for automated diagnostics and the development of physician decision support systems. However, data sets often require pre-processing, and many data attributes can reduce the accuracy of models. Therefore, an urgent task is to study methods for reducing data dimension for their application to medical data.

Therefore, this work aims to develop a model for reducing the dimensionality of these patients with suspected chronic kidney disease based on principal component analysis.

Research is part of a complex, intelligent information system for epidemiological diagnostics, the concept of which is discussed in [12].

2. Materials and Methods

Principal components analysis is one of the main methods of data dimensionality reduction, losing the least amount of information [13]. The method is used in many areas, including pattern recognition, computer vision, and data compression. The calculation of principal components is reduced to the calculation of eigenvectors and eigenvalues of the covariance matrix of the original data or the singular value decomposition of the data matrix.

Principal components analysis has several basic versions [14]:

- Approximate data by linear manifolds of lower dimension;
- Find subspaces of lower dimension in the orthogonal projection on which the data spread, is maximum;
- Find subspaces of lower dimension in the orthogonal projection onto which the root-mean-square distance between points is maximal;
- For a given multidimensional random variable, construct such an orthogonal transformation of coordinates that, as a result, the correlations between individual coordinates will vanish.

The first three options operate on finite data sets. They are equivalent and do not use any hypothesis about statistical data generation. The last option operates with random variables. Finite sets appear here as samples from a given distribution, and the solution of the first three problems approximates the true Karhunen-Loeve transformation [15].

Let there be *n* numerical features $f_j(x), j=1,...,n$. The objects of the training sample will be identified with their indicative descriptions:

$$x_i \equiv \big(f_1(x_i), \dots, f_n(x_i)\big), i = 1, \dots, l. \quad (1)$$

Consider the matrix *F*, the rows of which correspond to the indicative descriptions of training objects:

$$F_{l\times n} = \begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_1) & \dots & f_n(x_1) \end{pmatrix} = \begin{pmatrix} x_1 \\ \dots \\ x_1 \end{pmatrix}. \quad (2)$$

Denote by $z_i = (g_1(x_i), \dots, g_m(x_i))$ feature descriptions of the same objects in the new space $Z=R^m$ of lower dimension, $m < n$:

$$G_{l\times n} = \begin{pmatrix} g_1(x_1) & \dots & g_m(x_1) \\ \dots & \dots & \dots \\ g_1(x_1) & \dots & g_m(x_1) \end{pmatrix} = \begin{pmatrix} z_1 \\ \dots \\ z_1 \end{pmatrix}. \quad (3)$$

We require that the original feature descriptions can be restored from new descriptions using some linear transformation determined by the matrix $U=(u_{js})_{n\ x\ m}$:

$$\hat{f}_j(x) = \sum_{s=1}^{m} g_s(x)u_{js}, j = 1, \dots, n, x \in X. \quad (4)$$

or in vector notation.

$$\hat{x} = zU^T. \quad (5)$$

The reconstructed description of the vector form does not have to exactly match the original description x, but their difference on the objects of the training sample should be as small as possible for the chosen dimension m. We will search simultaneously for the matrix of new feature descriptions G and the linear transformation matrix U for which the total discrepancy $\Delta_2(G,U)$ of the restored descriptions is minimal:

$$\Delta^2(G,U) = \sum_{i=1}^{l} ||\hat{x}_i - x_i||^2 = \sum_{i=1}^{l} ||z_i U^T - x_i||^2 = ||GU^T - F||^2 \to \min_{G,U}, \quad (6)$$

where all norms are Euclidean.

Assume that the matrices G and U are non-degenerate: $rank\ G = rank\ U = m$. Otherwise there would be a representation

$$\bar{G}\bar{U}^T = GU^T, \quad (7)$$

with the number of columns in the matrix $\bar{G}$ less than m. Therefore, only cases where $m \leq rank\ F$ are of interest.

If $m \leq rank\ F$, then the minimum of $\Delta^2(G, U)$ is reached when the columns of the matrix U are the F^TF eigenvectors corresponding to the m maximum eigenvalues. Moreover, $G = FU$, the matrices U and G are orthogonal.

The main limitations of the principal component method are:

- The impossibility of semantic interpretation of the components, since they include dispersion from several initial variables;
- The method can only work with continuous data.

3. Results

The Python language was used to build the dimensionality reduction model. For experimental studies, a dataset of patients with suspected chronic kidney disease was used [16]. It contains measures of 24 features for 400 people. 14 features are numerical and 10 are categorical. Description of features is presented in Table 1.

Table 1
Dataset description

Feature	Scale type	Range
BloodPressure	Metric	50…180
SpecificGravity	Metric	1…1.02
Albumin	Metric	0…5
Sugar	Metric	0…5
RedBloodCell	Boolean	0,1
BloodUrea	Metric	1.5…391
SerumCreatinine	Metric	0.4…76
Sodium	Metric	4.5…163
Pottasium	Metric	2.5…47
Hemoglobin	Metric	3.1…17.8
WhiteBloodCellCount	Metric	2200…26400
RedBloodCellCount	Metric	2.1…8
Hypertension	Boolean	0,1
PredictedClass	Boolean	0,1

The dataset visualization is presented in Figure 1.

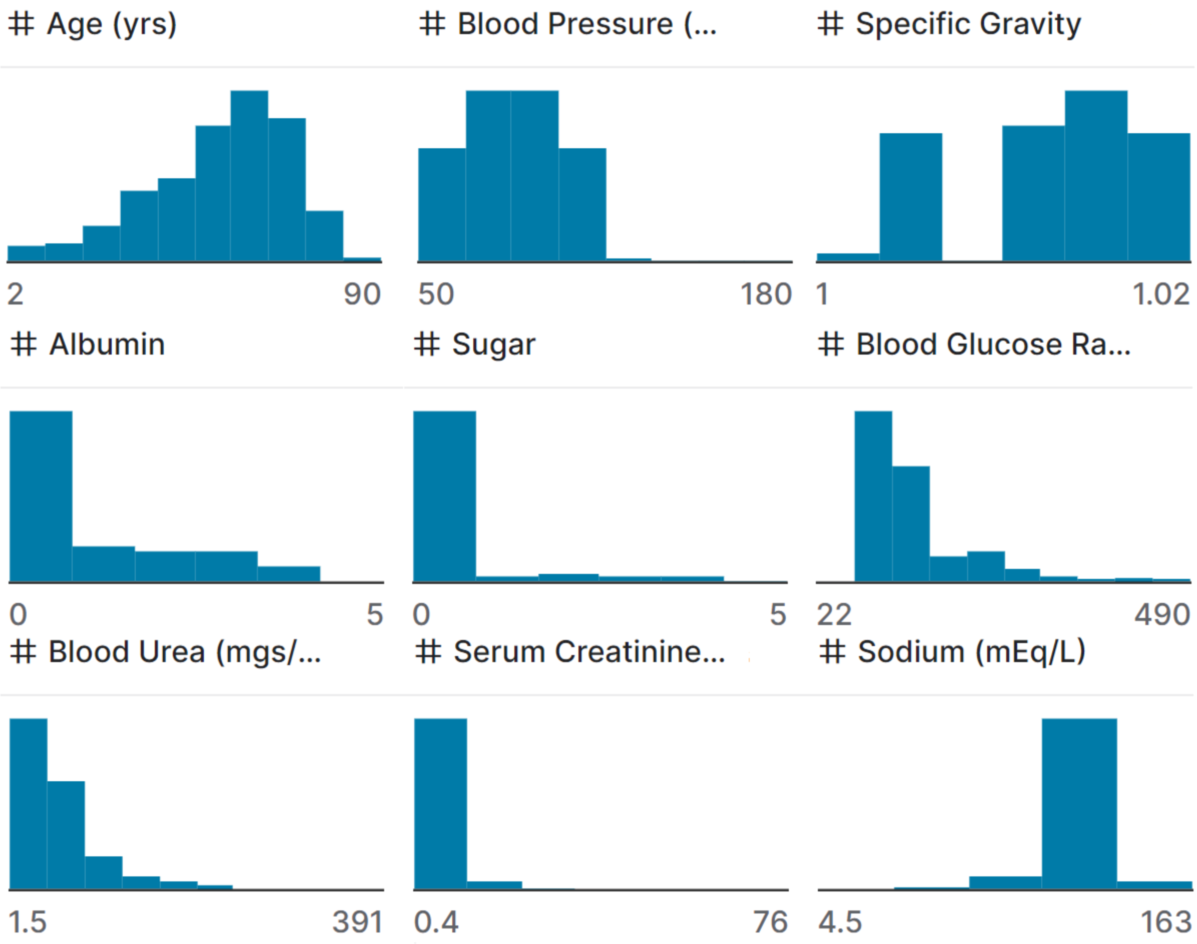

Figure 1: Data visualization

Initially, the data set was divided into objects and the objective function, and data processing was performed. Then an instance of the principal components analysis object was created and the data dimension was reduced while maintaining the two principal components. Thus, the new dimension of the dataset has the form (400, 2), principal component shape is (2, 13). Principal components analysis coefficients are presented in Table 2.

Table 2
Principal components coefficients.

0.1785287	-0.08201316
-0.2972433	0.30440845
0.34606806	-0.20975004
0.17004891	-0.30040035
-0.19416257	0.10444567
0.34101035	0.29990523
0.28353996	0.52222925
-0.25673523	-0.38659614
0.10679924	0.20726542
-0.39901579	0.06977409
0.09928799	-0.42844184
-0.3681989	0.05511764
0.33874907	-0.09441467

Visualization of the obtained results is shown in Figure 2.

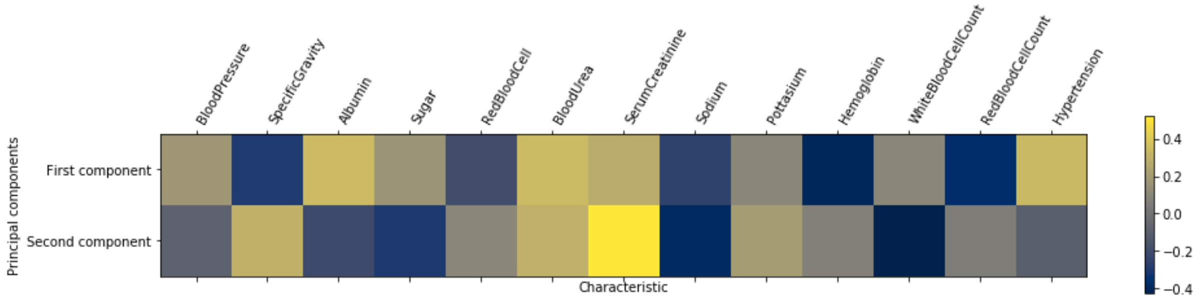

Figure 2: Results visualization

4. Conclusions

Chronic kidney disease is a pressing problem worldwide. An early diagnosis is an effective tool for reducing the development of chronic kidney disease. Machine learning methods make it possible to build models for the early detection of a disease, which allows for taking timely, practical measures to counteract the development of the disease. However, medical data requires preliminary preparation, and datasets containing many attributes can reduce the accuracy of diagnostic models. Therefore, within the framework of this study, a model for reducing the dimensionality of data from patients with suspected chronic kidney disease was developed based on the principal component analysis method. The constructed model made it possible to reduce the dimension of the dataset from 13 to 2.

Further research will combine the developed model with machine learning models to classify suspected chronic kidney disease patients.

5. Acknowledgements

The study was funded by the National Research Foundation of Ukraine in the framework of the research project 2020.02/0404 on the topic “Development of intelligent technologies for assessing the epidemic situation to support decision-making within the population biosafety management”

6. References

[1] C. Charles, A.H. Ferris, Chronic kidney disease, Primary care 47 (4) (2020) 585-595. doi: 10.1016/j.pop.2020.08.001

[2] M. Provenzano, et. al., Epidemiology of cardiovascular risk in chronic kidney disease patients: the real silent killer, Reviews in cardiovascular medicine 20 (4) (2019) 209-220. doi: 10.31083/j.rcm.2019.04.548

[3] V. Jha, et. al., Chronic kidney disease: global dimension and perspectives, Lancet 382 (9888) (2013) 260-72. doi: 10.1016/S0140-6736(13)60687-X.

[4] M.G. Shlipak, et al., The case for early identification and intervention of chronic kidney disease: conclusions from a kidney disease: improving global outcomes controversies conference, Kidney international 99 (1) (2021) 34-47. doi: 10.1016/j.kint.2020.10.012

[5] T.K. Chen, D.H. Knicely, M.E. Grams, Chronic kidney disease diagnostics and management: a review, Journal of American Medical Association 322 (13) (2019) 1294-1304. doi: 10.1001/jama.2019.14745

[6] M.A. Myszczynska, et. al., Applications of machine learning to diagnosis and treatment of neurodegenerative diseases, Nature Reviews Neurology 16 (2020) 440-456. doi: 10.1038/s41582-020-0377-8

[7] D. Chumachenko, et. al., Investigation of statistical machine learning models for COVID-19 epidemic process simulation: random forest, k-nearest neighbors, gradient boosting, Computation 10 (6) (2022) 86. doi: 10.3390/computation10060086

[8] D. Chumachenko, On intelligent multiagent approach to viral Hepatitis B epidemic processes simulation, Proceedings of the 2018 IEEE 2nd International Conference on Data Stream Mining and Processing, DSMP 2018 (2018) 415-419. doi: 10.1109/DSMP.2018.8478602

[9] R. Tkachenko, et. al., Committee of the SGTM neural-like structures with extended inputs for predictive analytics in insurance, Communications in Computer and Information Science 1054 (2019). doi: 10.1007/978-3-030-27355-2_9

[10] N. Davidich, et. al. Monitoring of urban freight flows distribution considering the human factor, Sustainable Cities and Society 75 (2021) 103168. doi: 10.1016/j.scs.2021.103168

[11] D. Chumachenko, et. al. Intelligent expert system of knowledge examination of medical staff regarding infections associated with the provision of medical care, CEUR Workshop Proceedings 2386 (2019) 321-330.

[12] S. Yakovlev, et. al. The concept of developing a decision support system for the epidemic morbidity control, CEUR Workshop Proceedings 2753 (2020) 265-274.

[13] I.T. Jolliffe, J. Cadima, Principal component analysis: a review and recent developments, Philosophical Transactions of the Royal Society 374 (2065) (2016) 20150202. doi: 10.1098/rsta.2015.0202

[14] D. Groth, S. Hartmann, S. Klie, J. Selbig, Principal components analysis, Methods in molecular biology 930 (2013) 527-47. doi: 10.1007/978-1-62703-059-5_22

[15] Y. Zhou, X. Ai, M. Lv, B. Tian, Karhunen-Loève Expansion for the Second Order Detrended Brownian Motion, Abstract and Applied Analysis 2014 (2014) 457051. doi: 10.1155/2014/457051

[16] Chronic Kidney Disease Data Set, UCI Machine Learning Repository (2019). Available at: https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease (accessed on 30.09.2022)