ICIC Express Letters
Volume **14**, Number **6**, June **2020**

ICIC International ©2020 ISSN 1881-803X
pp. 585–591

FACTOR ANALYSIS USING ARTIFICIAL INTELLIGENCE IN PATIENTS WITH CHRONIC KIDNEY DISEASE

Kenya Fukai[1], Hideaki Kawano[2] and Junki Enomoto[1]

[1]Graduate School of Engineering
[2]Faculty of Engineering
Kyusyu Institute of Technology
1-1, Sensui-cho, Tobata-ku, Kitakyusyu 804-8550, Japan
kawano@ecs.kyutech.ac.jp

Received December 2019; accepted March 2020

Abstract. *In recent years, the number of patients with chronic kidney disease in Japan has reached 13.3 million (1 in 8 of adults over 20 years old). The disease is said to be a new country illness. The important factors of the chronic kidney disease have been clarified by long-term clinical studies. In this paper, we propose a new method to discover important factors inherent to the chronic kidney disease patients composed by Topological Data Analysis (TDA) and Layer-wise Relevance Propagation (LRP). To verify the effectiveness of the proposed method, we compared the result obtained by our proposed method with the known important factors by the clinical studies.*

Keywords: Topological data analysis, Layer-wise relevance propagation, Chronic kidney disease, High dimensional data, Important factor

1. **Introduction.** In recent years, 13.3 million people (1 in 8 adults over 20) have chronic kidney disease in Japan, and it is said to be a new national disease. The unique features of this chronic kidney disease patient have been revealed by many years of clinical research [1]. In this study, we use Topological Data Analysis (TDA) and Layer-wise Relevance Propagation (LRP) to analyze important factors. We confirm whether these techniques are valid. In order to confirm the effectiveness of analysis of important factors using these, check if the result of this method is the same as the known result. Principal Component Analysis (PCA), Multi-Dimensional Scaling (MDS) and cluster analysis have been used as classical analysis methods for high dimensional data. However, after embedding data in low dimensions, these methods make it difficult to understand the relationship between data, and there is little information to be obtained. So we use Topological Data Analysis (TDA) [2] as a high dimensional analysis method.

TDA is unsupervised learning. Compared with conventional analysis methods, TDA can visually determine the degree of association between data. Because of its relevance, you can gain new insights that you cannot understand at a high dimension. The TDA used here is the TDA method known as mapper algorithm [2] developed by Singh et al. In this method, the high-dimensional data is dimensionally compressed to a low dimension. Then, create an edge that shows the relationship between the data. This mapper algorithm is currently attracting attention in the fields of bioinformatics and medicine [3]. And now, we use TDA to group patients with chronic kidney disease and discover key factors by group. Layer-wise Relevance Propagation (LRP) is used to analyze this important factor. LRP updates weights and biases in forward propagation and after that determine the degree of contribution of input by back propagation [4]. This method is mainly used for analysis of image judgment basis, but this time it is applied to numerical data. In this study, chronic kidney disease patients are classified into three groups using TDA,

DOI: 10.24507/icicel.14.06.585

and LRP is used to analyze the key factors of each group. In this paper, we will discuss related research in Section 2, proposed method in Section 3, experiment in Section 4, and conclusion in Section 5.

2. Research Questions.

To analyze the characteristics of clinically and genetically diverse patients with type 2 diabetes, TDA [5] is used to group patient data. Usage data are clinical data recorded in Electronic Medical Records (EMRs). The subjects were 2551 patients enrolled in Biobank of Mount Sinai Medical Center in New York, and the data combine clinical data on genotypes with clinical data on EMRs of those patients. The set was used. The breakdown of 2551 patients is 46% Hispanic, 32% African American, 20% European Caucasian, 2% others. The gender was 61% female, 39% male and the overall average age was 55.5. Figure 1 shows the results of visualization of 2551 people in TDA. Red indicates female and blue indicates male. This result shows that type 2 diabetes can be divided into three groups. After visualization with TDA, the authors perform mean comparisons to analyze the features of each of the three groups. However, this mean comparison may not find the features well. Therefore, after grouping TDA, we should use a more detailed comparison method. Therefore, this study uses LRP instead of average comparison after grouping with TDA. By using LRP, it is also possible to analyze data where it is difficult to understand the feature differences between groups.

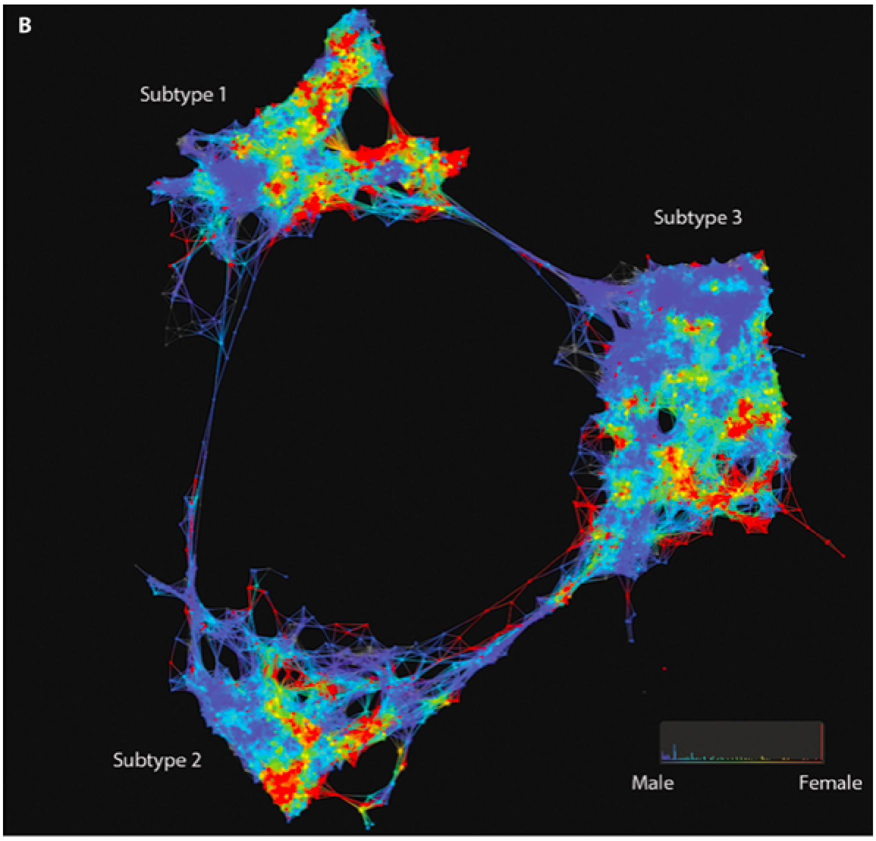

Figure 1. (color online) Visualization result by TDA of type 2 diabetes patients

3. Proposed method.

3.1. t-SNE (t-distributed Stochastic Neighbor Embedding).

The TDA process first dimensionally compresses high-dimensional data to low dimensions. In dimension compression, t-SNE [6] was used. t-SNE is an algorithm that represents high-dimensional spatial similarity and low-dimensional space (two-dimensional) with probability distributions, minimizes the KL information (Kullback-Leibler divergence) and reduces the error during compression. First, the normal distribution P_{ij} is expressed with the degree of similarity in high-dimensional space (using Euclidean distance) as the reference point x_i.

$$P_{ij} = \frac{P_{j|i} + P_{i|j}}{2n} \tag{1}$$

$$P_{i|j} = \frac{\exp\left(-\frac{|x_i - x_j|^2}{2\sigma_i^2}\right)}{\sum\limits_{k \neq i} \exp\left(-\frac{|x_i - x_k|^2}{2\sigma_i^2}\right)} \tag{2}$$

P_{ij} is a normal distribution, and $P_{i|j}$ is the conditional probability of similarity of x_i viewed from x_j. n is the number of observation points, and σ is the variance of the Gaussian that is centered on data point x. After that, we use t-distribution with one degree of freedom as the similarity in low-dimensional space. By using t-distribution, data with low similarity is positioned farther on the low-dimensional space during dimensional compression. The t-distribution equation is denoted by q_{ij}.

$$q_{ij} = \frac{\left(1 + |y_i - y_j|^2\right)^{-1}}{\sum\limits_{k \neq l} \left(1 + |y_k - y_l|^2\right)^{-1}} \tag{3}$$

y_i and y_j are parameters to be estimated. If the map points y_i and y_j correctly model the similarity between the high-dimensional data points x_i and x_j, the conditional probabilities $P_{j|i}$ and $q_{j|i}$ will be equal. t-SNE aims to find a low-dimensional data representation that minimizes the mismatch between $P_{j|i}$ and $q_{j|i}$. Therefore, t-SNE uses KL information as an evaluation method.

$$KL(P, Q) = \sum_i \sum_j P_{ij} \log \frac{P_{ij}}{q_{ij}} \tag{4}$$

in which P_i represents the conditional probability distribution over all other data points given data point x_i, and Q_i represents the conditional probability distribution over all other map points given map point y_i.

$$\frac{\delta C}{\delta y_i} = 4 \sum_j (P_{ij} - q_{ij})(y_i - y_j)\left(1 + |y_i - y_j|^2\right)^{-1} \tag{5}$$

Next, Equation (5) is updated using Equation (6).

$$Y^{(t)} = Y^{(t-1)} + \eta \frac{\delta C}{\delta Y} + \alpha(t)\left(Y^{(t-1)} - Y^{(t-2)}\right) \tag{6}$$

$Y^{(t)}$ indicates the solution at iteration t, η indicates the learning rate, and $\alpha(t)$ represents the momentum at iteration t.

3.2. **Preprocessing of edge creation.** Clustering is performed after dropping high-dimensional data into two dimensions. Unlike normal clustering, mapper [5] outputs nodes and edges. After compression in two dimensions, data with high similarity is stored in one node, and nodes with high similarity between nodes are connected by edges to facilitate visualization. Here, the parameters interval and overlap are specified. After dropping high-dimensional data into a two-dimensional space, divide the data into the number of intervals. At the time of the division, the peripheral data divided by overlap is arranged on both left and right.

3.3. **DBSCAN (Density-Based Spatial Clustering of Application with Noise).** After high-dimensional data is dropped into two dimensions, nodes and edges are created by clustering DBSCAN algorithm [7]. Set the parameters ε and minPts when dividing into clusters. Take the radius around the point specified by ε, and within that radius the cluster is determined by the number of minPts points. Here we divide it into three clusters. Core point is a cluster that contains $x >$ minPts points (x: number of points in radius). Reachable points are clusters that do not have $x <$ minPts points within the radius ε, but include core points within the radius ε. Outlier point is a cluster that has no adjacent points within the radius. Therefore, we create a cluster from a set of core points and consider reachable points as core nodes including one node. The edge connecting nodes is determined by the clustering preprocessing parameter. Nodes containing the same data are connected by edges to indicate the degree of association between the nodes.

3.4. **Layer-wise relevance propagation.** Layer-wise Relevance Propagation (LRP) is used to detect factors that classify two groups.

LRP first performs binary classification of data. After that, the relationship between layers is back-propagated, and the input is reached, and the contribution of the input is obtained [8]. The sum of the input contributions to the outputs is equal among the layers, taking advantage of the fact that their distribution changes only during propagation.

$$x_j^{(l+1)} = \sigma\left(\sum_i x_i^{(l)} w_{ij} + b_j^{(l+1)}\right) \tag{7}$$

The weights and bias are updated in (7), where w_{ij} is the weight, b_i is the bias, and σ is the activation function.

$$R_f = F(x) \tag{8}$$

$F(x)$ shows the output of class A in (8). Next, we use this output $F(x)$ to consider the contribution of the input in LRP.

$$z_{ij} = x^{(i)} w_{ij}^{(l,l+1)} \tag{9}$$

$$R_i^{(l)} = \sum_j \frac{z_{ij}}{\sum_{i'} z_{i'j}} R_j^{(l+1)} \tag{10}$$

LRP performs back propagation with the output $F(x)$ in forward propagation as R.

At the time of back propagation, the contribution of the input is output using Equation (10). The detailed structure is shown in Figure 2.

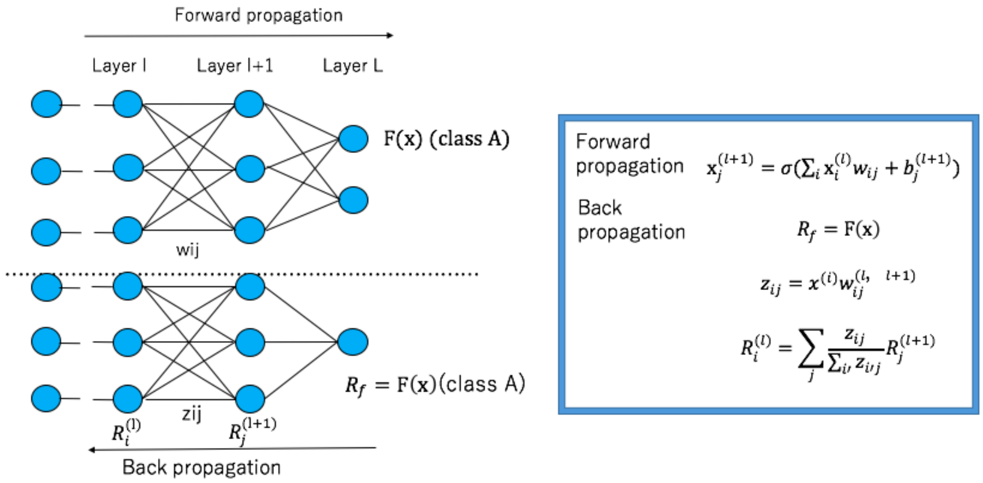

FIGURE 2. The structure of LRP is shown. Process and classify input values. Then, back propagation is performed to show the contribution of the input value.

4. **Experiment.** In this study, we first use TDA to visualize chronic kidney disease patient data. Visualize and find groups, and analyze the important factors of groups using LRP for each group. Usage data are 18 features of 490 chronic kidney disease patients. There are 250 chronic kidney disease patients and 240 non-chronic kidney disease patients. First, the entire data is visualized, and then only chronic kidney disease patient data is visualized to find the characteristics of each group. Figure 3 shows the results of visualizing the entire chronic kidney disease patient data. About 80% of the patients in group 1 showed the label of non-chronic kidney disease patients, and about 90% of the patients in group 2 showed chronic kidney disease patients. From this output result, it is clear that chronic kidney disease patients and non-chronic kidney disease patients can be roughly classified. Therefore, next we try to visualize the data group of chronic kidney disease patients only with TDA. The number of chronic kidney disease patient data here is 250.

The output results are shown in Figure 4. Figure 4 shows that chronic kidney disease patients are divided into three groups. This result does not focus on the color, but only on the shape at the nodes and edges. After grouping, LRP is used to compare these G1, G2, and G3 groups with patients with non-chronic kidney disease to determine feature contributions. Output results using LRP are shown in Figure 5 to Figure 7. The item

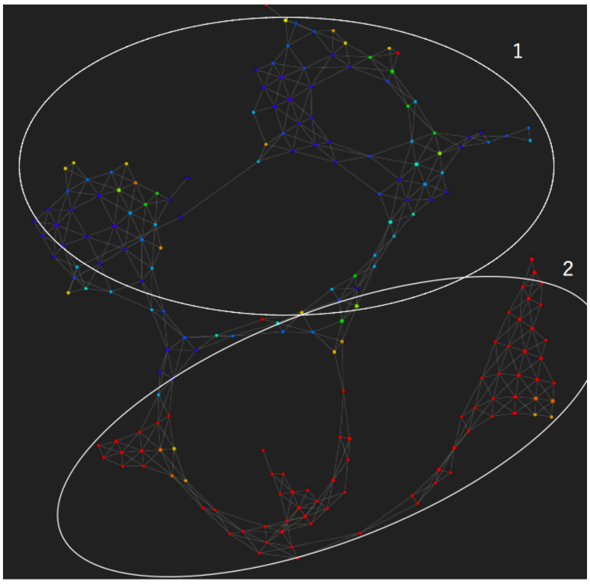

Figure 3. Visualization results of whole chronic kidney disease patient data in TDA

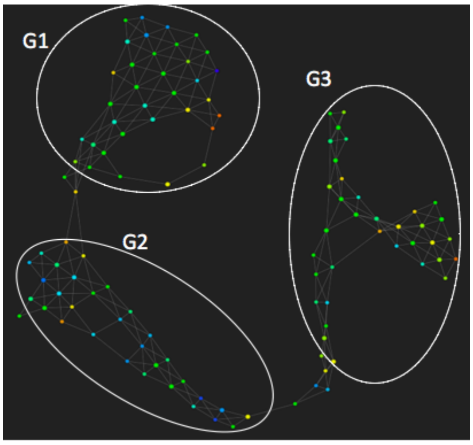

Figure 4. Visualization results of only chronic kidney disease patients by TDA

with the highest contribution rate is considered to indicate the characteristics of the group. In Figure 5, group 1 had the highest contribution of albumin value. Similarly, in Figure 6, group 2 has the highest contribution of the complication of hypertension, and in Figure 7, group 3 has the highest contribution of the blood glucose level. In addition, the contribution of hemoglobin value is high as a whole. Comparing the results of grouping with TDA with the output of LRP, the contribution of albumin level decreases and the contribution of blood glucose level increases as you trace the edges G1 → G2 → G3.

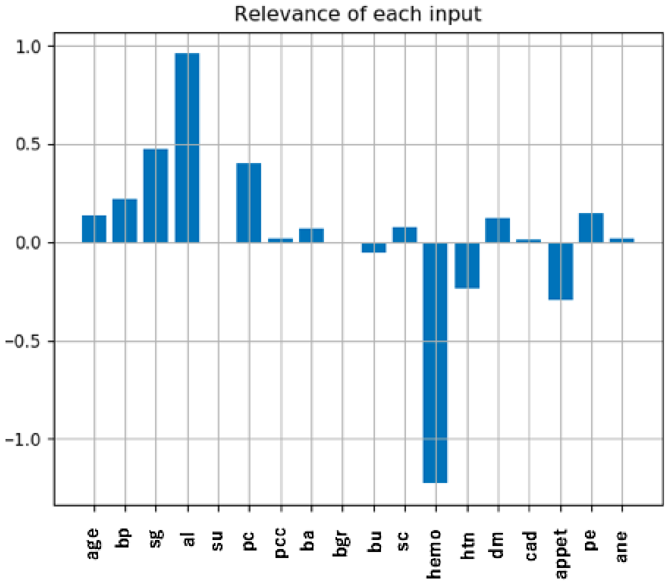

FIGURE 5. Comparison result of contribution of G1 and non-chronic kidney disease patients

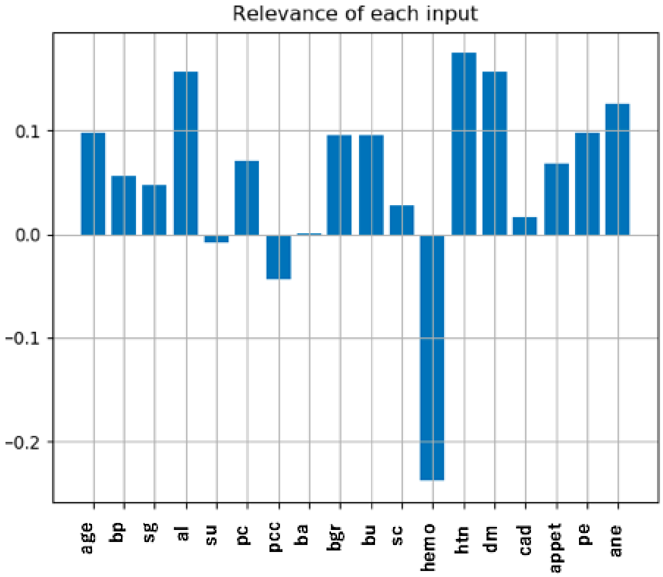

FIGURE 6. Comparison result of contribution of G2 and non-chronic kidney disease patients

5. **Conclusion.** Chronic kidney disease patient data were grouped using TDA, and LRP was used to calculate feature values for each group. In this study, we believe that the important factors in chronic kidney disease patients are albumin levels, blood glucose levels, complications of hypertension, and hemoglobin levels. These results are consistent

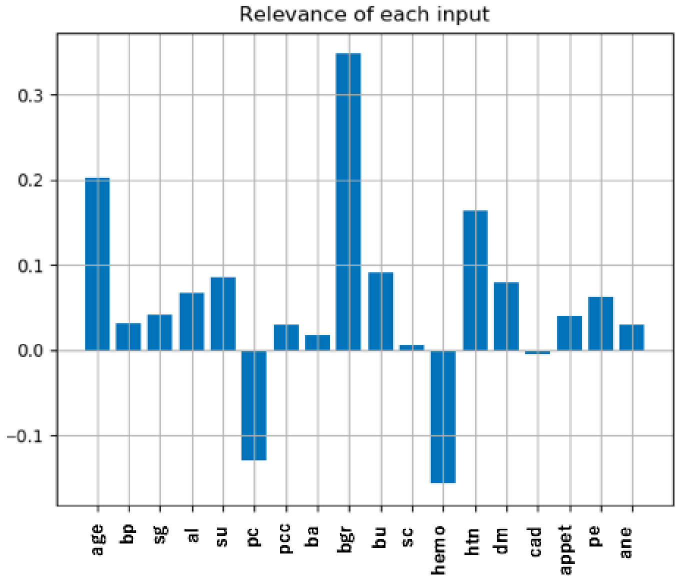

Figure 7. Comparison result of contribution of G3 and non-chronic kidney disease patients

with the characteristics of chronic kidney disease patients that have been analyzed for many years, and it was also found that TDA and LRP are useful for important factor analysis. In the future, we plan to analyze data with unknown determinants with these methods and expand it as a means to obtain new findings and insights.

REFERENCES

[1] The Japanese Society of Nephrology, *CKD Clinical Practice Guide 2009*, Revised 2nd Edition, Tokyo Medical, 2009.

[2] G. Singh, F. Memoli and G. Carlsson, Topological methods for the analysis of high dimensional data sets and 3D object recognition, *Symposium on Point Based Graphics 07*, The Eurographics Association, pp.91-100, 2007.

[3] W. Guo and A. G. Banerjee, Identification of key features using topological data analysis for accurate prediction of manufacturing system outputs, *Journal of Manufacturing System*, vol.43, no.P2, pp.225-234, 2017.

[4] S. Bach, A. Binde, G. Montavon, F. Klauschen, K.-R. Muller and W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, *PLoS One*, vol.10, no.7, 2015.

[5] L. Li, W.-Y. Cheng, B. S. Glicksberg, O. Gottesman, R. Tamler, R. Chen, E. P. Bottinger and J. T. Dudley, Identification of type 2 diabetes subgroups through topological analysis of patient similarity, *Science Translational Medicine*, vol.7, no.311, 2015.

[6] G. Hinton, Visualising data using t-SNE, *Journal of Machine Learning Research*, vol.9, 2008.

[7] M. Ester, H.-P. Kriegel and X. Xu, A density-based algorithm for discovering clusters in large algorithm for discovering noise, *Proc. of the 2nd International Conference on Knowledge Discovery and Data Mining*, Munchen, Germany, pp.226-231, 1996.

[8] A. Binder, S. Bach, G. Montavon, K.-R. Muller and W. Samek, Layer-wise relevance propagation for deep neural network architectures, *ICISA*, pp.913-922, 2016.