Amani Al-Mekhlafi, Sandra Klawitter and Frank Klawonn*

Standardization with zlog values improves exploratory data analysis and machine learning for laboratory data

https://doi.org/10.1515/labmed-2024-0051
Received March 31, 2024; accepted May 31, 2024; published online June 27, 2024

Abstract

Objectives: In the context of exploratory data analysis and machine learning, standardization of laboratory results is an important pre-processing step. Variable proportions of pathological results in routine datasets lead to changes of the mean (µ) and standard deviation (σ), and thus cause problems in the classical z-score transformation. Therefore, this study investigates whether the zlog transformation compensates these disadvantages and makes the results more meaningful from a medical perspective.

Methods: The results presented here were obtained with the statistical software environment R, and the underlying data set was obtained from the UC Irvine Machine Learning Repository. We compare the differences of the zlog and z-score transformation for five different dimension reduction methods, hierarchical clustering and four supervised classification methods.

Results: With the zlog transformation, we obtain better results in this study than with the z-score transformation for dimension reduction, clustering and classification methods. By compensating the disadvantages of the z-score transformation, the zlog transformation allows more meaningful medical conclusions.

Conclusions: We recommend using the zlog transformation of laboratory results for pre-processing when exploratory data analysis and machine learning techniques are applied.

Keywords: exploratory data analysis; laboratory data; machine learning; zlog value; z-score transformation

Introduction

Exploratory data analysis and machine learning (ML) for multivariate datasets have attracted considerable attention in numerous research areas, including laboratory medicine [1, 2]. For many statistical methods used in this context, it is important to standardize the values so that they can be compared with each other despite different scales.

In this study, we analyze a medical machine learning dataset [3] that includes 10 different analytes measured in blood donors and hepatitis C patients as well as demographic data such as sex and age. The most popular method to bring all numerical data onto a common scale, is the z-score transformation, which expresses each value \( x \) as a deviation from the mean value \( m \) in multiples of the standard deviation \( sd \):

\[
z = \frac{x - m}{sd}
\]

Since the z-score transformation is not robust against pathological outliers, we compare the results of this conventional normalization method with the more recent zlog transformation, in which the mean and standard deviation are not calculated from the data itself, but from the logarithms of the lower and upper reference limits LL and UL of the respective analytes [4]:

\[
zlog(x) = \left( \log(x) - \frac{\log(LL) + \log(UL)}{2} \right) \\
\times \frac{3.92}{\log(UL) - \log(LL)}
\]

The aim of our study is to find out which normalization method is superior for dimension reduction as well as unsupervised and supervised machine learning algorithms [4].

Materials and methods

The HCV dataset used in our study is freely available in the UC Irvine Machine Learning Repository [3] and has already been successfully used for machine learning applications [2, 5]. Here, it forms the basis for the comparison of the two standardization methods mentioned above. The data
contains measurements in 583 individuals aged 23–77 years. 61% thereof are male and 39% are female.

The participants are separated into five categories [5] with the following means and standard deviations of age: 519 healthy blood donors (47.2±9.6), seven blood donors with the suspicion of liver disease (57.6±11.1), 20 patients diagnosed with histologically inconspicuous hepatitis C (40.7±11.1), 12 patients diagnosed with liver fibrosis (49.7±12.1), and 24 patients diagnosed with liver cirrhosis (54.3±8.7).

Ten analytes, which are commonly used as parameters for the early- and late-stage functionality of the liver, are reported: albumin (ALB), bilirubin (BIL), alkaline phosphatase (ALP), alanine aminotransferase (ALT), aspartate aminotransferase (AST), cholesterol (CHOL), creatinine (CREA), gamma-glutamyltransferase (GGT), and total protein (PROT). For the analysis of the data and the preparation of the graphics, we used the statistical software environment R (Version 4.1.2, www.r-project.org) and the following packages: MASS, neuralnet, tsne, umap, class and e1071 [6–10].

To project all 10 analytes onto comparable scales, we transformed the absolute values as z-scores (Equation (1)) and as zlog values (Equation (2)). Both calculations result in transformed the absolute values as z-scores (Equation (1)) and as zlog values (Equation (2)).

Table 1: Reference intervals derived from the data set as 2.5th and 97.5th percentiles of blood donor measurements. LL, lower limit; UL, upper limit.

<table>
<thead>
<tr>
<th>Analyte</th>
<th>Unit</th>
<th>Female LL</th>
<th>Female UL</th>
<th>Male LL</th>
<th>Male UL</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALB</td>
<td>g/L</td>
<td>32.07</td>
<td>50.04</td>
<td>35.03</td>
<td>51.35</td>
</tr>
<tr>
<td>ALP</td>
<td>U/L</td>
<td>36.51</td>
<td>103.13</td>
<td>40.93</td>
<td>105.90</td>
</tr>
<tr>
<td>ALT</td>
<td>U/L</td>
<td>10.12</td>
<td>45.22</td>
<td>11.65</td>
<td>71.03</td>
</tr>
<tr>
<td>AST</td>
<td>U/L</td>
<td>14.92</td>
<td>41.15</td>
<td>17.28</td>
<td>46.98</td>
</tr>
<tr>
<td>BIL</td>
<td>µmol/L</td>
<td>2.40</td>
<td>17.63</td>
<td>2.90</td>
<td>25.40</td>
</tr>
<tr>
<td>CHE</td>
<td>kU/L</td>
<td>4.58</td>
<td>11.02</td>
<td>5.42</td>
<td>12.33</td>
</tr>
<tr>
<td>CHOL</td>
<td>mmol/L</td>
<td>3.63</td>
<td>7.56</td>
<td>3.71</td>
<td>7.76</td>
</tr>
<tr>
<td>CREA</td>
<td>µmol/L</td>
<td>52.00</td>
<td>93.65</td>
<td>63.75</td>
<td>111.00</td>
</tr>
<tr>
<td>GGT</td>
<td>U/L</td>
<td>7.49</td>
<td>73.92</td>
<td>10.48</td>
<td>99.63</td>
</tr>
<tr>
<td>PROT</td>
<td>g/L</td>
<td>62.61</td>
<td>79.63</td>
<td>64.18</td>
<td>80.75</td>
</tr>
</tbody>
</table>

We applied 10 ML methods to the transformed values, four of them being supervised and the other six unsupervised. The goal of supervised machine learning is to predict a known output, in our case the predefined classes (diagnoses) mentioned above, whereas unsupervised techniques try to find hidden structures or correlations in the data without knowing anything about these classes. Ideally, the structures or clusters identified by unsupervised algorithms should reflect the known medical entities.

For unsupervised ML, we applied the following dimensionality reduction methods (DRM): Principal Component Analysis (PCA) [12, 13], Sammon Mapping [14], Autoencoder [15, 16] in two variants (the first with only one hidden neuron layer and the second a more complex with several hidden layers), t-Distributed Stochastic NeighborEmbedding (t-SNE) [17, 18], and Uniform Manifold Approximation and Projection (UMAP) [19]. Dimensionality reduction aims to reduce the number of variables in high-dimensional data sets without substantial loss of information. In this study, we used dimensionality reduction to display the ten-dimensional data set as two-dimensional scatter plots (Figures 2–6). As another exploratory data analysis technique, we used Hierarchical Cluster Analysis (HCA) [20], which orders and groups the cases according to similarities in the measured values.

Finally, we applied the following supervised methods for the prediction of known categories: Linear Discriminant Analysis (LDA) [21], k=1 and 3 Nearest-Neighbor (kNN) [2], linear and non-linear Support Vector Machines (SVM) [22] and Artificial Neural Networks (ANN) with one and three hidden layers [23]. In the supervised section we restricted the experiment to the four well characterized classes listed in Table 3, excluding the group of suspected blood donors because this subgroup was too small (n=7) for reliable detection.

To cross-validate the results, we used the leave-one-out method, which is an iterative algorithm often applied to small data sets [5]. Given a total of n samples, the respective ML models were trained on n-1 samples, and their performance was tested on the single sample left out. This process was repeated n times, once for each sample in the dataset.

Results

Table 1 summarizes the results of the direct estimation of reference intervals. We deliberately determined the 2.5th and 97.5th percentiles from the blood donor data itself rather than from the literature in order to avoid the need for external sources for standardization. Our results reflect the
values given in the literature quite well [24] with only few deviations (e. g. the upper limits of GGT). For the purposes of the present study, this accuracy is absolutely sufficient.

Figure 1 gives an overview of the transformed values for the 10 analytes in a boxplot format. The blue boxes represent the zlog values and the red ones represent the respective z-score values. As expected, all boxes are within the common reference interval of $-1.96$ to $+1.96$ (vertical dashed lines), and all medians (vertical lines inside the boxes) are close to zero. It is noticeable, however, that the boxes (i. e. the central 50 %) of the z-score values are extremely narrow for ALT, AST, BIL, CREA, and GGT, whereas the boxes of the zlog values all have about the same size with a reasonable width compared to the common reference interval. In addition, the zlog values scatter quite symmetrically around the reference interval, whereas most z-score values are shifted to the right so that low values (e. g. ALB, ALT, CHE in liver cirrhosis) are not well represented.

In other words, the zlog values provide more information and reflect the clinical significance of the analytes better than the z-score values. In fact, most laboratory analytes do not conform to a normal distribution; instead, they tend to be right-skewed. Consequently, applying the zlog transformation can help approximate a normal distribution.

Figures 2–6 visualize the results of the dimensionality reduction techniques in two-dimensional scattergrams. The crucial question to be answered here is whether the original information about the different subgroups (inconspicuous and suspected blood donors, hepatitis with and without histological changes) is retained or blurred in the graphics.

The figures confirm the above expectation. Figure 2 shows the results of Principal Component Analysis (PCA), one of the oldest and most common dimensionality reduction techniques. It is based on feature extraction, which means that new features are created that capture the key information contained in the 10 original features (analytes). On the left side of Figure 2 the majority of z-score values for hepatitis and fibrosis are lost in the black cloud of normal blood donors. Only cirrhotic cases and suspected blood donors are clearly separated from the rest. The zlog values shown on the right
resolve significantly more pathological cases: more than half of hepatitis patients without histological signs and the majority of fibrosis patients are found outside the black cloud of healthy blood donors.

Although Sammon Mapping (Figure 3) and Autoencoders (Figure 4) work differently from PCA, the results are very similar. Sammon Mapping aims to preserve the pairwise distances between the points in the high-dimensional space when mapping them to a lower-dimensional space. Autoencoders are neural network architectures designed to learn compressed representations of the input data. Figure 4 shows that the Autoencoder results are independent of whether we use a simple (one layer) or a complex (several layers) autoencoder.

The results of t-SNE and UMAP are even more impressive. These advanced, relatively new methods assume that the multidimensional data lie on a nonlinear data structure (a so-called manifold) and try to preserve the local structures while projecting them to a low-dimensional representation. Figures 5 and 6 show that both techniques resolve the healthy blood donors as distinguishable points and not as dense black areas like in Figures 2–4. With z-score as the underlying standardization technique, the suspected blood donors and the three hepatitis categories are scattered somewhere between the black dots, whereas with zlog transformation they appear as clearly separated clusters. It can also be seen that UMAP takes the absolute distances into account in contrast to t-SNE which focuses only on the local neighborhood of the points. Therefore, UMAP places the diseased cases farther away from the healthy group.

The result of hierarchical clustering is shown in Figure 7. As expected from the “boxes” in Figure 1, the z-scores are mostly concentrated around zero, which results in pale colors in the upper graphic, whereas the zlog values reflect the distribution of normal and pathological values better leading to stronger colors in the lower graphic. Thus, not surprisingly, there is a better separation of the diseased groups from the healthy blood donors with the zlog transformation, especially visible in the cirrhosis patients who cluster well together.

Table 2 summarizes the results of the four classifiers kNN, SVM, LDA, and ANN. It is obvious that the accuracy with zlog values is consistently better than that with z-scores. This becomes even more evident, when we look at the individual results for the four classes. Table 3 takes the LDA method as an example. Here, the zlog values for the three pathological classes perform considerably better than the z-scores. This effect is particularly pronounced in case of fibrosis, which is difficult to diagnose with routine biomarkers [5]: more than 40 % are correctly identified using zlog values, but less than
Figure 4: Autoencoder with z-score and zlog transformation. The upper two figures use a simple and the bottom two figures a complex autoencoder. For more details see Figure 2.

Figure 5: t-Distributed stochastic neighbor embedding (t-SNE) with z-score and zlog transformation. For more details see Figure 2.

Figure 6: Uniform manifold approximation and projection (UMAP) with z-score and zlog transformation. For more details see Figure 2.
20% with z-scores. Interestingly, the high accuracy for healthy blood donors is independent of whether z-scores or zlog values are used.

Detailed results for all supervised methods are shown in the Supplementary Table 1. The distinction between blood donors and cirrhosis cases is consistently the most
Table 2: Percentage of correctly classified results with four supervised machine learning algorithms: A, k-nearest neighbour; B, support vector machine; C, linear discernment analysis; D, neural network with three hidden layers.

<table>
<thead>
<tr>
<th></th>
<th>A k=1</th>
<th>A k=3</th>
<th>B linear</th>
<th>B non-linear</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>z-score</td>
<td>94.6</td>
<td>94.3</td>
<td>94.8</td>
<td>95.1</td>
<td>94.6</td>
<td>98.6</td>
</tr>
<tr>
<td>zlog</td>
<td>96.2</td>
<td>95.7</td>
<td>96.2</td>
<td>95.8</td>
<td>97.0</td>
<td>99.1</td>
</tr>
</tbody>
</table>

Table 3: Comparison of z-score transformation and zlog transformation, taking LDA as an example for a classification method.

<table>
<thead>
<tr>
<th>Blood donor</th>
<th>Hepatitis</th>
<th>Fibrosis</th>
<th>Cirrhosis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>518</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>Blood donor</td>
<td>0</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>Fibrosis</td>
<td>1</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Cirrhosis</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Correctly classified, %</td>
<td>99.8</td>
<td>35.0</td>
<td>16.7</td>
</tr>
</tbody>
</table>

Table 3a: Classification with z-scores (overall accuracy 94.6 %)

<table>
<thead>
<tr>
<th>Blood donor</th>
<th>Hepatitis</th>
<th>Fibrosis</th>
<th>Cirrhosis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>518</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Blood donor</td>
<td>0</td>
<td>13</td>
<td>7</td>
</tr>
<tr>
<td>Fibrosis</td>
<td>1</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Cirrhosis</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Correctly classified, %</td>
<td>99.8</td>
<td>65.0</td>
<td>41.7</td>
</tr>
</tbody>
</table>

Table 3b: Classification with zlog values (overall accuracy 97.0 %)

successful, while the separation of fibrosis cases is the worst. The zlog values almost always perform better than the z-scores. Equally high accuracy results are only achieved with neural networks for blood donors and cirrhosis, as well as for blood donors with the other three methods. The poor performance of the Support Vector Machines in the detection of fibrosis is remarkable. Here, the Neural Networks are superior to all other methods.

Discussion

In medical diagnostics and prognostics, zlog transformation has already proven to be a useful scaling and normalization method for laboratory data. Published fields of application include standardization of electronic health records [4], plausibility testing of reference intervals [24], or outcome prediction in severely ill children [25]. This paper adds a new aspect to this list by introducing zlog values as an alternative to z-score scaling in the preparation of laboratory data for machine learning.

One main difference between the two approaches is that the calculation of z-scores is based on the entire data set including all pathological values, while the zlog transformation only refers to the central 95% of healthy individuals [4]. In the case of z-scores, this results in a comparatively large scatter in the denominator of Equation (1) and thus in a substantial loss of information due to the concentration of the majority of values around zero. In addition, the position and variance of z-scores is strongly influenced by individual extreme values so that they are frequently shifted to the right in a relatively unpredictable manner (Figure 1).

Another important advantage of zlog values over z-scores is that other important covariates are taken into account for the normalization. In our case, we only distinguished between women and men for computing the reference intervals and the zlog value. But especially when children are part of the cohort, age-dependent reference intervals play a crucial role and the zlog values would automatically account for the age-dependence.

The zlog values are based on comparatively stable parameters of position and variance derived from healthy individuals, so that they do not suffer from these disadvantages. They retain the information contained in the original data and are mostly projected onto a scale ranging from roughly −10 to +10. Both increased and decreased pathological values are about equally distributed, while the z-score normalization often loses information about disease states with decreased values. This is a possible explanation why our classification experiments are poor in detecting fibrosis and cirrhosis of the liver, since these severe disease states are characterized by low production rates of proteins such as albumin, cholinesterase or alanine aminotransferase [5, 24].

One limitation of the zlog approach is that the calculation requires reference limits, which may not always be available in data mining and machine learning projects. In our study, reference intervals were calculated according to the recommended standard method [11] as the 2.5th and 97.5th percentiles of a relatively large non-diseased cohort included in the dataset (Table 1). If no such well-defined subsets are available, reference limits can usually be derived from the assay insert sheet provided by the manufacturer or derived from the data itself with a so-called indirect method, as long as the proportion of diseased individuals is not too high [26]. The R packages reflimR or refineR can be used for this purpose [27, 28].

In our study, we provide many examples to demonstrate that the zlog standardization is clearly superior to z-scores, whenever exploratory data analysis and machine learning algorithms are applied to laboratory data. This is true both for supervised and unsupervised techniques. We achieved particularly impressive results with zlog values when using...
modern methods such as t-SNE (Figure 5) and UMAP (Figure 6) for dimensionality reduction, as well as in the differentiation of the three hepatitis states with LDA (Table 3) and artificial Neural Networks (Supplementary Table 1). Although our experiments with a publicly available HCV dataset are quite conclusive, the applicability of the zlog approach should be evaluated with a broader range of clinical examples in the future.

**Research ethics:** Not applicable.

**Informed consent:** Not applicable.

**Author contributions:** The authors have accepted responsibility for the entire content of this manuscript and approved its submission.

**Competing interests:** The authors state no conflict of interest.

**Research funding:** None declared.

**Data availability:** Open source data were used: https://archive.ics.uci.edu/dataset/571/hcv-dataset.

**References**


3. HCV data. UCI machine learning repository. Available at: https://archive.ics.uci.edu/ml/datasets/HCV.


**Supplementary Material:** This article contains supplementary material (https://doi.org/10.1515/labmed-2024-0051).