Standardisation of Data Set under Different Measurement Scales

K. Jajuga, M. Walesiak

Wroclaw University of Economics,
Komandorska 118/120, 53-345 Wroclaw, Poland

Abstract: Standardisation of multivariate observations is the important stage that precedes the determination of distances (dissimilarities) in clustering and multidimensional scaling. Different studies (e.g. Milligan, Cooper (1988)) show the effect of standardisation on the cluster structure in various data configurations. In the paper a survey of standardisation formulas is given. Then we consider the problem of different scales of measurement and their impact on:

- the selection of the standardisation formula;
- the selections of the appropriate dissimilarity (or similarity) measure.

1 The measurement scales of variables

In the theory of measurement four basic scales are distinguished: nominal, ordinal, interval and ratio. Among the four scales of measurement, the nominal is considered the lowest. It is followed by the ordinal scale, the interval scale, and the ratio scale, which is highest. They were introduced by Stevens (1959). The systematic of scales is related to the transformations, which retain the relations of respective scale. This is summarised in Table 1.

One of the basic rules in the measurement theory is the following one: the numbers being the results of the measurement on the stronger (higher) scale can be transformed to the numbers on the weaker (lower) scale. The transformation of values from weaker scale to stronger scale is not permissible, since this means increasing the amount of available information. Anderberg (1973) presents some approximation methods of transformation from weaker scale to stronger scale by using some additional information.

A general and important guideline is that the statistics based on a lower level of measurement can be used for a higher scale of measurement, since permissible functions for higher scale are also permissible for lower scale.

Hand (1996) discusses the problem of relationship between measurement scales and statistics. He presents the major theories of measurement and describes the different kinds of models which may be derived within each theory. He shows in this article several examples, which have been the source of confusion and controversy.

R. Decker et al. (eds.), *Classification and Information Processing at the Turn of the Millennium*

Scale	Basic Empirical Operations	Allowed Mathematical Transformations	Allowed Arithmetic Operations
Nominal	equal to ($x_A = x_B$), not equal to ($x_A \neq x_B$)	$z = f(x)$, $f(x)$–any one-to-one correspondence function	counting of events (numbers of relations equal to, not equal to)
Ordinal	above and greater than ($x_A > x_B$), smaller than ($x_A < x_B$)	$z = f(x)$, $f(x)$–any strictly increasing function	counting of events (numbers of relations equal to, not equal to, greater than, smaller than)
Interval	above and equality of differences $x_A - x_B = x_C - x_D$	$z = bx + a$ ($b > 0$), $z \in R$ for all possible values x in R. The zero value on this scale is usually assumed, either arbitrarily or by the convention	above and addition, subtraction
Ratio	above and equality of ratios ($\frac{x_A}{x_B} = \frac{x_C}{x_D}$)	$z = bx$ ($b > 0$), $z \in R_+$ for all possible values x in R_+. The natural origin of the ratio scale is zero (this scale is bounded from the left)	above and multiplication, division

Source: Adapted from: Stevens (1959), p. 25, 27; Walesiak (1995), p. 189-191.

Table 1: The Rules for Scales of Measurement

2 Standardisation of variables

Multivariate statistical methods often require that the scales of measurement of all variables are either the same or at least similar (as similar interval and ratio scale are considered as well as nominal and ordinal). In addition, in many multivariate statistical methods, like clustering or multidimensional scaling, one has to standardise the variables.

The purpose of standardisation is to adjust the size (magnitude) and the relative weighting of the input variables (see e.g. Milligan, Cooper (1988), p. 182). The standardisation is used when the variables are measured on interval or ratio scale. In the case of nominal and ordinal scales, standardisation is not necessary, because on nominal and ordinal values such relations as equality of differences and equality of ratios are not permitted.

The only permissible transformations on the interval and ratio scale are linear transformations, thus the standardisation formulas are of the following type (Walesiak (1990)):

$$z_{ij} = bx_{ij} + a \quad (b > 0), \tag{1}$$