Rescaling Non-metric Data to Metric Data Using Multi-Dimensional Scaling

Kelley M. Engle and Guisseppi A. Forgionne

Department of Information Systems
University of Maryland Baltimore County (UMBC)
Baltimore, MD 21250 USA
{kelley.engle,forgionn}@umbc.edu

Abstract. Rescaling of nominal- and ordinal-scaled data to interval-scaled data is an important preparatory step prior to applying parametric statistical tests. Without rescaling, the analyst typically must resort to non-parametric tests that are less robust statistically than the metric counterparts. Multi-dimensional scaling (MDS) is a procedure that can be used to perform the desired rescaling. This paper utilizes MDS to transform nonmetric data from the IAN (Interactive Autism Network) and illustrates the application of the results to autism. Two simulated distributions were created from the MDS procedure to determine the best transformation. The tests reveal that either a normal or uniform distribution is acceptable with the uniform distribution performing marginally better than the normal.

Keywords: rescaling techniques, MDS (Multi-dimensional scaling), parametric test requirements, autism, data mining.

1 Introduction

Mining techniques can detect important patterns in voluminous data. To explain the patterns, however, it is often necessary to employ statistical methods that rely on metric measurements. Rescaling of nominal- and ordinal-scaled data to interval-scaled data can facilitate the needed statistical analysis. The study of autism offers a pertinent illustration.

Autism is a significant medical disorder affecting many children, their parents, and society. Data on the disorder is available from the IAN (Interactive Autism Network) project[1] [1] and consists of 7,269 children between the ages of 0-18 that have been diagnosed with some form of autism or related disorder (i.e., PDD-NOS also referred to as Pervasive Developmental Disorder – Not Otherwise Specified). Understanding the factors involved in the disorder and the potential interrelationships between these factors can help medical professionals design appropriate preventative and rehabilitative programs to reduce the consequences of the disorder.

1 Data used in the preparation of this article were obtained from the Interactive Autism Network (IAN) Research Database at the Kennedy Krieger Institute and Johns Hopkins Medicine – Baltimore, sponsored by the Autism Speaks Foundation, version dated 2.0.2. For up-to-date information see www.ianproject.org

N. Aykin (Ed.): Internationalization, Design, LNCS 5623, pp. 245–253, 2009.
© Springer-Verlag Berlin Heidelberg 2009

A number of statistical tools can be used to facilitate understanding. The most robust tools, such as ANOVA and multiple regression, depend on interval-scaled data to properly perform the analyses [2]. Yet, much of the IAN data are ordinal and nominal (non-metric) in nature. There is value, then, in rescaling the ordinal/nominal data to interval-scaled data prior to applying any statistical analyses.

There are a number of different methods that can be used for the rescaling of ordinal/nominal data including IRT (Item Response Theory) [3], MDS (Multi-dimensional scaling), Thurstone's Case V method [4], and Osgood's semantic differential [4], among others. A preliminary correlation analysis with the available IAN data revealed that many of the variables were inter-correlated. Consequently, using a method, such as IRT (Item Response Test), was deemed inappropriate since it assumes variable independence. Instead, MDS was used to scale the data because of the variable inter-correlations. The research question was to determine which, among competing distributions, provided the best metrically scaled IAN data for analytical purposes.

2 Background

Scale construction refers to the process of developing an instrument designed to elicit information from a user group. This instrument should be designed in a way that preserves internal and external validity [5, 6, 7]. Numerous methods have been proposed for scale development, including Thurstone's psychophysics methods [8, 9, 10]. Rasch's model is also commonly used in item scaling [11, 12, 13].

The IRT (Item Response Test) [3] is an individual item analysis that was designed to address weaknesses in CTT (Classical Test Theory). These weaknesses included the CTT model's inability to accommodate nominal or ordinal data [14]. CTT is considered a group assessment model and consequently is deemed by some to be weak because the assumptions required by the theory are easily met [15]. Others, such as Harwell and Gatti [16], believe that CTT's deficiencies include "its inability to produce an interval scale for test scores and its failure to take the characteristics of items into account or to provide information about the reliability of estimated scores or proficiencies."

The are four basic steps involved in rescaling ordinal (or nominal) data to interval data using IRT: (1) Identifying an appropriate IRT model; (2) Estimate item parameters; (3) Estimate proficiency parameters and; (4) Assess model-data fit [16]. The model chosen will be determined by the type of item response data and the kinds of item parameters. For instance, one of the simplest models to use is the one-parameter Rasch model [13] [16]. IRT, depending on the model chosen, has specific requirements that must be met. Tests, such as the chi-square goodness-of-fit test for individual items, will determine if the IRT model chosen is a good fit for the data [16].

MDS (Multidimensional scaling) is a scaling technique which is primarily used to find hidden patterns among data by projecting the data into a 2 – or 3-dimensional space [17]. A simple analogy given by Kruskal and Wish [17] explains the basic concept quite well:

"Suppose you are given a map showing the locations of several cities in the United States and are asked to construct a table of distances between these cities. It is simple matter to fill in any entry in the table by measuring the distance between the cities with a ruler, and converting the ruler distance into the real distance by using the scale of the map (e.g., one cm. = 30 kilometers). Now consider the reverse problem, where you are given the table of distances between the cities, and are asked to produce the map. In essence, MDS is a method for solving this reverse problem."

MDS incorporates a proximity parameter which measures the distances between points. Depending on the type of MDS model, these measurements can be for metric or non-metric data Metric MDS is closely related to factor analysis and calculates the distance based on the Euclidean formula. Non-metric data assume an ordinal relationship among the points, and non-metric MDS analyzes only the rank order of the items [17]. Shepard [18] made evident that is was possible to derive metric results from ordinal data and developed the first computer program to conduct this "nonmetric" multidimensional scaling.

A central concept in MDS is defining the objective function, or as defined by Kruskal, the stress measure (a more detailed account of the stress function can be found in [19]). The stress function measures the badness-of-fit for the MDS results and is essentially a "residual sum of squares" which determines whether or not a perfect monotonic relationship exists between the dissimilarities and the distances [19].

In typical applications of MDS, methods for obtaining the proximities generally include asking respondents to judge the distance between two items (i.e., asking respondents to judge how closely Coke and Pepsi taste) [17]. To utilize MDS as a rescaling technique for secondary data, an ideal distribution of the ordinal data is required to determine the proximity or distance between the ideal and the actual data values. This distribution should match the population distribution (i.e., uniform, normal distribution), if known.

Preliminary data analysis suggested that MDS would be an appropriate rescaling technique for the IAN data. This conclusion was reached because of the nonmetric nature of the data and the interrelationships that existed between the variables.

4 Data Collection and Analysis

There were a number of data cleansing activities required to prepare the sample dataset for the MDS testing. These activities included determining the inclusion criteria; for example, we only included autistic children who had corresponding treatment data as well as SCQ (Social Communication Questionnaire) scores. As a result of these criteria, out of the 7,269 autistic children, only 3,926 were included. It was considered critical to have treatment data for each child since for future data mining/statistical analysis activities, linking treatment to outcomes, is a primary consideration. Table 1 lists all the 10 attributes that were selected from the IAN dataset as well as a scale nature and description.

Table 1. Data Description

ATTRIBUTE	SCALE	DESCRIPTION
First ASD Diagnosis	Nominal	Captures the first ASD diagnosis such as ASD (Autism Spectrum Disorder), CDD (Childhood Disintegrative Disorder), PDD-NOS (Pervasive Development Disorder – Not Otherwise Specified).
Influenced Decision	Nominal	Lists the professional(s) that most influenced the parent's decision to begin treatment (i.e., pediatrician, psychiatrist, teacher etc.).
Who Prescribed	Nominal	If a treatment required a prescription this attribute contains the information on the prescriber (i.e., Primary care pediatrician, clinical psychologist, speech pathologist etc.).
Funding Source	Nominal	If a treatment is covered under a funding source such as the public school system, state early childhood program or other source of public funding (excluding Medicaid), then this attribute will list the funding source.
Work up Satisfaction	Nominal	Rates parent's satisfaction with an evaluation or work-up prior to commencing treatment Consists of a Yes/No/Not Applicable response.
Expected Improvement	Ordinal	A 5-pt Likert scale response rating the parent's expectations for this treatment (prior to starting).
Potential Risk	Ordinal	A 5-pt Likert scale response rating the parent's perception of potential risks associated with a treatment.
Expected Burden	Ordinal	A 5-pt Likert scale response rating the parent's expectations for this treatment (prior to starting).
Insurance Coverage	Nominal	Determines if any of the cost associated with the treatment is being covered either by private health insurance or by Medicaid.
9 point Likert Scale Treatment Efficacy	Ordinal	This is a combination of three different attributes from the original dataset which captures the parent's assessment of the efficacy of the particular treatment.

Due to some anomalies in the data, a number of records were further excluded from the dataset. These anomalies included conflicting data values for the treatment efficacy or where the parent refused to rate the efficacy of a treatment. After completing all data cleansing tasks, there were 3,283 autistic children with a total of 14,351 corresponding treatment records – on average, each child is currently receiving 4.37 treatments.

Once all data cleansing was complete, the dataset was then randomly split into two separate samples – this was performed within SAS [20] using the surveyselect procedure. The creation of two separate samples was performed to avoid the Bonferroni correction [21] which states that for every 20 hypothesis tests, there will be one result, purely by chance, that has significant results (i.e., an alpha value = .05). Therefore, to account for the fact that chance may create statistical significance; two tests were run to confirm all results.

4.1 Results for Spearman's Chi-Square Test

Because of the nonmetric nature of the data, a chi-square test was performed to determine if there existed any correlation among the variables. This step was necessary to determine if IRT (Item Response Test) could be used for rescaling. The results of the test indicated that there were multiple correlations between the variables in both samples and therefore IRT was deemed not a good fit for this dataset.

4.2 MDS

Due to the correlations discovered in the chi-square test, multidimensional scaling (MDS) was chosen as the rescaling method in lieu of IRT (which requires independent variables). MDS does not have this limitation regarding variable independence.

Two simulated data files were created in SAS – one with a uniform distribution and one with a normal distribution. These files will become the ideal measure that the original dataset will be compared to during the MDS procedure.

Our research question was to determine which of the two ideal measures – uniform or normal – generated the best data fit for the MDS procedure. Best fit is further defined as the badness-of-fit criterion [19], which will be reported for each test. The following lists the null and alternative hypotheses:

Null and Alternative Hypotheses

H_0: There is no difference in the results between the normal and uniform distribution when using the MDS procedure.

H_A: There is a difference in results between the normal and uniform distribution when using the MDS procedure.

The hypotheses were tested through the following MDS comparisons:

1. MDS – Sample1 and uniform data
2. MDS – Sample1 and normal data
3. MDS – Sample2 and uniform data
4. MDS – Sample2 and normal data

For each MDS procedure, the data from each sample were combined with the distribution data and a new variable called subject was added to the dataset. The purpose of adding the subject attribute was to allow the MDS procedure to compare the two groups. The definition for subject is user-defined – for our purposes we called the sample data S and the distribution data either U for uniform or N for normal. MDS requires the number of records to be a multiple of the number of variables; consequently, the total records in each of the combined datasets was 14,350 (a multiple of 10).

Dimensionality Issue. Choosing the number of dimensions for the MDS procedure required a more detailed investigation. Dimensions refer to the number of coordinate axes, or as described in [17] "the number of coordinate values used to locate a point in the space." Typically, MDS is used with either 2- or 3-dimensions with little regard to the true dimensionality of the data. This typical use is due to an ease of use factor that considers result interpretation. The MDS procedure will produce plots of the transformed data and therefore, when using more than 2-dimensions, the ability to visualize the plotted data is obscured. This obscuring effect is particularly true for the plot of configuration and plot of dimension coefficients for each subject that is generated within SAS.

For the goal of rescaling non-metric data, the final outcome for this procedure was the actual transformed data – not the plots themselves. Therefore, we were more interested in determining the correct dimensionality for this group of variables then we were in preserving the "ease of interpretability [17]." To determine the true dimensionality, there is a "good statistical method" that is discussed in more detail in Kruskal and Wish [17].

The simplest method to determine dimensionality was to assume that each variable corresponds to one axis, or dimension. However, as stated by Kruskal et al. in [17], "Although a dimensional interpretation frequently involves one interpretation for each dimension of the space, the dimensionality is not necessarily the number of relevant characteristics involved." One or more variables may not greatly affect the MDS configuration due to correlation with other variables or because these variables only affect a small subset of the data.

Table 2. Dimensionality Test

Number of Dimensions	Badness-of-fit Criterion
2	24.24%
3	18.41%
4	14.32%
5	11.91%
6	9.25%
7	8.12%
8	6.24%
9	4.46%

Stress, also called the badness-of-fit criterion (or goodness-of-fit criterion in earlier literature), is crucial in determining the number of dimensions to choose. All other constants held equal and assuming complete convergence, increasing the number of dimensions should decrease the badness-of-fit measure. To verify this effect, multiple experiments were conducted utilizing the MDS procedure while varying the dimensions from two to nine. The data for this test were a combination of the data from sample 1 and the simulated uniform distribution. As can be seen from the results in Table 2, for every dimension added to the MDS procedure the stress level, or badness-of-fit criterion, decreases. With the maximum number of dimensions chosen (9),

Table 3. Kruskal's Stress Heuristics

Stress	Badness-of-fit
20%	Poor
10%	Fair
5%	Good
0%	Perfect

the stress level is at approximately 4.5%, which is considered good according to Kruskal [19]. As a result of these findings, the four MDS experiments were conducted using the maximum nine dimensions.

4.3 MDS Results

The results for each of the experiments are shown below in Table 4. The uniform distribution performed slightly better than the normal distribution in determining the badness-of-fit criterion. Due to the cross-comparisons by variable, the transformed data set consisted of 129,150 records. These data were outputted to the RES dataset and can be found in the TRANDATA column (See appendix for the MDS procedure code).

Table 4. Overall Results – Badness-of-fit Criterion

Sample	Uniform	Normal
Sample 1	5%	5%
Sample 2	5%	6%

5 Findings

After evaluating the badness-of-fit criterion for the four MDS experiments, there is notably very little difference between the uniform and normal distributions. The uniform distribution had slightly better results but not enough to definitely reject the null hypothesis. Therefore, for these experiments, the null hypothesis is accepted and for all practical purposes either distribution can be used to transform the data. These findings are limited to only the sample used; for future releases of the IAN data these results may or may not hold true. The experiments then should be recreated to confirm or refute these results.

Another issue to consider with the transformed data is the sheer number of records created. The IAN project is expected to register up to 100,000 autistic children. Given a larger sample size as is projected for the IAN project, the number of transformed records will grow exponentially. Statistical programs, such as SAS and SPSS, are typically not used for extremely large datasets. One such solution would be to randomly sample the transformed data prior to performing statistical analyses.

Alternatively, a random sample of the original data can be obtained prior to the transformation in order to decrease the total number of transformed records.

6 Conclusion

MDS (Multi-dimensional scaling) is a technique that is predominantly used for detecting patterns along a multi-dimensional space. However, MDS can be used to transform nonmetric scaled data to a metric scale. The details of this process have been described in this paper as a practical aide to implementation. Prior to actually executing the MDS procedure in SAS some preprocessing is necessary. This includes: (1) Data prep and cleansing and; (2) Creating a simulated file using either a uniform or normal distribution to match the original data. Ideally, the number of dimensions should be the number of variables included in the dataset (n) – 1 in order to assure the best fit. After executing the MDS procedure, the badness-of-fit criterion should be analyzed for sufficiency according to Kruskal's heuristics.

The overall purpose of this paper was to present a viable method for rescaling nominal and ordinal data to interval-scaled data in the presence of significant interrelationships among the studied variables. By rescaling the data, more powerful parametric statistical tests can be performed on the transformed data. The rescaling method used was MDS (Multi-dimensional scaling) and was carried out on the IAN (Interactive Autism Network) data, which is dataset of approximately 5 000 autistic children from across the United States. A number of variables were extracted from this dataset and rescaled using the MDS procedure. The research question was to determine whether to use a normal or uniform distribution for the ideal measure in MDS. The experiments indicate that there is no significant difference in the badness-of-fit criterion for these two distributions. Therefore, researchers requiring metric data for their analyses could use either the normal or uniform distribution with the MDS procedure to transform the IAN data.

References

1. IAN (Interactive Autism Network) (homepage on the Internet), Kennedy Krieger Institute, Baltimore (2007-2008), `http://www.ianproject.org` (cited November 13, 2008)
2. Cooper, D.R., Schindler, P.S.: Business Research Methods, 9th edn. Mc-Graw Hill Irwin, New York (2006)
3. DeVellis, R.F.: Scale Development Theory and Applications, 2nd edn. Sage Publications, Thousand Oaks (2003)
4. Green, P.E., Tull, D.S.: Research for Marketing Decisions, 4th edn. Prentice-Hall, Inc., Englewood Cliffs (1978)
5. Dawis, R.V.: Scale Construction. Journal of Counseling Psychology 34(4), 481–489 (1987)
6. Edwards, A.L.: Techniques of Attitude Scale Construction. Appleton-Century-Crofts, Inc., New York (1957)
7. Dunn-Rankin, P.: Scaling Methods. Lawrence Erlbaum Associates Publishers, Hillsdale (1983)
8. Thurstone, L.L.: A law of comparative judgment. Psychological Review 34, 273–286 (1927)

9. Thurstone, L.L.: The method of paired comparisons for social values. Journal of Abnormal and Social Psychology 21, 384–400 (1927)
10. Thurstone, L.L.: Psychophysical analysis. American Journal of Psychology 38(3), 368–389 (1927)
11. Andrich, D., Luo, G.: A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. Applied Psychological Measurement 17(3), 253–276 (1993)
12. Andrich, D.: Relationships between the Thurstone and Rasch approaches to item scaling. Applied Psychological Measurement 2(3), 449–460 (1978)
13. Wright, B.D., Masters, G.N.: Rating Scale Analysis. Mesa Press, Chicago (1982)
14. McCartney, K., Burchinal, M.R., Bub, K.L.: Best Practices in Quantitative Methods for Developmentalists. Blackwell Publishing, Boston (2006)
15. Sapp, M.: Basic Psychological Measurement, Research Designs, and Statistics without Math. Charles C. Thomas Publisher, Ltd., Springfield (2006)
16. Harwell, M.R., Gatti, G.G.: Rescaling ordinal data to interval data in educational research. Review of Educational Research 71(1), 105–131
17. Kruskal, J.B., Wish, M.: Multidimensional Scaling. Sage Publications, Beverly Hills (1978)
18. Shepard, R.N.: The analysis of proximities: multidimensional scaling with an unknown distance function. Psychometrika 27, 125–140, 219–246 (1962)
19. Kruskal, J.B.: Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 29(1), 1–27 (1964)
20. SAS Software (homepage on the Internet). Business Intelligence Software and Predictive Analytics, Cary, NC (2008), `http://www.sas.com` (cited November 13, 2008)
21. Abdi, H.: Bonferroni and Sidak corrections for multiple comparisons. In: Salkind (ed.) Encyclopedia of Measurement and Statistics, Sage, Thousand Oaks (2007)