Data Standardization for Enhanced Outcomes in Cluster Analysis Using JMP®

Mantosh Kumar Sarkar[1] , Karthik Nakkeeran[1] and Goutam Chakraborty[2]
[1]Management Information System, Oklahoma State University, Stillwater, OK 74078
[2]Professor (Marketing) and Founder of SAS and OSU Business Analytics Program, Oklahoma State University, Stillwater, OK 74078

Introduction

Data transformation plays a critical role in achieving better results from cluster analysis, a popular technique used for market segmentation. Customer survey data used for market segmentation via cluster analysis often exhibits response styles of respondents. Response styles like acquiescence and extreme response styles can bias the results of cluster analysis. Recent research has shown that double-standardization (standardizing across both row and column) and range standardization are good candidates for eliminating response styles from dominating the cluster results, Pagolu et al. (2011). But, JMP does not provide these special standardization features.

This paper shows how to achieve double-standardization in JMP by using its powerful scripting language (JSL). To test if our code is working properly, we have used customer survey data from a business-to-business company in hydraulic and pneumatic industry. A 9-point rating scale is used to measure customers' responses to questions as shown in Fig.1 below.

Methods

JSL Code

In a recent SAS GF paper, Pagolu et al. (2011) found that double-standardization often works well in survey data that is used for clustering. In double standardization method, the scores are first adjusted within the case and then the resulting scores are adjusted within the attribute (Fischer, 2004).

```
/*Code For Double Standardization*/
/*Authored by Mantosh Kumar Sarkar and Karthik Nakeerran*/
/*Open the dataset*/
SourceDt = Open( "C:\JMP_Discovery_Summit_2012\bfsurvey.jmp" );
/*Sort the data set by identification number*/
SourceDt << Sort(
    By( id ),
    order( Ascending ),
    Output Table
    Name( "C:\JMP_Discovery_Summit_2012\Temp1.jmp" )
);
/*Save the sorted dataset*/
SourceDt << save( "C:\JMP_Discovery_Summit_2012\Temp1.jmp" );
Close( SourceDt );
```

```
Temp1 = Open( "C:\JMP_Discovery_Summit_2012\Temp1.jmp" );

/*Aquire the column names from the data set*/
cols = Temp1 << Get Column Names( Continuous );

/*Column Standardize each data element and prefix column name
For( i = 1, i < N Items( cols ), i++,
    Eval(
        Eval Expr(
            col = Temp1 << New Column( "Std" || Char( cols[i
                Numeric,
                Continuous,
                Formula( Col Standardize( Expr( Cols[i] ) ) )
            )
        )
    );
    col << Set Selected;
);

Column( "ID" ) << Set Selected;
```

```
/*Save the column standardized dataset*/
```

Fig.1: Double Standardization JSL code

The main steps involved in writing JSL code(Fig.1) for double-standardization are as follows:

- Sort the dataset
- Column standardize the variables
- Transpose data
- Column standardize the data again
- Transpose data.

This procedure will provide you with double-standardized data, which can then be used for clustering.

Double standardized data helps in reducing the skewness, kurtosis in the data as evident from Fig.2.

The clusters are more refined and meaningful when hierarchical clustering is performed on double standardized data versus untransformed (raw) data(Fig.4). This is likely because of response styles in the raw data.

Fig.3 shows the profile of clusters obtained from original (raw) data versus the double standardized data. The profiles seem to be much better for the double standardized data than the original (raw) data.

Results

How important are the following issues to customers in choosing a supplier of automobile parts	Attribute	Scale	Not at all important	Extremely important
The reliability of the supplier	reliab	9 point	1	9
The timeliness of the deliveries by the supplier	time	9 point	1	9
The availability of a large breadth of products to choose from	av_br	9 point	1	9
The availability of well documented technical specification	av_spec	9 point	1	9
The price of products	price	9 point	1	9
The credit policy of the supplier	credit	9 point	1	9
The availability of electronic payment/debit option	av_pay	9 point	1	9
The return policy of the supplier	return	9 point	1	9
The warranty coverage provided by the supplier	warranty	9 point	1	9
The ability to talk directly to a salespeople about your needs	talk_dir	9 point	1	9

Table 1: Important factors for customers in selecting a supplier for the hydraulic and pneumatic products

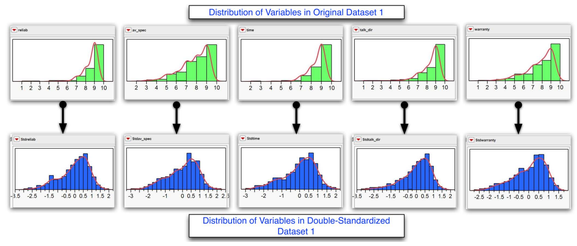

Fig.2: Distribution of few key variables

Fig 3: Cluster profile of original dataset vs Double standardized dataset

Fig4: Dendogram of original dataset vs. Double standardized dataset with Cluster history

Discussion

- The histograms and the hierarchical clustering results shows that double standardized data work better when compared to running clustering on untransformed (raw) data.
- At the recent analytics 2011 conference, a presenter suggested doing double-standardization not just once (as has been done here) but several times in succession until one finds no significant improvement. We would like to explore that in future research to see if that adds any significant value beyond doing double standardization once.

Reference

- Pagolu, & Chakraborty, (2011). Eliminating Response Style Segments in Survey Data via Double Standardization Before Clustering. SAS® Global Forum
- Fischer, R. (2004). Standardization to account for cross-cultural response bias - A classification of score adjustment procedures and review of research in JCCP. Journal of Cross-Cultural Psychology, 35(3), 263-282.
- SAS Institute Inc. 2012. *JMP® 10 Scripting Guide*. Cary, NC: SAS Institute Inc.

Acknowledgement

Authors would like to acknowledge the Business-to-Business company that provided the data for this research but wishes to remain anonymous.