Market Segmentation clustering methods and metrics

By

Narla Venkata durga bhavya sri

Algorithms with Integrated Variable Selection

Even after preprocessing techniques used on segmentation variables(binary data) for selecting best segmentation variables for efficient market segments there will be a problem of variables redundancy or noisy variables .for such problem we have presented two algorithms for binary segmentation variables:

1.biclustering

2.variable selection procedure for clustering binary data (VSBD)

Biclustering Algorithms

Biclustering, also known as co-clustering or two-way clustering, is a clustering algorithm used in market segmentation to identify subsets of customers who exhibit similar behaviors or preferences across multiple attributes or variables. It is a data mining technique which allows simultaneous clustering of the rows and columns of a matrix

Algorithm follows:

Step 1:rearrange the matrix with rows(consumers) and columns(variables).

Step 2: observe the common no of consumers with more no of variables covered and form them into a rectangle(form a rectangle of as many as possible).

Step 3:after clustering ,go with remaining rows and columns and continue the process until theres is no data samples without clustering.

Advantages of biclustering algorithms

1.No need to transform data: no need to preprocess large data.

2.Cpaturing niche markets: Because biclustering searches for identical patterns displayed by groups of consumers with respect to groups of variables, it is well suited for identifying niche markets.

Variable Selection Procedure for Clustering Binary Data(VSBD)

variable selection procedure for clustering binary datasets is based on the k-means algorithm as clustering method.

Algorithm follows:

Step 1:Select only a subset of observations with size $\phi \in (0, 1]$ times the size of the original data set.

Step 2: Given V (number of variables), perform an exhaustive search for the set of V variables that leads to the smallest within-cluster sum-of-squares criterion.

Step 3: Among the remaining variables, determine the variable leading to the smallest increase in the within-cluster sum-of-squares value if added to the set of segmentation variables.

Step 4: Add this variable if the increase in within-cluster sum-of-squares is smaller than the threshold. The threshold is δ times the number of observations in the subset divided by V.

Variable Reduction: Factor-Cluster Analysis

Variable reduction using factor cluster analysis is a statistical technique used to identify groups of variables that are highly correlated with each other, and to reduce the number of variables in a dataset while retaining the most important information. The technique involves two steps: factor analysis and cluster analysis.

Factor analysis is used to identify groups of variables that are highly correlated with each other, and to reduce the number of variables into a smaller set of underlying factors. This is done by analyzing the correlation matrix of the variables and extracting the principal components or factors that explain the most variance in the data.

Cluster analysis is then used to group the variables into clusters based on their similarities in terms of the underlying factors. This is done by using a clustering algorithm to group the variables based on their distances or similarities in terms of the factor loadings.

factor cluster analysis is a smaller set of variables that are highly correlated with each other, and that capture the most important information in the original dataset.

Gorge plots

Gorge plots, also known as density trace plots, are a type of data visualization tool used to display the distribution of a continuous variable in two or more groups or conditions.

Gorge plots typically consist of a series of density plots or kernel density estimates, each representing the distribution of the continuous variable in a different group or condition. The plots are arranged side-by-side, with a gap or "gorge" between each plot to visually highlight differences in the distributions.

Gorge plots are particularly useful for comparing the shape and spread of distributions between groups or conditions, and for identifying differences in key summary statistics such as mean, median, and standard deviation.

Data structure Analysis

Data structure analysis provides valuable insights into the properties of the data. These insights guide subsequent methodological decisions. Most importantly, stability-based data structure analysis provides an indication of whether natural, distinct, and well-separated market segments exist in the data or not. If they do, they can be revealed easily. If they do not, users and data analysts need to explore a large number of alternative solutions to identify the most useful segment(s) for the organisation.

Cluster indices

Cluster indices are quantitative measures used to evaluate the quality of clustering results. They provide a way to assess how well the clustering algorithm has performed and how well the resulting clusters are separated from each other.

Cluster validation indices are numerical measures that can be used to compare different clustering results, or to determine the optimal number of clusters for our data.

Internal indices

These indices do not require any external information or labels to evaluate the clustering.

A very simple internal cluster index measuring of clusters results from calculating the sum of distances between each segment member and their segment representative. Then the sum of within-cluster distances Wk for a segmentation solution with k segments is calculated using the following formula where we denote the set of observations assigned to segment number h by Sh and their segment representative by ch:

$$W_k = \sum_{h=1}^{k} \left(\sum_{x \in S_h} (d(x, c_h))\right)$$

An optimal market segmentation solution contains market segments that are very different from one another, and contain very similar consumers. This idea is mathematically captured by another internal cluster index based on the weighted distances between centroids (cluster centres, segment representative) Bk:

$$B_k = \sum_{h=1}^{k} \left(\sum_{x \in S_h} (n_h d(c_h, c))\right)$$

The Ratkowsky and Lance (1978) index is based on the squared Euclidean distance and uses the average value of the observations within a segment as centroid. The index is calculated by first determining, for each variable, the sum of squares between the segments divided by the total sum of squares for this variable. These ratios are then averaged and divided by the square root of the number of segments. The number of segments with the maximum index value is selected. This is method is used in VSBD.

External indices

External cluster indices using additional external information for market segmentation evaluation ; i.e, no enough with information contained in one market segment (different additional pieces of information can be used.).

For suppose, Selecting any two consumers, the following four situations can occur when comparing two market segmentation solutions P1 and P2:

- a: Both consumers are assigned to the same segment twice.
- b: The two consumers are in the same segment in P1, but not in P2.
- c: The two consumers are in the same segment in P2, but not in P1.
- d: The two consumers are assigned to different market segments twice.

Jaccard Index: If the segment solutions are similar, then the value of “a” would be large and the values of “b” and “c” would be small. This is Jaccard Index which is defined as follows:

$$J=a/(a+b+c)$$

The value of J = 0 indicates that the two market segments are completely different whereas the value of J = 1 indicates that the two market segments are same.

Rand Index and Adjusted Rand Index:Rand(1971) proposed a similar index using all the four values:

R=(a+d)/(a+b+c+d)

Both the external indexes suffer from the problem that their values lie between 0 and 1 and are therefore difficult to interpret. The values of both Jaccard and Rand depend upon the random assignment to segments and sizes of the extracted market segments

Global stability analysis is a mathematical technique used to analyse the stability of a system over its entire range of parameter values.

Resampling methods are statistical techniques that involve repeatedly drawing samples from a dataset in order to estimate population parameters or to assess the variability and uncertainty of statistical estimates. The two main types of resampling methods are:

We use bootstrapping technique for global stability from resampling methods.

Segment Level Stability Analysis

Segment Level Stability within Solutions (SLSW)

The criterion of segment level stability within solutions (SLSW) is similar to global stability The difference is that stability is computed at segment level.

Segment level stability within solutions (SLSW) measures how frequent a market segment with the same characteristics is identified across a number of repeated calculations of segmentation solutions with the same number of segments.

Segment Level Stability Across Solutions (SLSA)

This is the second criterion of stability at segment level proposed by Dolnicar and Leisch.

The purpose of this criterion is to determine the re-occurrence of a market segment across market segmentation solutions containing different numbers of segments. High values of segment level stability across solutions (SLSA) serve as indicators of market segments occurring naturally in the data. Natural segments show more attractiveness to organisations because they actually present no need of artificial segments construction.