Assessing Divide-and-Conquer Latent Class Analysis

Qiao Ma[1], Meimeizi Zhu[1], Edward Mulrow[1]

[1]NORC at the University of Chicago, 55 E. Monroe Street, Chicago, IL 60603

Abstract

Latent class analysis (LCA) is generally used to construct unobserved classes from observed indicator variables. Fitting an LCA to a large dataset is challenging. Abarda et al. (2017) proposed a Divide-and-Conquer approach for LCA when dealing with big data. Divide-and-Conquer partitions a data set into multiple subsets with equal size, and fits a model to each subset. We assess LCA results in such a setting by fitting LCA models with and without the Divide-and-Conquer approach, and measure the relationships between the Bayesian Information Criteria (BIC) for the whole dataset and its subsets.

Key Words: Latent class analysis, Big data, Multivariate Categorical data, Statistical methods, Structural equation modeling

1. Introduction

In nowadays, large volume of data are becoming very common. They are usually applied in various areas such as marketing, health, education, customer services, etc. Researcher are eager to explore solutions for processing and analyzing data with large volume. The use of statistical methods to solve the problem for massive data have been developing rapidly.

1.1 Divide-and-Conquer Approach

Zhang et al. (2013) studied the Divide-and-Conquer kernel ridge regression (KRR), which consists of three steps: 1) partitions a data set of size N into m subsets of equal size, 2) computes an independent kernel ridge regression estimator for each subset, and 3) averages the local solutions into a global predictor. Abarda et al. (2017) proposed a Divide-and-Conquer approach for LCA (DACL) when dealing with big data. They adopted the Divided-and-Conquer approach to LCA method and provided a stop condition for the algorithm in order to minimize the number of subsets processed.

1.2 Latent Class Analysis

Latent class analysis (LCA) is generally used to construct unobserved classes from observed indicator variables, which can uncover unobserved heterogeneity in a population. It can be seen as a mixture model of independent multinomial distributions.

Basic Latent Class Model:

$$P(y_i;\theta) = \sum_{k=1}^{K} P(x_i = k;\theta_x) P\left(y_i | x_i = k;\theta_y\right)$$

$$= \sum_{k=1}^{K} P(x_i = k;\theta_x) \prod_{j=1}^{J} P\left(y_{ij} \middle| x_i = k;\theta_{y_j}\right)$$

Where y_i is a vector containing the responses of person i on J categorical variables, x_i is a discrete latent variable, K is the number of categories of x_i or the total number of latent classes (Vermunt et al., 2008).

2. Description of Approach

2.1 Assess Divide-and-Conquer Latent Class Analysis

We first assessed LCA results by adopting the framework of Abarda et al. (2017) and then a simulation study was employed to generate a data set randomly for this assessment. We simulated a dataset randomly by using *poLCA* package in R, and simulated a LCA model having 2 classes and 4 variables by respecting the assumption of local independence. The population size is N=1,000,000. This data set was divided into 10 subsets by simple random sampling without replacement so the size of each sub-population is 100,000. Finally, we fit the LCA models using entire data and each subset.

Table 1: Estimated proportions and their standard errors for the population and 10 sub-populations

	Proportion		*Errors*	
	Class1	*Class2*	*Class1*	*Class2*
Entire Data	0.2733	0.7267	0.0119	0.0119
Subset1	0.2690	0.7310	0.0122	0.0122
Subset2	0.2686	0.7314	0.0119	0.0119
Subset3	0.2778	0.7222	0.0123	0.0123
Subset4	0.2670	0.7330	0.0125	0.0125
Subset5	0.2710	0.7290	0.0122	0.0122
Subset6	0.2870	0.7130	0.0126	0.0126
Subset7	0.2709	0.7291	0.0123	0.0123
Subset8	0.2975	0.7025	0.0128	0.0128
Subset9	0.2735	0.7265	0.0120	0.0120
Subset10	0.2851	0.7149	0.0123	0.0123

As you can see from the table above, the estimated proportions and their standard errors obtained by using DACL approach and those of the entire population are very similar.

2.2 Investigate AIC and BIC

Additionally, we measured the Akaike's Information Criterion (AIC) and Bayesian Information Criteria (BIC), and the relationships between the entire data and its subsets. We firstly simulated a dataset randomly with 4 classes and 8 variables by assuming local independence. The population size is N=10,000. This data set was divided into 10, 20, 50, 100 and 200 subsets respectively by simple random sampling without replacement.

Furthermore, we fit various LCA models using entire data and each subset by assigning different number of classes. We repeated this step for each model for 30 times and then evaluated how frequent the lowest BIC value was identified for the 4 class model. A value of 100 in the bolded column indicates perfect identification.

Table 2: Percentage of Times the Lowest Value Occurred in Each Class Model for the AIC and BIC

		AIC					*BIC*				
		Classes					*Classes*				
Number of Subsets	Subset Size	2	3	***4***	5	6	2	3	***4***	5	6
1	10000	0	1	***72***	15	12	0	0	***100***	0	0
10	1000	0	0	***65***	20	15	0	0	***100***	0	0
20	500	0	0	***68***	28	4	0	0	***100***	0	0
50	200	0	1	***59***	25	15	0	4	***96***	0	0
100	100	0	2	***58***	31	9	1	10	***89***	0	0
200	50	0	0	***63***	34	3	0	17	***83***	0	0

It is observed that BIC correctly identifies the number of classes more consistently across all models and all sample size, and BIC has sensitivity to small sample sizes.

References

Abarda, A., Bentaleb, Y. and Mharzi, H. (2017), A Divided Latent Class analysis for Big Data, International Workshop on Big Data and Networks Technologies Proceedings, 428-433

Hagenaars, J. A. and McCutcheon, A. L. (2002), Applied Latent Class Analysis, New York: Cambridge University Press.

McCutcheon, A. L. (1987), Latent Class Analysis, Beverly Hills and London: Sage Publications.

Shu, H. (2016), Big data analytics: six techniques, Geo-spatial Information Science, 19, 119-128.

Vermunt, J. K., Ginkel, J. R. V., der Ark, L. A. V., and Sijtsma, K. (2008), "Multiple imputation of incomplete categorical data using latent class analysis," Sociological Methodology, 38, 369-397.

Yang, CC. (2006), Evaluating latent class analysis models in qualitative phenotype identification, Computational Statistics and Data Analysis, 50, 1090-1104.

Zhang, Y, Duchi, J. and Wainwright, M. (2013), Divide and Conquer Kernel Ridge Regression, Workshop and Conference Proceedings, 30, 1-26.