Survey Methodology, June 1990
Vol. 16, No. 1, pp. 137–143
Statistics Canada

Use of Cluster Analysis for Collapsing Imputation Classes

E.R. LANGLET[1]

ABSTRACT

The problem of collapsing the imputation classes defined by a large number of cross-classifications of auxiliary variables is considered. A solution based on cluster analysis to reduce the number of levels of auxiliary variables to a reasonably small number of imputation classes is proposed. The motivation and solution of this general problem are illustrated by the imputation of age in the Hospital Morbidity System where auxiliary variables are sex and diagnosis.

KEY WORDS: Item nonresponse; Auxiliary variables; Imputation matrix; Donors; Disjoint techniques; Hierarchical techniques; Cluster seeds.

1. STATEMENT OF THE PROBLEM

In surveys, the problem of item nonresponse occurs when some but not all information is collected for a sample unit or when some information is deleted because it fails to satisfy edit constraints. In many surveys, this problem is handled by random imputation within classes, a common form of hot deck imputation method. For this type of imputation, a respondent is chosen at random within an imputation class defined by one or more auxiliary variables and the respondent's value is assigned to the nonrespondent.

The problem considered in this paper can be defined as follows. The classifications of the respondents according to certain auxiliary variables form a multi-dimensional imputation matrix where the number of imputation classes equals the number of cross-classification cells defined by the auxiliary variables. If the number of imputation classes is very large, few or no donors may be available in several classes. In addition, manipulation of this large matrix could be very cumbersome computationally. These problems can be alleviated by collapsing the cells of the matrix either by grouping the cells themselves, or the rows, columns or along some other dimension (or combination of dimensions) so that the resulting groups will be homogeneous with respect to the variables requiring imputation. We propose to use cluster analysis to achieve the desired level of collapsing. For this purpose, the values of the variables of interest from donors (or respondents) for each imputation class can be used to assign numerical scores to each class. In this paper, measures based on empirical distribution function for respondent data are used to quantify imputation classes. Cluster analysis can then be used to group the cells of the matrix according to these numerical scores. It will be shown that cluster analysis is appropriate for the problem under consideration. Related useful references concerning the application of cluster analysis to stratify primary sampling units are Drew, Bélanger and Foy (1985), Judkins and Singh (1981) and other references contained therein.

The above mentioned problem arose in the context of age imputation in the Hospital Morbidity System (HMS). This system uses the auxiliary variables sex and diagnosis as the basis for imputing the age. The number of imputation classes were over 5,000 for each sex. A solution based on the technique of cluster analysis was proposed in order to collapse the levels of

[1] E.R. Langlet, Social Survey Methods Division, Statistics Canada, Ottawa, K1A 0T6.

the diagnosis variable to 40 groups of related diagnoses. In section 2, a brief review of the commonly used cluster analysis techniques is presented. Use of cluster analysis for the problem of collapsing imputation classes is illustrated for the example of imputation of age for the HMS data in section 3 including the relative performance of the proposed method with respect to the current method. Both methods utilize a hot deck approach but the proposed method redefines the imputation classes using cluster analysis. Some concluding remarks including possible generalizations of the method are given in section 4.

2. CLUSTER ANALYSIS TECHNIQUES: A BRIEF REVIEW

The problem of classifying a given number of entities described by a number of quantitative variables into groups such that entities within the same groups or clusters will be similar to each other and dissimilar to entities in different groups is considered in this section. A good review of clustering techniques is given by Everitt (1980) mainly based on the work of Cormack (1971). Most clustering techniques can be classified into two groups, namely 'hierarchical techniques' and 'disjoint techniques', the latter one also known as 'optimization techniques'. These two groups of techniques will be described below. Some other methods, are density techniques where clusters are formed by searching for regions containing dense concentrations of entities. This is based on the fact that if entities are described as points in a metric space, there should be parts of the space in which the points are very dense, separated by parts of low density. Another class of techniques is called clumping techniques in which the clusters can overlap. In certain fields such as language studies, for example, classification must permit an overlap between the classes because words tend to have several meanings, and if they are classified by their meanings they may belong in several places.

Hierarchical techniques can be subdivided into 'fusion techniques' and 'divisive techniques'. In fusion methods, each entity begins in a cluster by itself. At each step, the two closest clusters are fused to form a new cluster until only one cluster containing all the observations is left. In divisive techniques, all entities are first grouped into one cluster. Then, at each step, groups of the entities are successively broken down into finer partitions until each entity constitutes a cluster by itself. Hierarchical techniques differ with respects to the definition of the distance measure between observations or groups of observations. An advantage of hierarchical techniques is that a single run can produce results for one cluster to as many as you like by stopping the fusion or division process at the desired level of the hierarchy. Obviously, hierarchical techniques can be used for only small data sets since there are $n(n-1)/2$ possibilities to fuse two entities in a group of n entities and $2^{n-1}-1$ possibilities to break a group of n entities in two groups.

In contrast to hierarchical techniques where observations belong to a series of clusters depending on the level of the hierarchy, disjoint techniques divide observations into a number of clusters (generally predetermined) such that each observation belongs to one and only one cluster. They also differ from hierarchical techniques in that they admit relocation of the observations so that a poor initial partition can be corrected at a later stage. Disjoint techniques are clearly more appropriate than hierarchical techniques to handle large data sets. Disjoint techniques are also called optimization techniques because they seek for a partition of the data which optimizes some predefined criterion. Various disjoint techniques differ in the way the methods obtain an initial partition and in the clustering criterion they try to optimize. Usually, disjoint techniques start by selecting a set of points called cluster seeds as a first guess of the means of the clusters. A number of procedures have been suggested for choosing these points

(Anderberg 1973). Once the cluster seeds have been selected, the entities are then assigned to the closest cluster seeds (usually, the Euclidean distance is used). Estimates of the cluster means might be updated after each allocation (MacQueen 1967) or after all entities have been allocated (Ball and Hall 1967). Once an initial partition has been found (which is equivalent to finding a set of cluster seeds and to allocating each entity to the closest cluster seed), a search is made for entities whose re-allocation to some other group will improve the clustering criterion. This procedure is repeated until no further move of a single entity improves the clustering criterion. A local optimum is then reached. This is what Anderberg (1973) calls 'nearest centroid sorting'. In general, there is no way to know whether a global optimum has been reached.

3. APPLICATION: FORMING IMPUTATION CLASSES FOR THE HMS

3.1 Background

The Hospital Morbidity System (Statistics Canada 1987) consists of a count of inpatient cases, discharged during the year from general and allied special hospitals in Canada except Yukon and Northwest Territories. Each record of the system contains at least one diagnosis code, the age and sex of the patient, the length of stay, *etc.* The first valid diagnosis on the record is called the tabulating diagnosis and is the diagnosis on which tabulations are based in the publications. This diagnosis can be seen as the main cause for which the patient is hospitalized and is coded according to the 9th Edition of the International Classification of Diseases (World Health Organization 1977) which contains more than 5,000 diagnoses.

The age imputation problem in the HMS is currently treated by a hot deck method. In this imputation problem to predict the age of the patient y, two auxiliary variables are used, namely the tabulating diagnosis d which is always present on the record and the sex of the patient s. The sex itself needs to be imputed first if it is missing according to the observed male/female proportions of d over previous years. Classification of the patients according to d and s forms an imputation matrix with the number of imputation classes larger than 5000 × 2. In order to reduce the dimension of the imputation matrix, diagnoses were regrouped or collapsed, based on the age distribution of each diagnosis. Let F_d denote the age distribution in the population of the patients with tabulating diagnosis d. Then, diagnoses A and B would be collapsed together if F_A is close to F_B. Estimates of F_d from available data can be used for this purpose. It should be noted that the sex variable was not used in defining imputation classes (see section 4 for details on how it could be used) although it was used in the imputation scheme. By not using the sex variable for defining imputation classes, the number of imputation classes of the imputation matrix is reduced by half.

In order to motivate the proposed method for collapsing imputation classes, we will first describe the current method and its limitations. The collapsed groups were created by comparing manually (using histograms) the shapes of the empirical age frequency distributions, $\hat{F}_d$ of all diagnosis codes corresponding to 1974 HMS data. Thirty six groups were obtained and a 37th group was created for those diagnoses for which less than 200 observations were available. The number of groups was determined a posteriori arbitrarily. The main deficiency of the current method comes from the fact that no statistical criterion was used to group diagnoses which makes the method labour intensive and somewhat subjective. These groups were obtained by simply comparing histograms. An evaluation of the current imputation method indicated that the resulting groups of diagnoses were, in a few cases, not homogeneous with respect to $\hat{F}_d$ and consequently needed to be updated.

3.2 Proposed Method

The proposed method can be briefly described as follows. We shall consider the case when only one quantitative variable needs to be imputed. Extension to cases where more than one variable requires imputation is discussed in section 4. Let's denote by y the variable to be imputed and by F_i the distribution of variable y in class i. Note that the classes are defined by the cross-classification of one or more auxiliary variables which are suitably categorized if necessary. The first step is to find an appropriate set of parameters to represent F_i in each class, for example, the first three or four moments of the F_i's or the percentiles. The next step is to estimate these parameters from the respondent data. Finally, a suitable technique of cluster analysis on the set of estimated parameters can be used to condense the number of classes such that classes grouped together will be similar with respect to the parameters representing the F_i's.

A justification for the choice of the proposed method in the context of the age imputation for the Hospital Morbidity System (HMS) will now be presented. First, consider some possible alternative strategies to the collapsing problem. One strategy for this problem might be similar to the original method that was used for 1974 data, that is, to group diagnoses according to the distributions $\hat{F}_d$ but using a statistical criterion for grouping instead of manually comparing histograms. Data would be cross-classified by tabulating diagnoses, sex and a number of age groups, say 10. Two diagnoses would be grouped together if the proportion of cases in each of these ten age groups, $p_1, \ldots, p_{10}$ were judged to be close to each other according to some criterion such as the Euclidean distance or a chi-square measure. Note that the use of a chi-square measure would cause serious computational burden since no commonly available cluster analysis program uses this distance measure. This would imply the calculation of the chi-square distance for all possible pairs of diagnoses. Another possible strategy would be to first use data reduction techniques such as principal components to reduce the dimension of age groups and then decide whether two diagnoses are close based on principal component scores. An obvious disadvantage to all these methods is the number of observations required to obtain a reliable estimate of the categorical age distribution for each diagnosis.

In view of the above problem, we decided to use the first two or three moments to approximately describe F_d. We started with three – the mean m_d, the standard deviation s_d and the skewness coefficient b_d. However, it was found by means of principal component analysis that it was not necessary to include b_d. The approach then is to collapse diagnoses according to the sample mean, m_d, and the sample standard deviation s_d. Cluster analysis can be used to provide a suitable statistical technique for this purpose. An obvious advantage with this approach over other strategies based on the categorical distribution of age is that a reliable estimation of two moments requires much fewer observations than the estimation of the proportion of cases over several age groups. In section 4, implementation of this approach is described for the problem of age imputation.

3.3 Procedure Steps in the Implementation of the Proposed Method for HMS Data

There are four steps in implementing the proposed collapsing method based on cluster analysis for the age imputation problem for HMS data.

Step I: Selection of a clustering method

Before selecting a clustering method, it should be noted that our goal is primarily to partition the diagnoses into homogeneous groups without trying to uncover 'natural' or 'real' clusters. This is called 'data dissection' in the literature (Everitt 1980). Another important consideration is the availability of a well tested clustering program using an efficient clustering method.The determinant consideration for the selection of a clustering method was the number of observations in our data set which resulted in the selection of a disjoint technique rather than a hierarchical technique.

Taking into consideration the above points, the disjoint clustering technique used in the FASTCLUS procedure of SAS (1985) was chosen to do the analysis. This procedure performs a disjoint cluster analysis based on the usual Euclidean distances computed from a given set of quantitative variables. The FASTCLUS procedure combines an effective method for finding initial clusters (or initial clusters can be given by the user) with a standard iterative algorithm for minimizing the sum of squared distances from the cluster means. FASTCLUS was directly inspired by Hartigan's leader algorithm (1975) and MacQueen's *k*-means algorithm (1967). A set of cluster seeds is first selected as a guess of the means of the clusters. Each observation is assigned to the nearest cluster seed to form temporary clusters. The cluster seeds are replaced by the means of the temporary clusters each time an observation is assigned (this is an option chosen for our application). After each pass through the data set, the observations are assigned to the nearest cluster seed until the changes in the cluster seeds become small or null (chosen to be null for our application). The final clusters are formed by assigning each observation to the nearest cluster seed.

Step II: Estimation of parameters

Two years of HMS data from 82–83 and 83–84 fiscal years were gathered to get estimates m_d and s_d for each diagnosis d. These estimates were the usual weighted estimates over the two year period. Each diagnosis is represented by two variables, m_d and s_d. The problem is now reduced to finding an appropriate partition of the diagnoses according to m_d and s_d. Three special groups of diagnoses judged as outliers were removed. These three special groups will form the first three rows of the imputation matrix (the columns are defined by the sex variable). A catch-all category was created in the last row of the imputation matrix for those diagnoses with, say, fewer than ten observations available over the two years of data and not included in the three special groups. The choice for the upper bound of ten observations was made arbitrarily. Cluster analysis can then be used to group the remaining diagnoses not included in the three special groups with at least ten observations available.

Step III: Determination of the number of clusters

The determination of the number of clusters was dictated by operational constraints since the imputation module of the program doing the imputation will accept a maximum number of rows not larger than 40. Since there are already three rows for special diagnoses and one row for diagnoses with fewer than ten observations, the maximum number of other rows that would not affect the program is then 36. A small empirical study calculating the R^2 coefficient for different numbers of clusters indicated that the R^2 coefficient was already above 98% for 36 clusters, suggesting that 36 clusters was acceptable. Note that even with 15 clusters, the R^2 could be made as high as 95%. The definition of the R^2 coefficient is given in section 3.4.

Step IV: FASTCLUS implementation

First, an initial partition of the observations into 36 groups was chosen (equivalent to choosing a set of 36 cluster seeds). Better results were obtained by selecting an initial set of cluster seeds than by letting FASTCLUS find initial cluster seeds. Note that different initial cluster seeds and different orders of the input data set will yield different results due to the fact that the method produces only locally optimal partitions. To select cluster seeds, diagnoses were divided into nine groups of roughly the same size according to m_d and four groups of roughly the same size according to s_d. This procedure produced 36 homogeneous groups of diagnoses of approximately the same size. The means of the two variables m_d and s_d in each group were taken as initial cluster seeds. Several other variations were tried and the procedure giving the largest R^2 was chosen.

Second, since m_d and s_d were based on very different numbers of observations for different diagnoses, it was judged preferable to perform a weighted cluster analysis, the weights being the number of observations available for each diagnosis. Note that, in this case, FASTCLUS would minimize the weighted within cluster sum of squares instead of an unweighted within-cluster sum of squares.

3.4 Relative Performance of the Proposed Method

One way to compare the current and proposed method for collapsing imputation classes is to use the R^2 coefficient pooled over all variables (in our case, it would be the mean and the standard deviation). The pooled R^2 coefficient is the proportion of the total variance explained by the between cluster pooled sum of squares (which should be as large as possible). Each pooled sum of squares is defined as $(SSQ_m + SSQ_s)/2$ where SSQ_m and SSQ_s are the sums of squares of the mean and the standard deviation respectively. The R^2 coefficients obtained from FASTCLUS were 0.993 for m_d and 0.929 for s_d for a pooled R^2 value of 0.986. The current classification of diagnoses into groups would yield an R^2 of 0.735 for m_d and 0.466 for s_d producing a pooled R^2 value of 0.705. Thus, in terms of R^2, results indicated that the groups of diagnoses formed using cluster analysis were much more homogeneous with respect to the variable being imputed than in the case where classes were formed by the earlier method.

4. CONCLUDING REMARKS

A methodology based on cluster analysis for collapsing the imputation classes of an imputation matrix defined by the cross-classification of several auxiliary variables was proposed. This methodology was applied to the imputation of age for the Hospital Morbidity System where diagnosis and sex were used as auxiliary variables.

It should be noted that in this specific application, only one variable, namely the diagnosis, was used to collapse the original imputation classes. The variable sex is, however, used later in the imputation scheme so that a recipient will be matched to a donor of the same sex. In a generalization of the proposed method, one may consider using the two variables, sex and diagnosis, in the collapsing process. For this purpose one might also impose some constraints that male and female cases of the same diagnosis belong to the same row in the final imputation matrix. Alternatively, one could produce two final imputation matrices, one for each sex. In either one of these alternatives, the number of initial imputation classes would clearly be much higher and hence the collapsing problem more complex. In this situation, it is more likely for many classes to have a small number of donors and therefore many of the imputation classes would have to be assigned to the catch all category. This, however, may not be desirable in practice. This problem can be simplified if one could make the assumption that, for most diagnoses, the male and female age distributions are similar to each other. There is some evidence based on significance tests that this is not an unreasonable assumption. In the HMS example considered, it was decided to group diagnoses based on estimates of μ_d and σ_d from the data pooled over sex.

It should also be noted that the choice of mean and standard deviation of age distribution to assign numerical scores to each imputation class was not investigated. Other choices might be percentiles or some other parameters of the age distribution. Clearly, the results of using cluster analysis for collapsing purpose would depend on the choice of the above scores.

Finally, generalization of the proposed method to the case where $k \geq 1$ variables need to be imputed and where $p \geq 2$ auxiliary variables are available follows in a straightforward manner from the simpler case considered in this paper.

ACKNOWLEDGEMENTS

This work was presented at the annual meeting of the "Association canadienne-française pour l'avancement des sciences" in May 1988. I thank Avi Singh for his helpful comments which have led to improvements in this paper. I would like to thank Cyril Nair of the Health Division and his staff for their support especially concerning the production of the computer files required to complete this work.

REFERENCES

ANDERBERG, M.R. (1973). *Cluster Analysis for Application.* New York: Academic Press.

BALL, G.H., and HALL, D.J. (1970). Some implications of interactive graphic computer systems for data analysis and statistics. *Technometrics,* 12, 17-31.

CORMACK, R.M. (1971) A review of classification. *Journal of the Royal Statistical Society,* Series A, 134, 321-367.

DREW, J.D., BÉLANGER, Y., and FOY, P. (1985). Stratification in the Canadian Labour Force Survey. *Survey Methodology,* 11, 95-110.

EVERITT, B.S. (1980). *Cluster Analysis.* Second Edition, London: Heineman Education Books Ltd.

JUDKINS, D.R., and SINGH, R.P. (1981). Using clustering algorithms to stratify primary sampling units. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 274-284.

HARTIGAN J.A. (1975). *Clustering Algorithms.* New York: John Wiley & Sons.

MacQUEEN J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings 5th Berkeley Symposium* 1, 281-297.

SAS INSTITUTE Inc. (1985). *SAS User's Guide: Statistics,* Version 5.

STATISTICS CANADA (1986). *Hospital Morbidity 1981-82, 1982-83.* Catalogue No. 82-206, Statistics Canada, Ottawa.

WORLD HEALTH ORGANIZATION (1977). *International Classification of Diseases.* 1975 Revision, Volume 1, Geneva.