International Journal of Management Science
Vol 14, No 2, November 2008

An Exploratory Study on Donor Location Strategies in Data Fusion

Jonathan S. Kim
School of Business, Hanyang University
17 Hangdang-Dong, Sungdong-Gu, Seoul, Korea

Sungbin Cho*
School of Business, Sogang University
1 Sinsu-Dong, Mapo-Gu, Seoul, Korea

(Received: August 23, 2007 / Revised: July 1. 2008 / Accepted: September 4. 2008)

ABSTRACT

This study explores several donor location strategies and discusses experiment results, which contributes to the saving of time and effort required in designing data fusion processes. In particular, three concepts are introduced. The Mahalanobis distance is applied to locate the nearest neighbors more effectively, which incorporates the covariance structure of attributes. The ideal point helps reduce the dimensionality problem that arises in conjoint-type experiments. The correspondence analysis is used to derive the coordinates from non-metric attributes. The Monte Carlo simulation results show that the proposed donor location strategies provide better fusion performance, compared to the currently-in-use methods.

Keywords: Data Fusion, Donor Location Strategy, Mahalanobis Distance, Correspondence Analysis

1. Introduction

Many leading corporations have adopted customer relationship management (CRM) and database marketing (DBM) as their major strategic tools to enhance both effectiveness and efficiencies of the performance of their business operation in the competitive business environment. Both the CRM and the DBM require current and accu-

* Corresponding author, E- mail: sungbincho@sogang.ac.kr

rate information about customers. A traditional, yet still popular method to build a competitive product positioning is to make an efficient use of survey data. Despite the time and effort involved in designing a questionnaire and collecting data through it, the data sets finally collected often look disappointing. In many cases, surveys end up with low return rates and insincere responses [4]. In parallel with the advancement of computers and the Internet, customer transaction data in today's market place are collected and utilized in a variety of ways through the framework of e-CRM. While consumers participate in Internet-based surveys, increasingly efficient and convenient form of data collection these days, it is true that they can be easily distracted in the cyber space, resulting in unsatisfactory responses. Similar problem also arises in conjoint data collection process, which involves a great amount of evaluation of a number of hypothetical product profiles.

Regardless of the source of customer information, either by surveys or by log file analysis, a common symptom is that both academic researchers and practitioners are mainly concerned with the ex-post analysis of the existing data. Data fusion plays an important role in estimating the value of missing information. Data fusion has been a popular approach in social sciences as an ex-post remedy for treating unsatisfactory database containing missing information. From an ex-ante perspective, data fusion process starts with the smaller design of a survey questionnaire and the structure of data sets to be analyzed. Reducing the amount of information to be collected helps respondents to maintain higher level of attention and to respond more faithfully to the questionnaire.

Despite the value of data fusion in estimating the missing information as well as in reducing the amount of information to be collected, however, not much research has been conducted in marketing and other business areas. Recently, Kim et al. [14] showed that the systematic selection of a set of common variables in data fusion yields better fusion performance. Building on their research, this study further examines the effectiveness of various donor location strategies in ex-ante data fusion processes. Data fusion can be used for planning data collection as well as creating an integrated data set in a more efficient way, which contributes in consequence to more accurate prediction of the unknown behavior of customers. This study explores various donor location strategies for intentional, ex-ante data fusion processes and evaluates their performance.

2. Data Fusion

2.1 Introduction to Data Fusion

Data fusion is a set of activities that researchers and practitioners must go through in an attempt to create an integrated data set upon combining a few ones [12, 13]. Since the term "data fusion" was created in Great Britain in 1988, it has been used primarily in marketing and media planning [11]. Data fusion is often synonymous to integrated targeting because statistical procedures such as matching, merging, and linking are dealt with in the data fusion process [1]. However, data fusion differs from traditional statistical treatments of missing data in the following sense. Unanswered questions or data filtering such as removing outliers result in missing values in the database construction. Traditional statistical methods have been used to impute such missing values whose patterns are, in general, random. In contrast, data fusion evolves from integrating data sets, in that some variables are not investigated in one data set while other variables are not investigated in another data set. Thereby, missing data happen systematically and in a massive way in data fusion processes.

Data fusion has been applied in three distinct fields. First, data fusion is applicable to market surveys. Two different survey results can be successfully fused together on the condition that some common questions exist between the two surveys. The second application area is media planning [18]. For example, a data set that contains customers' purchase behavior of a certain commodity can be fused together with another data set that contains television watching patterns of another cohort. Demographic similarities would play a key role for combining these two different data sets. Potential purchase behavior can be predicted for the group of customers who tend to watch a particular television program, and this can be linked to successful advertising and promotion campaigns. Finally, data fusion can be applied to direct marketing. For those customers who did not purchase a specific product or service, we can predict their preferences by fusing their profile with the existing buyers' profile and finally identify a group of potential customers for the specific product or service that they did not purchase.

2.2 Data Fusion Process

The following terminologies have been generally used in data fusion. "Donor"

refers to the one who gives his/her value to the one who has missing value on the same variable(s). "Recipient" refers to the counter part of the donor. "Common variables" means a set of variables that both donors and recipients have without an incomplete (missing) record. The common variables provide a basis to figure out proximity (such as similarity, dissimilarity, or distance) between the donors and the recipients. Variables other than the common variables are called "non-common variables" or "unique variables."

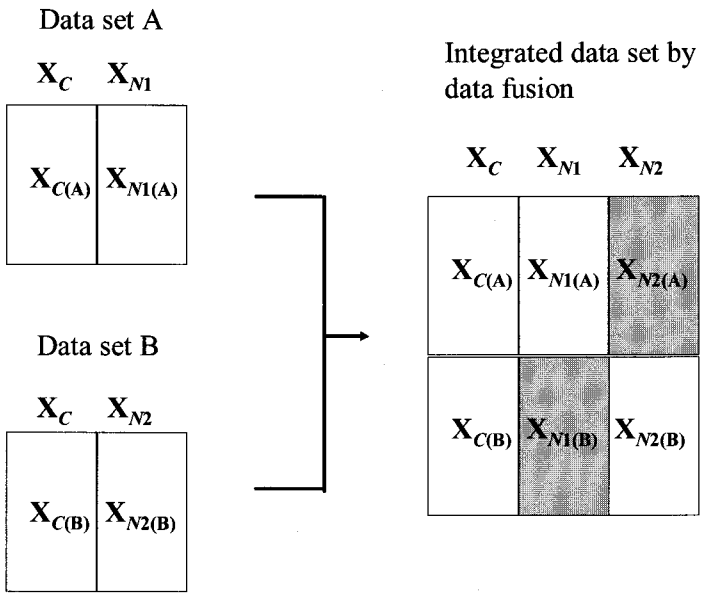

Figure 1. Fusion processes in an ex-ante data fusion process

As depicted in Figure 1, missing data are created in a systematic way in data fusion. Both data sets A and B share the vector of common variables $\mathbf{X}_C$. Data set A contains the vector of non-common variables $\mathbf{X}_{N1}$ whereas data set B contains the vector of non-common variables $\mathbf{X}_{N2}$. Depending on the proximity structure discovered from the vector of common variables across respondents, data fusion creates $\mathbf{X}_{N2}$ for data set A and $\mathbf{X}_{N1}$ for data set B, respectively. It is well known that conditional independence is the underlying assumption of data fusion. Given the common vector variables $\mathbf{X}_C$, the vector of non-common variables, $\mathbf{X}_{N1}$ and $\mathbf{X}_{N2}$, are conditionally independent of each other. Due to this assumption, substituting $\mathbf{X}_{N1}$ and $\mathbf{X}_{N2}$ for each party is reasonable on a statistical basis. Roger's study [17] pointed out that the assumption of

conditional independence is quite valid for variables with a normal density and that this assumption fits better for a planned data fusion that is based on the demographics and other general common variables of respondents. Jophcott and Bock [11] insisted that this assumption is quite realistic in planned media fusion considering conditional independence, compared to unplanned fusion studies.

2.3 Applying Data Fusion to Conjoint Experiment

Conjoint analysis has been widely used for measuring consumer preferences over the last few decades [6, 7, 10] since Green and Rao [5] introduced the methodology to marketing and decision-making problems. A consumer's preference for products or services normally encompasses multi-attributes where each attribute includes multi-attribute levels. A consumer's utility or subjective preference is evaluated for each level of an attribute and the sum represents the overall product utility. The product with higher utility is presumed to be a better choice for the consumer. The strengths of conjoint analysis include the general assumptions about the relationships between independent and dependent variables and the flexibility of accommodating metric or non-metric dependent variables [8].

On the other hand, conjoint analysis is notorious for requiring a large number of comparisons of the hypothetical product profiles. The number of hypothetical product profiles increases exponentially as the size of attributes and attribute levels grows. Designing a commercial product usually includes more than ten attributes with several attribute levels for each attribute. For example, if eleven attributes are considered with four attribute levels for each attribute, the number of products to evaluate is 4^{11}. While a fractional factorial design is used in most conjoint experiments as a way to reduce the number of profiles to evaluate, Kim and Hamano [15] has considered the value of data fusion to reduce the number of product profiles to evaluate in conjoint experiment.

3. Research Objectives and Method

3.1 Research Objectives

The main objective of this study is to explore several donor location strategies

and compare their performances in terms of forecasting accuracy of data fusion. We consider the Mahalanobis distance in the context of data fusion as a way of defining the dissimilarity among respondents. In the analysis, correspondence analysis is employed to locate the donor for a recipient. By increasing the amount of missing values, we evaluate, through a Monte Carlo simulation, the degree to which data fusion recovers the original values using two real world data sets.

3.2 Data

Two data sets with respect to utility measures are used: one data set includes part-worth from a conjoint experiment and the other contains satisfaction level from a general survey. The former includes the preference for a credit card service from a sample of 480 respondents, including 12 attributes and 35 attribute levels. The part-worth is measured on an attribute level as a numeric value ranging from zero to one. The attributes of the credit card service data include annual fee (with 6 attribute levels), cash refund (3 levels), message delivery (3 levels), purchase item insurance (2 levels), air travel insurance (3 levels), rental car insurance (2 levels), luggage insurance (3 levels), airport lounge/club (3 levels), credit card coverage (4 levels), emergency vehicle (2 levels), limousine service (2 levels), and 24 hour customer assistance (2 levels).

The other data set contains the customer satisfaction level for Internet services from 500 respondents. The level of satisfaction is measured on a Likert scale from one to five. The data consist of 11 attributes only (without attribute levels), including dimensions such as: security of Internet shopping and auction, personal data leakage, legal regulations, damage by hacking, time spent, access and line congestion, provider's response, service charge, information search, easiness of operation, and method of use.

3.3 Experimental Design

Apparently, selecting the common variables is an important subject in data fusion process, since fusion performance is affected by the proximity between the donor and the recipient as well as by the composition of common variables. In this research, the optimal set of common variables is constructed by the attributes that have greater variances, as suggested by Kim *et al.* [14]. Six attributes are selected out of 12 for the credit card service data set, while five attributes are selected out of 11 for the Internet

service data set, respectively. Respondents are divided in half for each data set. Non-common attributes (with missing data) are selected at random and their values are then deleted. To simulate planned data fusion, the missing attributes differ in the two divided groups. In other words, the two groups contain different sets of unique attribute variables. Proximity is computed by the common variable vectors, either by calculating similarity such as correlation coefficient, or by dissimilarity such as distance.

For the credit card service data, the following five strategies are adopted for locating donors. Note that the six common attributes selected include 19 attribute levels in total.

- Strategy 1: From the correlation coefficient matrix for the 19 common attribute level variables, the subject with the highest correlation coefficient becomes the donor.
- Strategy 2: The 19 common attribute level variables are used. The donor is found by the least distance method where the distance is defined by the Euclidean distance, using Eq. (1).

$$d_{E(ij)} = \sqrt{(\mathbf{X}_{C(i)} - \mathbf{X}_{C(j)})(\mathbf{X}_{C(i)} - \mathbf{X}_{C(j)})^T} \tag{1}$$

  where $d_{E(ij)}$ denotes the Euclidean distance of respondents i and j.
- Strategy 3: The 19 common attribute level variables are used, and the Mahalanobis distance is employed to locate the closest respondent, using Eq. (2).

$$d_{M(ij)} = \sqrt{(\mathbf{X}_{C(i)} - \mathbf{X}_{C(j)})\Gamma^{-1}(\mathbf{X}_{C(i)} - \mathbf{X}_{C(j)})^T} \tag{2}$$

  where $d_{M(ij)}$ denotes the Mahalanobis distance of respondents i and j, and Γ refers to the covariance matrix of the common variables.
- Strategy 4: As is well known, searching an association rule or statistical dependence in a high dimensional space often ends up with a poor result. To reduce this chance, we introduce "ideal point" variable, in that the attribute level with the maximum value represents the corresponding attribute in a categorical way. That is, the 19 numeric common variables are transformed into six categorical common variables. To obtain the coordinates from these variables, we apply the chi-square metric method from multiple correspondence analyses [2, 3, 9], a special multidimensional scal-

ing technique [16]. Given the coordinates in five dimensions (one less dimension resulting from the correspondence analysis), the Euclidean distance measure, as in Eq. (1), is applied.

- Strategy 5: Like Strategy 4, given the coordinates in five dimensions resulting from the correspondence analysis upon the ideal point variables, and the donor is found by the Mahalanobis distance measure, Eq. (2).

The accuracy of data fusion strategies is evaluated by increasing the number of missing attributes; deleting all attribute levels of one, two, or three attributes. Therefore, a 5 (donor location strategy) × 3 (missing attributes) full factorial design is employed for the credit card service data set.

For the Internet service data, the dimensionality problem does not matter because only five common attribute variables are analyzed. The correspondence analysis is applied to obtain the coordinates of respondents from the five non-metric common variables, resulting in four metric common variables. In this case, the following three donor location strategies are considered.

- Strategy 1: From the correlation coefficient matrix for the derived metric common variables, the respondent with the highest correlation coefficient becomes the donor.
- Strategy 2: The Euclidean distance measure, using Eq. (1), is applied to find the donor.
- Strategy 3: The Mahalanobis distance measure, using Eq. (2), is applied to find the donor.

Like the credit card service data set, the level of missing attributes increases from one to three. These non-common attributes are selected at random and then deleted. Thus, a 3 (donor location strategy) × 3 (missing attributes) factorial design is made for the Internet service data set. For each combination of the factor levels, the Monte Carlo simulations are carried out 20 times for both data sets.

4. Results

The performance of the donor location strategies is evaluated using two statistical

measures-the root mean square error (root MSE) of the original data and the fused data for the credit card service data set, and the misclassification ratio for the Internet service data set, respectively. Table 1 summarizes the results of the analysis of variance for the credit card service data set. "Strategy" factor and "Missing" factor are both significant with p-values of less than 0.0001. According to the Scheffe grouping as a posterior test, we can be reasonably confident that Strategy 5 (by Mahalanobis distance over the ideal points) yields a better result than Strategy 1 (by correlation coefficient), Strategy 2 (by Euclidean distance), and Strategy 3 (by Mahalanobis distance). Also, Strategy 4 (by Euclidean distance over the ideal points) outperforms Strategy 3 (by Mahalanobis distance). On the other hand, as anticipated, the performance of data fusion becomes worse as the number of missing values increases, which is intuitively understandable.

Table 1. Test results for the credit card service data

ANOVA Test					
Source	d.f.	Sum of squares	Mean square	F statistic	p-value
Model	6	0.00139933	0.00023322	1141.69	< 0.0001
Error	293	0.00005985	0.00000020	-	-
Total	299	0.00145919	-	-	-
Strategy	4	0.00001308	0.00000327	16.01	< 0.0001
Missing	2	0.00138625	0.00069313	3393.03	< 0.0001
A Posteriori Test					
	Factor	Scheffe grouping		Mean	Standard deviation
Strategy	Strategy 3	A		0.00836	0.002317
	Strategy 2	A	B	0.00813	0.002253
	Strategy 1		B	0.00802	0.002246
	Strategy 4	C	B	0.00790	0.002098
	Strategy 5	C		0.00774	0.002149
Missing	3	A		0.01050	0.000334
	2	B		0.00833	0.000532
	1	C		0.00526	0.000585

For the Internet service data set, the two experimental factors, "Strategy" and "Missing" are statistically significant (see Table 2). The Scheffe grouping indicates that Strategy 1 (by correlation coefficient) and Strategy 3 (by Mahalanobis distance)

lead to lower misclassification ratio than Strategy 2 (by Euclidean distance). The performance of data fusion becomes worse for increasing missing values, which is consistent with the case of the credit card service data.

Table 2. Test results for the Internet service data

ANOVA Test					
Source	d.f.	Sum of squares	Mean square	F statistic	p-value
Model	4	0.41510610	0.10377652	28608.5	< 0.0001
Error	175	0.00063481	0.00000363	-	-
Total	179	0.41574090	-	-	-
Strategy	2	0.00010782	0.00005391	14.86	< 0.0001
Missing	2	0.41499827	0.20749914	57202.2	< 0.0001
A Posteriori Test					
	Factor	Scheffe grouping	Mean	Standard deviation	
Strategy	Strategy 2	A	0.11878	0.048895	
	Strategy 3	B	0.11741	0.048249	
	Strategy 1	B	0.11696	0.048227	
Missing	3	A	0.17676	0.001497	
	2	B	0.11725	0.002199	
	1	C	0.05915	0.002347	

5. Discussion

This study considered various donor location strategies in data fusion and evaluated their performance in terms of forecasting accuracy of the fused data compared to the original data. Although the test results of this study may not be sufficient to make a confirmatory assertion with regard to the value of donor location strategies, the strategies proposed in this exploratory study can shed a light such that a systematic ex-ante data fusion process can save time and cost by reducing the amount of data to collect.

From an exploratory point of view, this study contributes to the field of data fusion in the following sense. First, we considered the Mahalanobis distance for locating donors in data fusion. It is natural to assume that the attributes are correlated with one another in a large data set. Thereby, a better way to measure proximity

among respondents can be achieved by reflecting the covariance structure of the attributes into the model. The second contribution is that this study suggests a way of reducing the number of attribute levels by transforming them into the ideal point variables. The test results of the credit card service data show that this can be a good approach to cope with the dimensionality problem that might arise in conjoint experiments as well as in data fusion. Third, various donor location strategies were applied to the two kinds of utility data sets-the conjoint part-worths and the general satisfaction measures. To measure the proximity for the non-metric data, the correspondence analysis was used for assigning subjects on the metric dimensional space. Note that the common variables used in traditional data fusion approaches have been typically demographic data. In other words, we explored the possibility of expanding the scope of common variables into the judgmental data such as the conjoint part-worths or customer satisfaction measures.

The limitation of this study can be pointed out as the generalizability of the findings. Although we tried to assure the minimal level of external validity by applying into the two data sets, more data sets in various fields should be examined in the future studies to validate the current findings. On the other hand, it would be worthwhile to investigate whether an interaction effect exists between the donor location strategies and the common variable selection strategies. Although the attribute variables with greater variance were selected as a set of common variables in the current study, demographics or psychographic variables might be considered as a viable set of common variables in the future research. Also, the impact of the number of common variables selected on the data fusion performance should be investigated, along with the donor location strategies.

References

[1] Baker, K., P. Harris, and J. O'Brien, "Data fusion: An appraisal and experimental evaluation," *Journal of the Marketing Research Society* 39, 1 (1997), 225-271.

[2] Benzecri, J. P., *Correspondence analysis handbook*, Marcel Dekker Inc 1992.

[3] Carroll, D. J., P. E. Green, and C. M. Schaffer, "Interpoint distance comparisons in correspondence analysis," *Journal of Marketing Research* 23 (1986), 271-280.

[4] Craig, S. C. and J. M. McCann, "Item nonresponse in mail surveys: Extent and correlates," *Journal of Marketing Research* 15, 2 (1978), 285-289.

[5] Green, P. E. and V. R. Rao, "Conjoint measurement for quantifying judgmental data," *Journal of Marketing Research* 8, 3 (1971), 355-363.

[6] Green, P. E. and V. Srinivasan, "Conjoint analysis in consumer research: Issues and outlook," *Journal of Consumer Research* 5 (1978), 103-123.

[7] Green, P. E. and V. Srinivasan, "Conjoint analysis in marketing: New developments with implications for research and practice," *Journal of Marketing* 54 (1990), 3-19.

[8] Hair, J. F. Jr., R. E. Anderson, R. L. Tatham, and W. C. Black, *Multivariate data analysis*, Prentice Hall, New Jersey (NJ) (1998), 387-437

[9] Hoffman, D. L. and G. R. Franke, "Correspondence analysis: Graphical representation of categorical data in marketing research," *Journal of Marketing Research* 23, 3 (1986), 213-227.

[10] Hofstede, F. T., Y. Kim, and M. Wedel, "Bayesian prediction in hybrid conjoint analysis," *Journal of Marketing Research* 39, 2 (2002), 253-261.

[11] Jephcott, J. and T. Bock, "The application and validation of data fusion," *Journal of the Marketing Research Society* 40, 3 (1998), 185-205.

[12] Kamakura, W. A. and M. Wedel, "Statistical data fusion for cross-tabulation," *Journal of Marketing Research* 34, 4 (1997), 485-498.

[13] Kamakura, W. A. and M. Wedel, "Factor analysis and missing data," *Journal of Marketing Research* 37, 4 (2000), 490-498.

[14] Kim, J. S., S. Baek, and S. Cho, "A preliminary study on common variable selection strategy in data fusion," *Advances in Consumer Research* 31 (2004), 716-720.

[15] Kim, J. S. and M. Hamano, "A preliminary study of data fusion approach in conjoint analysis," World Marketing Congress VII: Proceedings of the Seventh Bi-Annual World Marketing Congress, Melbourne, Australia, 10 (1995), 95-101.

[16] Kruskal, J. B. and M. Wish, *Multidimensional scaling*, Sage Publications, London 1986.

[17] Roger, W. L., "An evaluation of statistical matching," *Journal of Business and Economic Statistics* 2, 1 (1984), 91-105.

[18] Wiegand, J., "Combining different media surveys: The German partnership model and fusion experiments," *Journal of the Marketing Research Society* 28, 2 (1986), 189-208.