Original Article

Improving your sales with data fusion

Received (in revised form): 23rd December 2008

Pascal van Hattum

studied Business Mathematics and Computer Science at the Free University Amsterdam. After his studies (2000) he started as a statistical consultant at The SmartAgent Company, where he is currently responsible for all the statistical processes and research and development. In 2004, he started a PhD project in the Faculty of Social and Behavioural Sciences of Utrecht University under the supervision of Professor Herbert Hoijtink. In 2009, he hopes to finalise his thesis 'Market Segmentation Using Bayesian Model Based Clustering'. The knowledge behind his PhD project is successfully used in the day-to-day business of The SmartAgent Company.

Herbert Hoijtink

is Professor of Applied Bayesian Statistics at the Department of Methodology and Statistics in the Faculty of Social and Behavioural Sciences of Utrecht University. In addition to Bayesian Statistics, his areas of expertise are, among others, model-based clustering for large data sets, regression, (repeated) measures, analysis of variance, multiple hypotheses testing and structural equation modelling.

ABSTRACT This paper shows how a European mail order company uses data fusion in order to improve sales. To select the best data fusion algorithm, two traditional data fusion methods, that is, polytomeous logistic regression and nearest neighbour algorithms, are compared with two model-based clustering approaches. Finally, it is shown how internal and external validation criteria are used in order to evaluate the results of the data fusion algorithms.

Journal of Database Marketing & Customer Strategy Management (2009) 16, 7–14.

doi:10.1057/dbm.2009.4

Keywords: data fusion; differentiated marketing; model-based clustering; internal validation; external validation

INTRODUCTION

In this paper, the following problem is addressed: a European mail order company, which specialises in gardening products, wants to send differentiated catalogues to all the customers in the database. In order to do this, information about customer gardening preferences and interests from an external database are fused to the customer database. Using the fused information, differentiated catalogues can be sent to all the customers.

The problem described above can best be illustrated using the schematic representation

in Figure 1.¹ In this representation, data set **A** is the customer database and contains knowledge and information (represented by J items) from all customers. External data set **B** contains knowledge and information (represented by $J+1$ items) from a group of customers who are not present in data set **A**. The knowledge and information represented by the first J items is available for each customer in each data set. For the group of customers in external data set **B**, however, there is some additional knowledge and information, that is, item $J+1$. The goal of this paper is to fuse the

Correspondence:
Pascal van Hattum
Department of Methodology
and Statistics, Faculty of
Social Sciences, Utrecht
University, PO Box 80140,
Utrecht NL-3508 TC,
the Netherlands
E-mail: p.vanhattum@uu.nl

Figure 1: Schematic representation of data fusion in marketing (derived and adjusted from Van der Putten *et al.*).

extra knowledge and information in external data set **B**, that is, item $J+1$, to data set **A**. As a result of this data fusion, the knowledge and information about item $J+1$ becomes 'known' for all customers in the database, data set **A**.

The structure of this paper is as follows. The next section describes how data fusion can be used in the context of the European mail order company. In addition, the marketing goals and the data sets used are described. The subsequent section shows how data fusion is evaluated using internal and external validation criteria. This paper concludes with a discussion in the final section.

APPLICATION

Improving sales

A European mail order company, specialised in all kinds of gardening products, such as flowers, bulbs, plants, and so on, wants to increase the number of buying customers.^{2,3} In order to do this, they decide to develop a direct marketing strategy. From an external motivational research study, the mail order company knows that each individual has a different attitude towards gardening and gardening products (www.tuinbeleving.nl).

From the motivational study, it is known that there are actually four groups (or clusters) of customers who have more or less the same attitude towards gardening and

gardening products. Short descriptions (see www.tuinbeleving.nl for detailed descriptions) of these four motivational clusters are as follows:

- *cluster 1:* gardens in this cluster are different from other gardens. They are surprising, wild, romantic and stylish, gardens meant for relaxation and unwinding. For customers in this cluster, gardening brings that relaxation;
- *cluster 2:* gardens in this cluster are more often large patios, easy to maintain and cluttered. Gardens are outdoor spaces to hang out with family and friends. Customers in this cluster think of gardening as strenuous, rather than a relaxing activity;
- *cluster 3:* in this cluster, the true gardener can be found. Gardens in this cluster are nice, neat, full of atmosphere and fit in with the rest of the neighbourhood. Gardening is relaxing, a passion and the main hobby;
- *cluster 4:* gardens in this cluster are practical and easy to maintain. Customers in this cluster do not feel like gardening and cannot find the time for gardening.

These motivational clusters provide a basis for developing a company's vision and/or a company's marketing directions on the strategic, tactical and operational levels, aligning the total marketing mix around the consumers needs in the domain of gardening. Table 1 displays the frequencies of the resulting motivational clustering. Furthermore, each individual deals with, handles and perceives gardening catalogues in a different way. The mail order company assumes that giving customers (some sort of) a tailor-made offer will eventually increase the number of buying customers.

Using the descriptions of the four motivational clusters, for each of the four clusters a separate catalogue can be made by a specialised communication agency. The content of the catalogues, that is,

Table 1: Frequency of respondents in motivational clusters concerning gardening

<i>Cluster</i>	<i>Frequency respondents</i>	<i>Percentage respondents</i>
Cluster 1	246	21.6 (23.0)
Cluster 2	228	20.0 (21.3)
Cluster 3	343	30.1 (32.0)
Cluster 4	254	22.3 (23.7)
No cluster	70	6.1
Total	1141	100.0

Between brackets are the percentages based on the total number of respondents classified to one of the four motivational clusters.

the gardening products offered to the customers, is the same for each catalogue. Only the layout (colours and pictures used on the front and the back pages) and the tone of voice of the catalogue's introduction are different for the cluster-specific catalogues.

Because the mail order company wants to send differentiated catalogues to their customers, data fusion is used. Using the common items in both the supplier's customer database and the external motivational research study, data fusion methods are used to fuse the motivational cluster to the supplier's customer database.

Description of the data sets used

The available data sets are the customer database of the company, and the data set with the external motivational research study. Or, translated to Figure 1, data set **A** and data set **B**, respectively. The content of data set **A** is data set $\mathbf{X}^A$. This data set $\mathbf{X}^A$ contains $J=7$ items, which are house ownership, number of vehicles, education, socio-demographic typology, household stage, prosperity and spending behaviour, from all $N=66\,549$ customers.

The content of data set **B** is data set $\mathbf{X}^B$ and data set **Y**. Data set $\mathbf{X}^B$ contains the same $J=7$ items as data set $\mathbf{X}^A$. The content of data set **Y** is one item, that is, the motivational clusters. In total, $M=1141$ respondents have participated in the motivational study, and for them the

motivational clusters are known. Note that these 1141 customers are not a fraction of the 66 549 customers from data set **A**.

As described in the first section and illustrated in Figure 1, the goal of the data fusion process is to fuse the information in data set **Y** to data set **A** using the common items in $\mathbf{X}^A$. Or, in the context of this application, using data fusion methods, all the 66 549 customers are classified into one of the four motivational clusters using the seven common items. How well or badly this data fusion process is carried out is described in the following sub-sections.

VALIDATION

Internal validation

Van Hattum and Hoijtink⁴ compare four data fusion methods: polytomeous logistic regression, a nearest neighbour algorithm, a fusion value-specific probabilities method and a model-based clustering approach. In the nearest neighbour algorithm, the missing motivational clusters in data set **A** are duplicated from data set **B** using cases with similar values on the explanatory items. In polytomeous logistic regression, the relationship between motivational cluster and the explanatory items as determined in data set **B** is used to fuse information to data set **A**. The fusion value-specific probabilities model is based on latent cluster analysis, where the role of the latent clusters is taken by the fusion value, in this case the motivational cluster, and the explanatory variables are the items. The model fitted in data set **B** is used to predict the motivational clusters in data set **A**. The assumption in the model-based clustering approach is that the data are generated from a mixture of fusion value-specific probabilities models. As a result of the model-based clustering approach, there will be a fusion value-specific probabilities model for each latent cluster found. Translated to the application at hand, the number of latent clusters found is nine;

for each of the nine latent clusters a fusion value-specific probabilities model is estimated. The interested reader is referred to the study by Van Hattum and Hoijtink⁴ for a full description of the four data fusion approaches.

In order to select the best data fusion method, the statistics Total Correct Classification Rate (TCCR) and model lift are calculated. The TCCR is the percentage of respondents who are classified to the right motivational cluster. Furthermore, a percentage of correct classifications based on a random chance model can be obtained. Or, in other words, the percentage of correct classifications that can be expected when the motivational clusters are randomly assigned to the respondents. This percentage is called $TCCR_{chance}$.

The statistic model lift is calculated as $(TCCR/TCCR_{chance}) \times 100$ per cent, and can be interpreted as the percentage of more correct classifications than would be obtained by chance. All statistics can be calculated for the overall model and for each motivational cluster separately.

The statistics described above are calculated for each data fusion method, applied to one training data set

(approximately 50 per cent of the cases) and two test data sets (each approximately 25 per cent of the cases). Models fitted on a data set tend to predict much better for that data set than for a new data set sampled from the same population. As we want to fuse values from data set **B** to data set **A**, we will evaluate the predictive performance of the four fusion methods by fitting them on the training data set, and using the resulting models to make predictions in the two test data sets. As will be illustrated below, this prevents model overfitting. The interested reader is referred to the study by Verstraeten⁵ for a further discussion of model overfitting.

Frequencies, TCCRs and model lifts can be used to determine which data fusion method performs the best. There must be a proper comparison between the actual and the predicted frequencies of the motivational clusters. These frequencies are displayed in Table 2. Looking at this table, it is not clear which data fusion method to choose. When looking at, for example, the data fusion method 'logistic regression', it can be seen that the predicted frequency for motivational cluster 1 can be compared with the actual frequency. The difference

Table 2: Classification percentages after applying data fusion methods for domain gardening

Method		Data set	#Records	Cluster 1 (%)	Cluster 2 (%)	Cluster 3 (%)	Cluster 4 (%)
	Actual	train	539	23.0	18.2	34.3	24.5
		test1	257	21.4	22.2	30.4	26.1
		test2	275	24.4	26.5	29.1	20.0
Nearest neighbour	Predicted	train	539	—	—	—	—
		test1	257	25.7	17.1	36.2	21.0
		test2	275	25.5	18.5	33.5	22.5
Logistic regression	Predicted	train	539	24.7	10.9	38.6	25.8
		test1	257	23.3	11.3	38.9	26.5
		test2	275	25.8	12.7	39.3	22.2
Fusion value-specific approach	Predicted	train	539	33.2	13.7	31.7	21.3
		test1	257	31.1	16.3	29.6	23.0
		test2	275	32.0	15.3	34.5	18.2
Model-based clustering approach	Predicted	train	539	25.4	13.2	39.3	22.1
		test1	257	23.3	16.0	34.2	16.5
		test2	275	21.8	17.8	35.3	25.1

Table 3: Total correct classification rates after applying data fusion methods for domain gardening

<i>Method</i>		<i>Data set</i>	<i># Records</i>	<i>Cluster 1 (%)</i>	<i>Cluster 2 (%)</i>	<i>Cluster 3 (%)</i>	<i>Cluster 4 (%)</i>	<i>Total (%)</i>	<i>Chance (%)</i>
Nearest neighbour	Predicted	train	539	—	—	—	—	—	26.4
		test1	257	31.8	20.5	38.7	37.0	33.5	25.5
		test2	275	38.6	35.3	37.0	22.6	33.8	25.4
Logistic regression	Predicted	train	539	46.6	50.8	55.8	51.1	51.8	26.4
		test1	257	43.3	48.3	43.0	52.9	46.3	25.5
		test2	275	42.3	57.1	40.7	37.7	42.5	25.4
Fusion value-specific approach	Predicted	train	539	39.1	45.9	55.0	47.0	46.8	26.4
		test1	257	41.2	45.2	44.7	55.9	46.3	25.5
		test2	275	44.3	52.4	46.3	36.0	44.7	25.4
Model-based clustering approach	Predicted	train	539	48.9	47.7	58.5	52.9	54.7	26.4
		test1	257	40.0	31.7	42.0	44.1	40.5	25.5
		test2	275	43.3	42.9	39.2	31.9	38.9	25.4

between the predicted frequency for motivational cluster 2 and the actual frequency for this motivational cluster is, however, quite large. Similar observations can be made for other data fusion methods.

As can be seen from the TCCRs (displayed in Table 3) and the model lifts (displayed in Table 4), the data fusion method ‘nearest neighbour’ performs the worst of all. Both the TCCRs and the model lifts are the lowest compared to the other methods. The method ‘model-based clustering approach (with nine latent clusters)’ has the highest statistics applied to the train data set, but, due to model overfitting, these statistics drop when applied to the test data sets. The other data fusion methods also suffer from model overfitting, but not as dramatically as the model-based clustering approach. From the two tables, it can be seen that the statistics for the data fusion methods ‘logistic regression’ and ‘fusion value-specific probabilities approach’ are among the highest. Both methods are good models; however, the statistics for the latter model are more consistent on the train data set and the two test data sets.

As in Van Hattum and Hoijsink’s study,⁴ the fusion value-specific probabilities approach turns out to be the best-performing data fusion method.

Consequently, this method is used to fuse the motivational clusters to the company’s customer database. As a result of this data fusion, the motivational clusters become known for all the customers in the database. This is the starting point for differentiated marketing strategies, as described in the next sub-section.

External validation

In the case of the European mail order company, the initial goal was to increase the number of buying customers. From past experiences, the mail order company knows that 3.58 per cent of the customers who received a catalogue bought something from this catalogue within four weeks of receipt. Using differentiated catalogues, the goal of the mail order company is to increase this number of buying customers.

As a result of the data fusion, the total customer database, with 66 549 customers, is classified. The columns ‘Frequency customers’ and ‘Percentage customers’ in Table 5 show the resulting motivational cluster frequencies of the fused data set $\hat{Y}$. For 6056 (=9.1 per cent) customers, there are no or insufficient common items available to classify into one of the four motivational clusters. From the percentages between brackets in Tables 1 and 5, it is clear that the mail order company has more

Table 4: Model lifts after applying data fusion methods for domain gardening

<i>Method</i>		<i>Data set</i>	<i>#Records</i>	<i>Cluster 1</i>	<i>Cluster 2</i>	<i>Cluster 3</i>	<i>Cluster 4</i>	<i>total</i>
Nearest neighbour	Predicted	train	539	—	—	—	—	—
		test1	257	149	92	128	142	131
		test2	275	158	133	127	113	133
Logistic regression	Predicted	train	539	203	280	162	209	196
		test1	257	202	218	142	203	182
		test2	275	173	215	140	189	167
Fusion value-specific approach	Predicted	train	539	170	253	160	192	177
		test1	257	193	204	147	215	182
		test2	275	182	197	159	180	176
Model-based clustering approach	Predicted	train	539	213	318	170	216	208
		test1	257	187	143	139	169	159
		test2	275	178	161	135	159	153

customers classified into motivational cluster 1 and 3, compared with the external motivational research study. This is completely consistent with the description of these two motivational clusters; both motivational clusters contain, in general, more true and passionate gardeners.

Using the descriptions of the four motivational clusters, for each cluster a separate catalogue can be made. In the first step of the differentiated marketing strategy, however, the mail order company wants to concentrate on just two of the four motivational clusters: motivational clusters 1 and 3. Thus, for only these two motivational clusters, cluster-specific catalogues are made by a specialised communication agency. The content of the catalogues, that is, the gardening products offered to the customers, is the same. Only the layout (colours and pictures used on the front page and the backpage) and the tone of voice of the catalogue's introduction are different for the cluster-specific catalogues. The focus of the catalogue for motivational cluster 1 is on getting inspired by the catalogue, self-creation of gardens, and exotic and adventurous gardens. The focus of the catalogue for motivational cluster 3 is on traditional and hobby gardening, and on the amount of information that is being given about gardening products and services.

Table 5: Frequency clusters in application gardening

<i>Cluster</i>	<i>Frequency customers</i>	<i>Percentage customers</i>
Cluster 1	16938	25.5 (28.0)
Cluster 2	8409	12.6 (13.9)
Cluster 3	24076	36.2 (39.8)
Cluster 4	11070	16.6 (18.3)
No cluster	6056	9.1
Total	66549	100.0

Between brackets are the percentages based on the total number of customers classified to one of the four motivational clusters.

Furthermore, in order to validate the two cluster-specific catalogues, the mail order company decides to compare the results with the standard catalogue. This is achieved using a randomised experiment among a sample from the customers who are classified into either motivational cluster 1 or 3. The set-up for the randomised experiment is as follows (see also Table 6):

- to 6250 customers, who are classified into motivational cluster 1, standard catalogues are sent;
- to 6250 customers, who are classified into motivational cluster 1, cluster 1-specific catalogues are sent;
- to 6250 customers, who are classified into motivational cluster 3, standard catalogues are sent;

Table 6: Number of catalogues sent

		<i>Catalogue sent</i>	
		<i>Standard</i>	<i>Cluster-specific</i>
Predicted	Cluster 1	6250	6250
	Cluster 3	6250	6250

— to 6250 customers, who are classified into motivational cluster 3, cluster 3-specific catalogues are sent.

From the 25 000 customers, who are in the randomised experiment, the percentages of customers, as displayed in Table 7, bought something from the (un)differentiated catalogues. From this table, it is clear that the differentiated catalogue approach rendered more buying customers than the standard catalogues, not only compared with the customers who received undifferentiated catalogues, but also with the 3.58 per cent buying customers from past experiences. It can be concluded that the company's goal with the differentiated marketing strategy is attained.

Also from Table 7, it is interesting to see the difference between the two motivational clusters. From the table, it is clear that the effect of the differentiated catalogue is larger for motivational cluster 1 than for motivational cluster 3. This difference can be explained by the following three arguments. First, from past experiences with the motivational clusters, it is known that customers classified into motivational cluster 1 are, in general, more sensitive to (perceived) tailor-made offerings than customers classified into motivational cluster 3. Second, from past experiences with the motivational clusters, it is known that customers classified to motivational cluster 1 buy, in general, more products from mail order companies. And finally, the specialised communication agency thinks that the layout of the standard catalogue and the tone of voice of the standard catalogue's introduction are more likely to attract

Table 7: Percentage buying customers

		<i>Catalogue sent</i>	
		<i>Standard (%)</i>	<i>Cluster-specific (%)</i>
Predicted	Cluster 1	3.70	4.46
	Cluster 3	3.65	3.83

customers classified into motivational cluster 3.

For both motivational clusters 1 and 3, however, the percentage of buying customers is higher using a differentiated marketing approach than using an undifferentiated or standard approach. With the differentiated catalogue, the European mail order company is able to acquire more buying customers, and, consequently, increase its turnover.

The researcher must keep in mind that it is impossible to conclude that the increase in number of buying customers can be fully attributed to the differentiated catalogues. When sending the catalogues, it was impossible to control for all kinds of side effects that may be associated with customer buying behaviour. For this application, however, the goal of increasing the number of buying customers is attained. Furthermore, by using the results of an external motivational research study, instead of conducting their own research study, the mail order company saved dollars on marketing research activities.

DISCUSSION

In this paper, the customer database of a mail order company was fused to a motivational research study about gardening. In order to fuse the data sets, different traditional and new data fusion methods were used to fuse the data sets.

In the internal validation step, the different data fusion methods were compared and the fusion value-specific probabilities approach was found to be the 'best' method. The TCCR for the overall

model was around 45–46 per cent, whereas the TCCR with a random chance mode was around 25 per cent. The resulting model lift was around 180 per cent, which means that the fusion value-specific probabilities approach provided around 80 per cent more correct matches than would be obtained by chance. The conclusion that the fusion value-specific probabilities approach performed the ‘best’ was not only drawn in this paper, but also in past research.⁴ This makes this data fusion method a method with stable results.

As a result of the internal validation step, the motivational clusters were estimated for all the customers in the database. This was the starting point for differentiated marketing strategies, or, in the case of the mail order company, differentiated catalogues were made by a specialised company.

Using a randomised experiment, different marketing goals were tested. In the case of the mail order company, more customers bought something from the catalogue when receiving the right (differentiated) catalogues.

Given the large number of customers involved, the increase in buying behaviour gave the company a tremendous amount of extra sales. Furthermore, by using an external domain study, a lot of dollars were saved on marketing research costs. In all cases, the data fusion project was profitable, and, consequently, was successful.

REFERENCES

- 1 Van der Putten, P., Kok, J. N. and Gupta, A. (2002) Data Fusion through Statistical Matching. MIT Sloan School of Management. Paper 185.
- 2 Lattin, J. M. and Bucklin, R. E. (1989) Reference effects of price and promotion on brand choice behavior. *Journal of Marketing Research* 26(3): 299–310.
- 3 Feinberg, F. F., Krishna, A. and Zhang, J. Z. (2002) Do we care what others get? A behaviorist approach to targeted promotions. *Journal of Marketing Research* 39(3): 277–291.
- 4 Van Hattum, P. and Hoijtink, H. (2009) The proof of the pudding is in the eating. Data fusion: An application in marketing. *Journal of Database Marketing & Customer Strategy Management* 15(4): 267–284.
- 5 Verstraeten, G. (2005) Issues in predictive modelling of individual customer behavior: Applications in targeted marketing and consumer credit scoring. PhD thesis, Marketing, Gent University, Belgium.