Data Fusion for Direct Marketing

Peter van der Putten

1 Introduction

With no data, there is nothing to mine in. Multiple of sources of data can exist, and linking this data together can be non trivial. Assume we are given an instance, representing for example a customer. The problem of merging information from different sources about this particular instance, assuming it can't be done with simple joins, is also called the exact matching problem (Radner et al. (1980)). In contrast, enriching the data for this instance with information from other instances is called a statistical matching or data fusion problem, which is the topic of our research.

In literature data fusion is almost exclusively used in a market research or socio-economic survey context, to merge information from samples with different sets of questions, to reduce the response burden or to connect survey data that has previously not been studied jointly. The resulting surveys are then typically mined using simple techniques such as cross tabulations and correlation analysis. However for direct marketing, information is required for every single customer to allow for personalized one to one communication, so typically a customer database is fused with surveys. The fused data can then directly be exploited, for instance for rule based segmentation (van Hattum & Hoijtink (2008)), or the enriched data can be used to build prediction models with an improved performance or that are easier to understand (van der Putten et al. (2002)).

The goal of our paper is not to present new (or any) algorithms, but it is an application oriented position paper to discuss opportunities and challenges for applying data fusion in a data mining for direct marketing context. To guide the discussion we will use a a proof of principle example that demonstrates data fusion can add value for predictive modeling for direct marketing (van der Putten et al. (2002), van der Putten (2010)). The results will generalize to any case where there is an interest to enrich data that will then be used for predictive modeling.

Peter van der Putten,
LIACS, Leiden University, The Netherlands, e-mail: putten@liacs.nl

Table 1 External evaluation results: using enriched data generally leads to improved performance in this example.

	Only common variables	**Common and correlated fusion variables**	**Common and all fusion variables**
SCG neural network	c=0.692 ± 0.012	c=0.703 ± 0.015 p=0.041	c=0.694 ± 0.019 p=0.38
Linear regression	c=0.692 ± 0.014	c=0.724 ± 0.012 p=0.000	c=0.713 ± 0.013 p=0.002
Naive Bayes Gaussian	c=0.701 ± 0.015	c=0.720 ± 0.012 p=0.003	c=0.719 ± 0.012 p=0.005
Naive Bayes multinomial	c=0.707 ± 0.015	c=0.720 ± 0.011 p=0.200	c=0.704 ± 0.009 p not relevant
***k*-nearest neighbor**	c=0.702 ± 0.012	c=0.716 ± 0.013 p=0.0093	c=0.720 ± 0.012 p=0.0023

2 Main results

In Table 1 the results of the proof of principle experiment can be found. Using real world data we simulated a use case in which a customer database was enriched with survey data through data fusion, and models were built to predict credit card ownership with or without fused survey data using a variety of algorithms. Using the survey data generally leads to better models (the *c* measure is similar to AUC). This example can then be used to discuss challenges and opportunities for applying data fusion for direct marketing.

References

Radner, D., Rich, A., Gonzalez, M., Jabine, T. & Muller, H. (1980), Report on exact and statistical matching techniques. statistical working paper 5, Technical report, Office of Federal Statistical Policy and Standards US DoC.

van der Putten, P. (2010), On Data Mining in Context: Cases, Fusion and Evaluation, PhD thesis, Leiden Institute of Advanced Computer Science (LIACS), Leiden University.

van der Putten, P., Kok, J. N. & Gupta, A. (2002), Why the information explosion can be bad for data mining, and how data fusion provides a way out, *in* R. L. Grossman, J. Han, V. Kumar, H. Mannila & R. Motwani, eds, 'Second SIAM International Conference on Data Mining (SDM 2002)', SIAM, pp. 128–138.

van Hattum, P. & Hoijtink, H. (2008), 'The proof of the pudding is in the eating. data fusion: an application in marketing', *Journal of Database Marketing and Customer Strategy Management* **15**(4), 267–284.