Knowledge Discovery Using Least Squares Support Vector Machine Classifiers: A Direct Marketing Case

S. Viaene[1], B. Baesens[1], T. Van Gestel[2], J.A.K. Suykens[2], D. Van den Poel[3], J. Vanthienen[1], B. De Moor[2], and G. Dedene[1]

[1] K.U.Leuven, Dept. of Applied Economic Sciences,
Naamsestraat 69, B-3000 Leuven, Belgium
{Stijn.Viaene,Bart.Baesens,Jan.Vanthienen,Guido.Dedene}@econ.kuleuven.ac.be
[2] K.U.Leuven, Dept. of Electrical Engineering ESAT-SISTA,
Kardinaal Mercierlaan 94, B-3001 Leuven, Belgium
{Tony.Vangestel,Johan.Suykens,Bart.Demoor}@esat.kuleuven.ac.be
[3] Ghent University, Dept. of Marketing,
Hoveniersberg 24, B-9000 Ghent, Belgium
{Dirk.Vandenpoel}@rug.ac.be

Abstract. The case involves the detection and qualification of the most relevant predictors for repeat-purchase modelling in a direct marketing setting. Analysis is based on a wrapped form of feature selection using a sensitivity based pruning heuristic to guide a greedy, step-wise and backward traversal of the input space. For this purpose, we make use of a powerful and promising least squares version (LS-SVM) for support vector machine classification. The set-up is based upon the standard R(ecency) F(requency) M(onetary) modelling semantics. Results indicate that elimination of redundant/irrelevant features allows to significantly reduce model complexity. The empirical findings also highlight the importance of Frequency and Monetary variables, whilst the Recency variable category seems to be of lesser importance. Results also point to the added value of including non-RFM variables for improving customer profiling.

1 Introduction

The main objective of this paper involves the detection and qualification of the most relevant variables for repeat-purchase modelling in a direct marketing setting. This knowledge is believed to vastly enrich customer profiling and thus contribute directly to more targeted customer contact.

The empirical study focuses on the *purchase incidence*, i.e. the issue whether or not a purchase is made from any product category offered by the direct mailing company. Standard R(ecency) F(requency) M(onetary) modelling semantics underly the discussed purchase incidence model [3]. This binary (buyer vs. non-buyer) classification problem is being tackled in this paper by using least squares support vector machine (LS-SVM) classifiers. LS-SVM's have recently been introduced in the literature [11] and excellent benchmark results have been

D.A. Zighed, J. Komorowski, and J. Żytkow (Eds.): PKDD 2000, LNAI 1910, pp. 657–664, 2000.

reported [13]. Having constructed an LS-SVM classifier with all available predictors, we engage in a feature selection experiment. Feature selection has been an active area of research in the data-mining field for many years now. A compact, yet highly accurate model may come in very handy in (on-line) customer profiling systems. Furthermore, elimination of redundant and/or irrelevant features often improves the predictive power of a classifier, in addition to reducing model complexity. On top, by reducing the number of input features, both human understanding and computational performance can often be vastly enhanced.

Section 2 briefly elaborates on some response modelling issues including the description of the data set. In Section 3, we discuss the basic underpinnings of LS-SVM's for binary classification. The feature selection experiment and corresponding results are presented and discussed in Section 4.

2 The Response Modelling Case for Direct Marketing

2.1 Response Modelling and RFM

For mail-order response modelling, several alternative problem formulations have been proposed based on the choice of the dependent variable. In purchase incidence modelling [1,12] the main question is whether a customer will purchase during the next mailing period. Other authors have investigated related problems dealing with both the purchase incidence and the amount of purchase in a joint model [7]. A third alternative perspective for response modelling is to model interpurchase time through survival analysis or (split-)hazard rate models [4]. The purchase incidence model in our experiment uses the traditionally discussed (R)ecency, (F)requency and (M)onetary variables as the main predictor categories. In addition, some extra historical customer profiling variables have been included in the data set. This choice is motivated by the fact that most previous research cites them as being most predictive and because they are internally available at very low cost.

Cullinan [3] is generally considered as being the pioneer of RFM modelling in direct marketing. Since then, the literature has accumulated so many uses of these three variable categories, that there is overwhelming evidence both from academically reviewed studies as well as from practitioners' experience that the RFM variables are the most important set of predictors for modelling mail-order repeat purchasing [1,6]. However, when browsing the vast amount of literature, it becomes evident that only very limited attention has been devoted to selecting the right set of variables (and their operationalisations) for inclusion into the model of mail-order repeat buying.

2.2 The Data Set

We obtained Belgian data on past purchase behaviour at the order-line level, i.e. we know when a customer purchased what quantity of a particular product at what price as part of what order. The total sample size amounts to 5000 customers, of which 37.9% represented buyers. The (R)ecency, (F)requency and

Table 1. A listing of all features (both RFM and non-RFM) included in the direct marketing case.

Recency	Frequency	Monetary	Other
RecYearR	FrYearR	MonHistR	ProdclaT
RecYearN	FrYearN	MonHistN	ProdclaM
RecHistR	FrHistR	MonYearR	GenCust
RecHistN	FrHistN	MonYearN	GenInfo
		Ln(MonHistR)	Ndays
		Ln(MonHistN)	IncrHist
		Ln(MonYearR)	IncrYear
		Ln(MonYearN)	RetMerch
			RetPerc

(M)onetary variables have then been modelled as described in detail in [12]. Here, we briefly cover the basic semantics of the variables included in Table 1.

We used two time horizons for all RFM variables. The Hist horizon refers to the fact that the variable is measured between the period 1 July 1993 until 30 June 1997. The Year horizon refers to the fact that the variable is measured over the last year. All RFM variables have been modelled both with and without the occurrence of returned merchandise, indicated by R and N, respectively. Taking into account both time horizons (Year versus Hist) and inclusion versus exclusion of returned items (R versus N), we arrive at a 2×2 design in which each RFM variable is operationalised in 4 ways. The Recency variable is operationalised as the number of days since the last purchase. The Monetary variable is modelled as the total accumulated monetary amount of spending by a customer. Additionally, we include the natural log transformation (Ln) of all monetary variables as a means to reduce the skewness of the data distribution. The Frequency variable measures the number of purchase occasions in a certain time period. Apart from the RFM variables, we also included 9 other customer-profiling features, which have also been discussed in detail in [12]. The ProdclaT respectively ProdclaM variables represent the Total respectively Mean forward-looking weighted productindex. The weighting procedure represents the 'forward-looking' nature of a product category purchase, derived from another sample of data. The GenCust and GenInfo variables model the customer/company interaction on the subject of information requests and complaints. The length of the customer relationship is quantified by means of the Ndays variable. The IncrHist and IncrYear variables measure the increased spending frequency over the entire customer history and over the last year, respectively. The RetMerch variable is a binary variable indicating whether the customer has ever returned an item, that was previously ordered from the mail-order company. The RetPerc variable measures the total monetary amount of returned orders divided by the total amount of spending.

Notice that all missing values were handled by the mean imputation procedure [8] and that all predictor variables were normalized to zero mean and unit variance prior to their inclusion in the model.

3 Least Squares SVM Classification

3.1 The LS-SVM Classifier

Given a Training set of N data points $\{y_k, x_k\}_{k=1}^N$, where $x_k \in \Re^n$ is the k-th input pattern and $y_k \in \{-1, 1\}$ is the k-th output pattern, Vapnik's SVM classifier formulation [2,9,14] is modified by Suykens [11] into the following LS-SVM formulation:

$$\min_{w,b,e} \mathcal{J}(w, e) = \frac{1}{2} w^T w + \gamma \frac{1}{2} \sum_{i=1}^{N} e_i^2 \tag{1}$$

subject to the equality constraints

$$y_i \left[w^T \varphi(x_i) + b \right] = 1 - e_i, \;\; i = 1, ..., N. \tag{2}$$

This formulation now consists of equality instead of inequality constraints and takes into account a squared error with regularization term similar to ridge regression. The solution is obtained after constructing the Lagrangian $\mathcal{L}(w, b, e; \alpha) =$

$$\mathcal{J}(w, e) - \sum_{i=1}^{N} \alpha_i \{ y_i [w^T \varphi(x_i) + b] - 1 + e_i \} \tag{3}$$

where α_i are the Lagrange multipliers. After taking the conditions for optimality, one obtains the following linear system [11]:

$$\left[\begin{array}{c|c} 0 & Y^T \\ \hline Y & \Omega + \gamma^{-1} I \end{array} \right] \left[\begin{array}{c} b \\ \hline \alpha \end{array} \right] = \left[\begin{array}{c} 0 \\ \hline \mathbf{1} \end{array} \right] \tag{4}$$

where $Z = [\varphi(x_1)^T y_1; ...; \varphi(x_N)^T y_N]$, $Y = [y_1; ...; y_N]$, $\mathbf{1} = [1; ...; 1]$, $\alpha = [\alpha_1; ...; \alpha_N]$, $\Omega = Z Z^T$ and Mercer's condition [11] is applied within the Ω matrix

$$\begin{aligned} \Omega_{ij} &= y_i y_j \, \varphi(x_i)^T \varphi(x_j) \\ &= y_i y_j \, K(x_i, x_j). \end{aligned} \tag{5}$$

For the kernel function $K(\cdot, \cdot)$ one typically has the following choices: $K(x, x_i) = x_i^T x$ (linear kernel), $K(x, x_i) = (x_i^T x + 1)^d$ (polynomial kernel of degree d), $K(x, x_i) = \exp\{-\|x - x_i\|_2^2 / \sigma^2\}$ (RBF kernel), $K(x, x_i) = \tanh(\kappa \, x_i^T x + \theta)$ (MLP kernel), where d, σ, κ and θ are constants. Notice that the Mercer condition holds for all σ and d values in the RBF and the polynomial case, but not for all possible choices of κ and θ in the MLP case. The LS-SVM classifier is then constructed as follows:

$$y(x) = \text{sign} \left(\sum_{i=1}^{N} \alpha_i y_i K(x, x_i) + b \right). \tag{6}$$

Note that the matrix in (4) is of dimension $(N+1) \times (N+1)$. For large values of N, this matrix cannot easily be stored, such that an iterative solution method for solving it is needed. A Hestenes-Stiefel conjugate gradient algorithm is suggested in [10] to overcome this problem. Basically, the latter rests upon a transformation of the matrix in (4) to a positive definite form [10].

3.2 Calibrating the RBF LS-SVM Classifier

All classifiers were trained using RBF kernels. Estimation of the generalisation ability of the RBF LS-SVM classifier is then realised by the following experimental set-up [13]:

1. Set aside $\frac{2}{3}$ of the data for the training/validation set and the remaining $\frac{1}{3}$ for testing.
2. Perform 10-fold cross validation on the training/validation data for each (σ, γ) combination from the initial candidate tuning sets Σ and Γ typically chosen as follows :
   $\Sigma = \{0.5, 5, 10, 15, 25, 50, 100, 250, 500\} \cdot \sqrt{n}$,
   $\Gamma = \{0.01, 0.5, 1, 10, 50, 100, 500, 1000\} \cdot \frac{1}{N}$.
   The square root $\sqrt{n}$ of the number of inputs n is introduced since $||x - x_i||_2^2$ in the RBF kernel is proportional to n and the factor $1/N$ is introduced such that the misclassification term $\gamma \sum_{i=1}^{N} e_i^2$ is normalized with the size of the data set.
3. Choose optimal (σ, γ) from the initial candidate tuning sets Σ and Γ by looking at the best cross validation performance for each (σ, γ) combination.
4. Refine Σ and Γ iteratively by means of a grid search mechanism in order to further optimize the tuning parameters (σ, γ). In our experiments, we repeated this step three times.
5. Construct the LS-SVM classifier using the total training/validation set for the optimal choice of the tuned hyperparameters (σ, γ).
6. Assess the generalization ability by means of the independent test set.

4 The Feature Selection Experiment

Feature selection effectively starts at the moment the LS-SVM classifier has been constructed on the full set of n available predictors. The feature selection procedure is based upon a (greedy) best-first heuristic, guiding a backward search mechanism through the feature space [5]. The mechanics of the implemented heuristic for assessing the sensitivity of the classifier to a certain input feature are quite straightforward. We apply a strategy of constant substitution in which a feature is perturbed to its mean while all other features keep their values and compute the impact of this operation on the performance of the obtained LS-SVM classifier without re-estimation of the LS-SVM parameters α_k and b. This assessment is done using the separate Pruning set, in order to obtain an unbiased estimate of the change in classification accuracy of the constructed classifier. Fig. 1 provides a concise overview of the different steps of the experimental procedure.

Starting with a full feature set F_1, all n inputs are pruned sequentially, i.e. one by one. The first feature f_k to be removed, is determined at the end of *Step* 1 (task (4)). After having removed this feature from F_1, the reduced feature set $F_2 = F_1 \setminus \{f_k\}$ is used for subsequent feature removal. At this moment, an iteration of identical *Steps* i is started, in which, in a first phase, the LS-SVM parameters α_k and b are re-estimated on the Training set (task (1) of *Step* i),

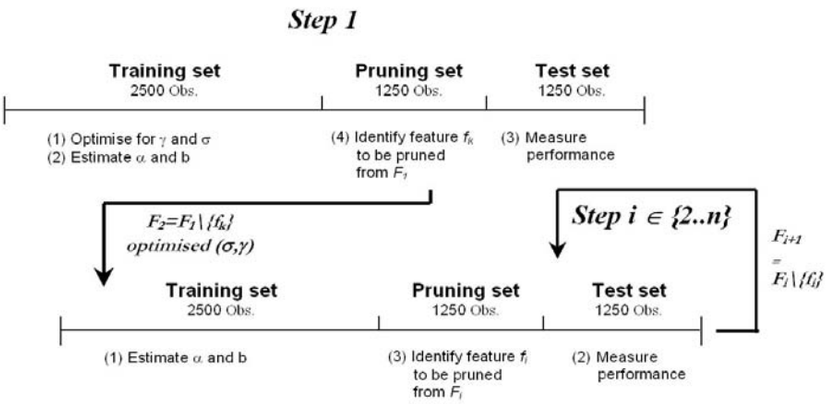

Fig. 1. Experimental set-up consisting of a first step for constructing an optimised LS-SVM classifier on a full feature set and of a subsequent iteration of pruning *Steps i*.

Table 2. Empirical assessment of the RBF LS-SVM classifier trained on the full feature set, i.e. *Full model*, vis-a-vis the RBF LS-SVM classifier trained on the reduced feature set, i.e. *Reduced Model*. *Majority* stands for the majority prediction error.

Results	*Full Model*	*Reduced Model*	*Majority*
Training (2500 Obs.)	77.36%	76.04%	62%
Pruning (1250 Obs.)	76.72%	77.20%	62%
Test (1250 Obs.)	73.92%	73.52%	62%
Features	25	9	0

however without re-calibrating for σ and γ[1], and generalisation ability of the classifier is quantified on the independent Test set (task (2) of *Step i*). Again, feature sensitivities of the resulting classification model (without re-estimation of α_k and b) are assessed on the Pruning set to identify the feature to which the classifier is least sensitive when perturbed to its mean (task (3) of *Step i*). This feature is then pruned from the remaining feature subset and disregarded for further analysis. The pruning procedure is thereupon resumed with a reduced feature set (*Step* $i+1$), until all input features are eventually removed. Once all features have been pruned, the prefered reduced model is then determined by means of the highest Pruning set performance.

[1] Notice that the originally optimised γ and σ values obtained in task (1) of *Step* 1 remain unchanged during the entire feature selection phase. Experimental evaluation showed that this implementation heuristic significantly speeds up the pruning procedure, without having a detrimental effect on the predictive performance of the reduced feature set.

Table 3. Order of feature removal using the pruning procedure presented in section 4. Each feature is qualified by its category with r, f, m, o respectively standing for recency, frequency, monetary and other (cf. Table 1).

Pruning Steps									
1-5		**6-10**		**11-15**		**16-20**		**21-25**	
RetPerc	o	ProdclaM	o	RecHistN	r	FrYearN	f	MonYearR	m
Ln(MonHistN)	m	MonHistR	m	IncrHist	o	Ln(MonHistR)	m	MonYearN	m
RecHistR	r	IncrYear	o	RecYearR	r	MonHistN	m	GenInfo	o
Ndays	o	Ln(MonYearR)	m	RecYearN	r	GenCust	o	FrHistR	f
ProdclaT	o	Ln(MonYearN)	m	FrHistN	f	RetMerch	o	FrYearR	f

Table 2 summarises the empirical findings of the pruning procedure for the RFM case. We contrasted the full model results with those of a binary logistic regression and concluded that the RBF LS-SVM classifier outperformed the latter significantly. Observe how the suggested feature selection method allows to significantly reduce the model complexity (from 25 to 9 features) without any significant degradation of the generalisation behaviour on the independent Test set. The Test set performance amounts to 73.92% for the full model and 73.52% for the reduced model.

The order of feature removal as depicted in Table 3, provides further insight into the relative importance of the predictor categories (cf. Table 1). The reduced model consists of the 9 features that are underlined in Table 3. This reduced set of predictors consists of Frequency, Monetary value and other (non-RFM) variables. It is especially important to note that the reduced model includes information on returned merchandise. Furthermore, notice the absence of the Recency component in the reduced feature set. Inspection of the order of removal of features, while further pruning this reduced feature set, highlights the importance of the Frequency variables. More specifically, the last two variables to be removed belong to this predictor category. Remark that a feature set consisting of only these two features, still yields a percentage correctly classified of 72% on the Test set, which might be considered quite satisfactory.

5 Conclusion

In this paper, we applied an LS-SVM based feature selection wrapper to a real-life direct marketing case involving the modeling of repeat-purchase behaviour based on the well-known R(ecency) F(requency) M(onetary) framework. The sensitivity based, step-wise feature selection method constructed as a wrapper around the LS-SVM classifier allows to significantly reduce model complexity without degrading predictive performance. The empirical findings highlight the role of Frequency and Monetary variables in the reduced model, whilst the Recency variable category seems to be of lesser importance within the RFM model. Results also point to the beneficial effect of including non-RFM customer profiling variables for improving predictive accuracy.

Acknowledgements. This work was partly carried out at the Leuven Institute for Research in Information Systems (L.I.R.I.S.) of the Dept. of Applied Economic Sciences of the K.U.Leuven in the framework of the KBC Insurance Research Chair. This work was partially carried out at the ESAT laboratory and the Interdisciplinary Center of Neural Networks ICNN of the KULeuven and supported by grants and projects from the Flemish Gov. (Res. Counc. KULeuven: GOA-Mefisto; FWO-Flanders: res. projects and comm (ICCoS & ANMMM); IWT: STWW Eureka); the Belgium Gov. (IUAP-IV/02, IV-24); the Eur. Comm. (TMR, ERNSI). S. Viaene, holder of the KBC Research Chair, and B. Baesens are both Research Assistants of L.I.R.I.S. T. Van Gestel is a research assistant, J. Suykens is a postdoctoral researcher, B. De Moor is a senior research associate with the Fund for Scientific Research Flanders (FWO-Flanders), resp. D. Van den Poel is an assistant professor at the department of marketing at Ghent University. J. Vanthienen and G. Dedene are Senior Research Associates of L.I.R.I.S.

References

1. Bauer, A.: A direct mail customer purchase model. *Journal of Direct Marketing*, 2(3):16–24, 1988.
2. Cristianini, N. and Shawe-Taylor, J.: *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.
3. Cullinan, G. J.: *Picking them by their batting averages' recency-frequency-monetary method of controlling circulation*. Manual release 2103. Direct Mail/Marketing Association. N.Y., 1977.
4. Dekimpe, M. G. and Degraeve, Z.: The attrition of volunteers. *European Journal of Operations Research*, 98:37–51, 1997.
5. John, G., Kohavi, R. and Pfleger, K.: Irrelevant features and the subest selection problem. *Machine Learning: proceedings of the Eleventh International Conference, San Francisco*, pages 121–129, 1994.
6. Kestnbaum, R. D.: Quantitative database methods. *The Direct Marketing Handbook*, pages 588–597, 1992.
7. Levin, N. and Zahavi, J.: Continuous predictive modeling: a comparative analysis. *Journal of Interactive Marketing*, 12(2):5–22, 1998.
8. Little, R.J.A.: Regression with missing x's: a review. *Journal of the American Statistical Association*, 87(420):1227–1230, 1992.
9. Schölkopf, B., Burges, C., and Smola, A.: *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1998.
10. Suykens, J. A. K., Lukas, L., Van Dooren, P., De Moor, B. and Vandewalle, J.: Least squares support vector machine classifiers: a large scale algorithm. *ECCTD'99 European Conf. on Circuits Theory and Design, Stresa, Italy*, pages 839–842, 1999.
11. Suykens, J. A. K. and Vandewalle, J.: Least squares support vector machine classifiers. *Neural Processing Letters*, 9(3):293–300, 1999.
12. Van den Poel, D. : *Response Modeling for Database Marketing using Binary Classification*. Phd. Dissertation. K.U. Leuven, 1999.
13. Van Gestel, T., Suykens, J.A.K., Baesens, B., Viaene, S., Vanthienen, J., Dedene, G., De Moor, B. and Vandewalle, J.: Benchmarking least squares support vector machine classifiers. *CTEO, Technical Report 0037, K.U. Leuven, Belgium*, 2000.
14. Vapnik, V.: *Statistical learning theory*. John Wiley, New-York, 1998.