Subgroup evaluation and decision support for a direct mailing marketing problem

Peter Flach[1] and Dragan Gamberger[2]

[1] University of Bristol, The Merchant Venturers Building Woodland Road, Bristol BS8 1UB, United Kingdom E-mail: Peter.Flach@bristol.ac.uk
[2] Rudjer Bošković Institute, Bijenička 54,10000 Zagreb, Croatia E-mail: dragan.gamberger@irb.hr

Abstract. In this work we use ROC (Receiver Operating Characteristic) analysis to evaluate customer subgroups detected by different machine learning approaches in a marketing database. A direct mailing model with a marginal cost per mailing and an average expected profit per new customer has been assumed. In order to identify optimal mailing strategies for different marketing situations, we introduce the *normalised profit curve*, which extends the ROC curve by not only identifying the optimal subgroup in a given context, but also indicating the expected profit. In this sense, the analysis presents a link between data mining and decision support.

1 Introduction

The dataset investigated in this work is a relational database obtained by interviewing potential customers. Customer answers about how they recognize, appreciate, and use tested brands are the main part of the database. The subject of the interviews was 300 brands selling food and beverages. The customers are described by their answers about age, educational level, profession, place of living, preferences, and habits like what TV programmes they watch and what newspapers they read regularly.

The dataset presents a very good starting point for different kinds of marketing analysis. Interesting questions are which brands have the potential to improve their recognition or usage rate, what are characteristics of the people appreciating or using a specific brand, what is the nature of the relation between brand recognition and brand usage, and so on. The subject of this paper is selection of potential customer subgroups that can be targeted by advertising campaigns. The people were classified according to their recognition of the brand in two groups. The group of people who do not know the brand was selected as the target (positive) class for the data mining process. The task was to find their significant characteristics relative to the characteristics of the population that recognizes the brand (negative class) and to determine if and how an advertising campaign could increase the expected brand profit.

We use various forms of subgroup discovery to characterise interesting subgroups of the target class (prospective customers). We use ROC (Receiver Operating Characteristic) analysis to evaluate the interest of the subgroups, and also to evaluate the interest of multiple combined subgroups. Since we employ a direct mailing model, the main

interest of targeting a subgroup for direct mailing is that it may result in new customer and thus profit. We propose a new measure of *normalised profit*, which corresponds to the profitable proportion of true positives.

The outline of the paper is as follows. In Section 2 we review ROC analysis in the context of subgroup discovery and direct mailing. In Section 3 we define normalised profit. Section 4 presents some experimental results, and Section 5 concludes.

2 Subgroup discovery and ROC analysis

ROC analysis [3] is usually applied in domains such as medicine which are cost-sensitive: e.g., the cost of a false negative (an ill patient who is considered healthy) is much higher than the cost of a false positive (a healthy patient who is considered ill). Rather than fixing these cost parameters, ROC analysis portrays the performance of a classifier under several possible parameter settings. It allows, through the construction of the convex hull of a set of points, identification of classifiers that are optimal under certain parameter settings [4]. Once the application context is known (i.e., distribution of positives and negatives and misclassification costs) the optimal classifier can be determined from the convex hull.

In this paper we work with a direct mailing model, where a mailing is sent out to all people covered by a rule or set of rules [2]. We assume a marginal cost c per mailing, and an average profit g per true positive (people who receive a mailing and may become customers); this includes the cost of the mailing. So, for instance, if 10% of the true positives reached by the mailing are expected to become customers and spend 1000 Euros each, while the marginal cost per mailing is 1 Euro, then g is 99 Euros. The default decision would be to send a mailing to everybody in the population, resulting in a default profit $g * Pos - c * Neg$, where Pos (Neg) is the number of people who will (will not) become a customer after receiving a letter, and $N = Pos + Neg$ is the size of the population. The context here is defined as a four-tuple (c, g, Pos, Neg). Note that the default profit may be negative if positives are rare and/or mailings are expensive. We are only interested in rules that improve upon this default profit (or yield positive profit if the default profit is negative).

This is usually characterised as a subgroup discovery problem: identify a subset of the population whose class distribution is significantly different from the distribution over the whole population (in this case, we are only interested in subgroups, which have a larger proportion of positives than default). While subgroup discovery is usually seen as different from classification because we can tolerate many more false positives, it can be unified under the umbrella of cost-sensitive classification because for deciding which rules are optimal in a given context it doesn't matter whether we punish false negatives as in classification or reward true positives as in subgroup discovery (it only matters for determining expected profit in a given context) [6].

As an illustration, Figure 1 shows the performance of several rules found by different algorithms in ROC space. The horizontal axis in this figure is the false positive rate fp defined as the fraction of false positives FP (negative cases classified erroneously by the rule as positive cases) relative to the number of all negative examples Neg. The vertical axis is the true positive rate tp defined as the fraction TP/Pos where TP are

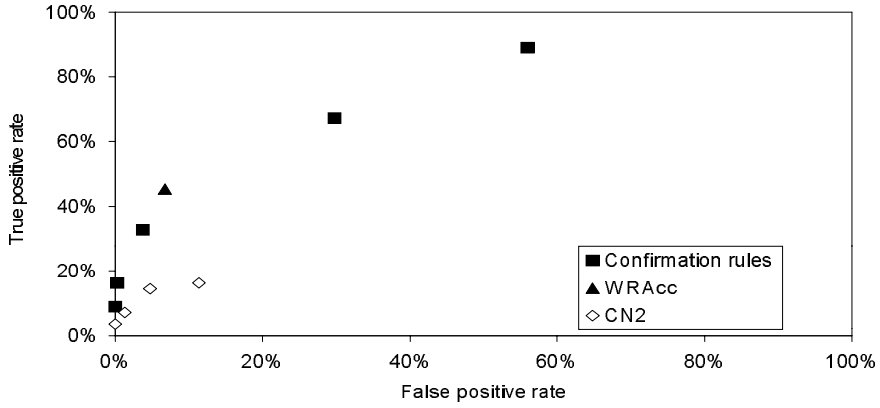

Fig. 1. Rules plotted in ROC space.

positive examples correctly predicted by the rule. Confirmation rules are found with an ILLM system algorithm [1], CN2 is a sequential covering rule learner, and WRAcc is a modified version of CN2 using weighted relative accuracy as a rule evaluation heuristic [5]. From this graph it can be seen that the rules found by CN2 are too specific and sub-optimal in any context. For instance, the leftmost CN2 rule covers 2 true positives and no false positives, but the leftmost confirmation rule A covers 5 true positives and no false positives and will therefore always be preferable.

The convex hull of these points or ROC curve is shown in Figure 2. It includes all 5 confirmation rules plus the WRAcc rule, and the trivial rules, which classify everything as negative (0,0) and positive (1,1), respectively. Each of these 8 subgroups is optimal in a particular context. For instance, if mailings are cheap enough and/or positives are frequent enough the optimal strategy will be to mail everyone.

3 Normalised profit

In the direct mailing model the lines of equal profit are important in the ROC space. The expected profit of any rule can be determined by the relation $Profit = g * TP - c * FP$ which is equal to $Profit = g * Pos * tp - c * Neg * fp$. Lines with equal profit $Profit$ are defined by the equation $tp = (c * Neg/g * Pos) * fp + Profit/(g * Pos)$. In other words, in a given context (c, g, Pos, Neg) rules with different (tp, fp) values will have the same profit values if lying on the same line with slope $c * Neg/(g * Pos) = (c/g) * (Neg/Pos)$ in ROC space. Movements in ROC space along lines of equal profit will not change the amount of total profit while movements upward or downward will increase or decrease the profit, respectively. Movements upward will increase the total profit because in such case the number of TP cases can increase while the number of FP cases remains the same.

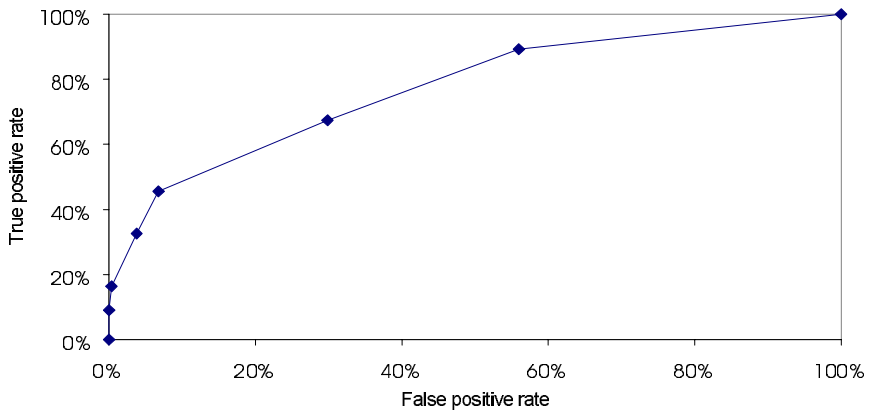

Fig. 2. The ROC curve corresponding to the points in Figure 1.

Notice that the slope of equal profit lines is completely defined by parameters that describe the intended context: cost c per mailing, average expected profit g per new customer, and total estimated numbers Pos and Neg of potential customers and non-customers in the population, respectively. We call the resulting slope the *intended slope*. The main advantage of ROC analysis is that we can train our learning algorithms without knowledge of the intended context and slope; once these are known, we can select the rule that is optimal for that context. The optimal rule is the point which has an equal profit line as its tangent (if such a point exists). This follows from the convexity of the ROC curve (all other points are downwards with respect to the equal profit line through the so selected rule).

As an example, assume $c = 1$, $g = 9$, $Pos = 55$ and $Neg = 1070$, then the intended slope is 2.16. If there is a line segment with this slope then the point on either end of the line segment would be optimal; if the slope of the line segment is slightly higher (lower) than the intended slope, then the right (left) point would be optimal. In other words, we are looking for the point connecting a line segment with higher slope to a line segment with lower slope. The ROC curve provides a solution to the decision problem: which rule to apply in a given context. Thus, ROC analysis provides a link between data mining and decision support.[1]

In order to make decisions easier and to show the expected profit explicitly, the ROC curve from Figure 2 can be transformed into the *normalised profit curve*. This curve is presented in Figure 3. Its horizontal axis is intended slope and the vertical axis is normalised profit. The base for profit normalization is maximal profit that could be obtained in the domain. It represents the profit when all potentially good customers

[1] Here, we use the term decision support in the narrow sense of selecting the decision with maximum expected utility.

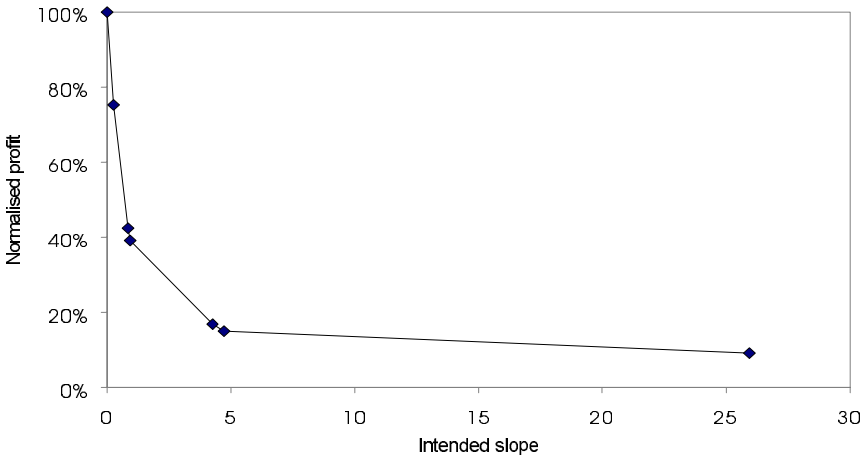

Fig. 3. Normalised profit curve corresponding to the ROC curve in Figure 2.

could be attracted without any expense lost on non-potential customers. Maximal profit corresponds to the ideal rule with $tp = 100\%$ and $fp = 0\%$, and its value is $g * Pos$. From the expression for the expected profit $Profit = g * Pos * tp - c * Neg * fp$ it follows that normalised profit value is $Profit/(g * Pos) = tp - (intended\ slope) * fp$, where tp and fp are true and false positive rates of the rule used to select a subgroup as the mailing target.

The importance of the last expression is that it enables us to determine the optimal subgroup **and** calculate its normalised profit once the parameters of the concrete mailing environment are known; thus, it contains more information than the ROC curve. In a given context, the normalised profit demonstrates how much of the positives can be reached with the best rule, corrected with the cost of addressing negatives. For instance, a normalised profit of 40% may mean that we reached 50% of positives, but 1/5 of the profit was spent on the negatives addressed; or it may mean that we in fact reach 40% of the positives and no negatives. From the perspective of profit maximisation, both situations are equivalent. Similarly to a ROC curve, Figure 3 enables the user to select the optimal rule, which is the point immediately to the left of the intended slope. The leftmost point in Figure 3 is the default most general rule (send mail to everybody), followed by the other points from the ROC curve but in the reverse order (from most general to most specific rules). The last point is typically the point corresponding to the default most specific rule (send no letters). In Figure 3 the last point is the first rule from the ROC curve because it has $fp = 0$. Only in this specific case the normalised profit is always greater than zero.

Subgroup	False positive rate		True positive rate	
	mean	stdev	mean	stdev
CG0.9	0.03	0.05	0.14	0.12
CG0.7	0.13	0.03	0.41	0.13
CG0.5	0.42	0.09	0.50	0.20
CG0.3	0.60	0.09	0.64	0.09
CG0.2	0.72	0.21	0.71	0.13
w1	0.03	0.05	0.14	0.12
w3	0.10	0.05	0.27	0.24
w10	0.43	0.07	0.58	0.17
w30	0.45	0.05	0.68	0.18
w100	0.68	0.06	0.74	0.12
CG0.9+w1	0.03	0.05	0.14	0.12
CG0.9+w3	0.10	0.05	0.27	0.24
CG0.9+w10	0.46	0.08	0.61	0.20
CG0.9+w30	0.47	0.05	0.71	0.19
CG0.9+w100	0.68	0.06	0.76	0.13
CG0.7+w1	0.13	0.03	0.42	0.13
CG0.7+w3	0.17	0.02	0.42	0.13
CG0.7+w10	0.49	0.02	0.71	0.28
CG0.7+w30	0.51	0.02	0.79	0.26
CG0.7+w100	0.72	0.05	0.84	0.11
CG0.5+w1	0.45	0.13	0.54	0.23
CG0.5+w3	0.48	0.09	0.59	0.27
CG0.5+w10	0.47	0.05	0.64	0.22
CG0.5+w30	0.47	0.05	0.70	0.19
CG0.5+w100	0.68	0.06	0.74	0.12
CG0.3+w1	0.60	0.09	0.71	0.13
CG0.3+w3	0.62	0.10	0.75	0.19
CG0.3+w10	0.66	0.03	0.85	0.13
CG0.3+w30	0.64	0.02	0.85	0.13
CG0.3+w100	0.69	0.08	0.84	0.08
CG0.2+w1	0.72	0.21	0.74	0.16
CG0.2+w3	0.74	0.21	0.78	0.20
CG0.2+w10	0.78	0.12	0.90	0.11
CG0.2+w30	0.76	0.15	0.90	0.11
CG0.2+w100	0.82	0.09	0.83	0.19

Table 1. Prediction results for 35 subgroups measured by 3-fold cross-validation for the brand ACI.

4 Some experiments

As an illustration of the approach outlined above, we have selected three brands from the questionnaire database, each with different percentages of people who really recognize the brand (i.e., negatives). The first is ACI, which is recognized by about 50% of the tested population. The second is FANTA with about 90% of good recognition (i.e., few positives) and the last is brand YO with only about 15% of successful recognition. For all brands the number of people who were explicitly asked about the firm is about 100 (N = 100).

Rules have been induced by two different systems. The first is the CN2 system that has been modified for this purpose so that a new cost-sensitive evaluation of the induced rules was introduced. In this mode the system uses a parameter called CG which reflects the target ratio of the cost of a single false positive prediction and the gain connected with a true positive prediction. In the experiments the following five parameter values have been used: 0.2, 0.3, 0.5, 0.7, and 0.9. This enabled induction

of rules of different generality for the same data set, where low CG values correspond to more general rules. The other system is ILLM used to induce confirmation rules. Here the generalization level parameter *w* is used to induce rules of different generality, low *w* values corresponding to more specific rules. The following *w* values have been used in the experiments: 1, 3, 10, 30, 100. For all firms five rules were induced by the CN2 system and five rules by the ILLM system. 25 additional subgroups have been obtained by disjunctive combination of one induced rule by the CN2 system and one rule obtained by ILLM (model combination). In this way a total of 35 points were obtained in ROC space. For every rule or pair of rules, true and false positive rates have been measured by cross-validation. 3-fold cross validation has been used because of the relative small number of examples in the training sets.

4.1 Results for ACI

The ACI dataset has about 50% of positives and negatives. It represents the brand for which many useful rules could be induced (Figure 4). Four of the subgroups defined by these rules are selected by the ROC analysis as significant (Figure 5) and, as normalised profit curve in Figure 6 demonstrates, the profit increase is possible for very different marketing situations

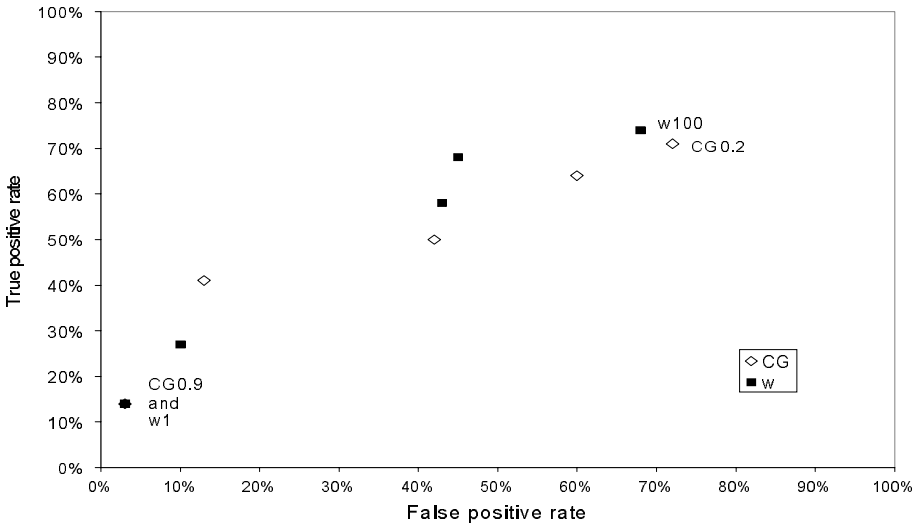

Fig. 4. Properties of 5 rules induced by CN2 system with cost-gain ratio values 0.9, 0.7, 0.5, 0.3, and 0.2 (CG series from left to right) and 5 rules induced as confirmation rules with generalization levels 1, 3, 10, 30, and 100 (*w* series from left to right).

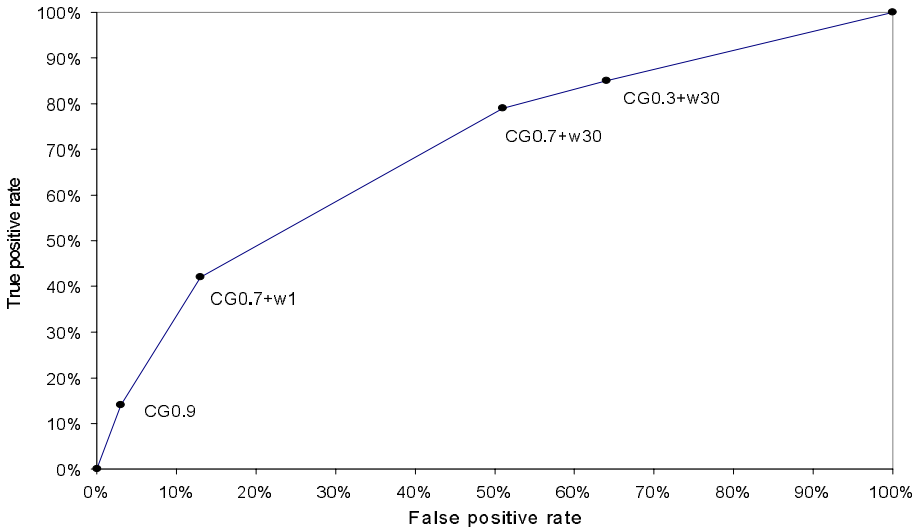

Fig. 5. ROC curve for brand ACI with four subgroups between default points (0,0) and (1,1). The selected subgroups are CG0.9 , CG0.7+w1 , CG0.7+w30 , and CG0.3+w30 from Table 1.

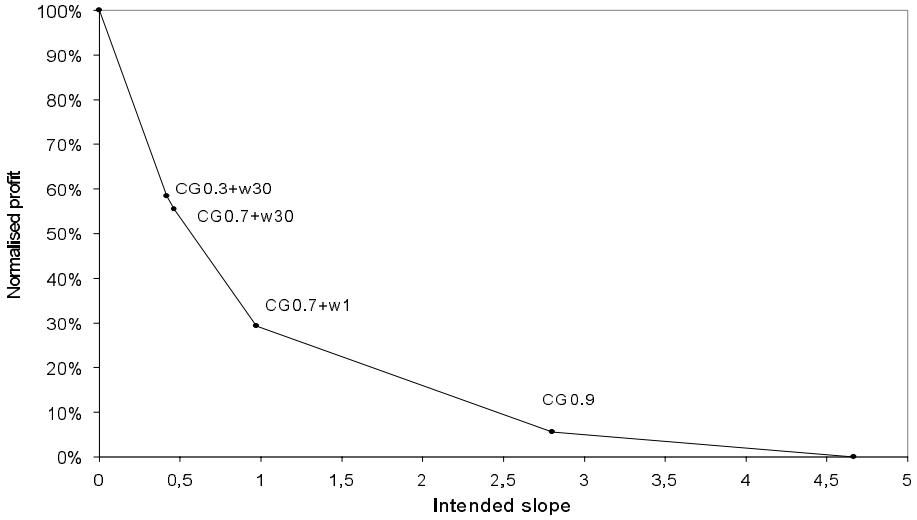

Fig. 6. Normalised profit curve for the brand ACI. The interpretation is as follows: for intended slopes up to about 0.4 use the default rule and send mail to everybody. The normalised profit will be between 100% and 60%, depending on the actual value of the intended slope. For intended slopes between about 0.4 and 0.5 select the subgroup CG0.3+w30 to send mail; normalised profit will be about 60% - 55%. For intended slopes higher than 0.5 mail subgroup CG0.7+w30 and so on. For intended slopes between 2.8 and 4.6 mail subgroup CG0.9 and normalised profit will be up to 8% and for intended slopes above 4.6 the best strategy is to send no letters.

4.2 Results for FANTA

The FANTA dataset has approximately 10% of positives and 90% negatives. As it turns out, neither ILLM nor cost-sensitive CN2 were able to find subgroups that would perform better than the default groups in some context. In ROC space, all points lie below the main diagonal. Interestingly, this could be remedied by taking the *complement* of the identified subgroups, which would mirror them through the $(0.5, 0.5)$ point, but we haven't followed that up in this work.

4.3 Results for YO

The YO dataset is also skewed, with approximately 15% negatives and 85% positives. Here we were able to identify one confirmation rule and one combined model to participate in the ROC curve and the normalised profit curve.

Subgroup	False positive rate		True positive rate	
	mean	stdev	mean	stdev
CG0.9	0.41	0.21	0.37	0.18
CG0.7	0.30	0.32	0.46	0.13
CG0.5	0.76	0.21	0.73	0.21
CG0.3	0.96	0.07	0.92	0.07
CG0.2	0.96	0.07	0.92	0.07
w1	0.26	0.17	0.33	0.17
w3	0.56	0.20	0.42	0.24
w10	0.39	0.48	0.71	0.22
w30	0.74	0.17	0.84	0.14
w100	0.98	0.04	0.91	0.06
CG0.9+w1	0.49	0.16	0.52	0.31
CG0.9+w3	0.60	0.23	0.52	0.18
CG0.9+w10	0.53	0.41	0.76	0.24
CG0.9+w30	0.76	0.21	0.87	0.17
CG0.9+w100	1.00	0.00	0.95	0.04
CG0.7+w1	0.52	0.19	0.65	0.17
CG0.7+w3	0.67	0.31	0.63	0.15
CG0.7+w10	0.42	0.52	0.73	0.25
CG0.7+w30	0.76	0.21	0.86	0.16
CG0.7+w100	1.00	0.00	0.94	0.06
CG0.5+w1	0.76	0.21	0.85	0.25
CG0.5+w3	0.81	0.17	0.80	0.20
CG0.5+w10	0.76	0.21	0.86	0.08
CG0.5+w30	0.81	0.17	0.90	0.06
CG0.5+w100	1.00	0.00	0.96	0.04
CG0.3+w1	0.96	0.07	0.96	0.04
CG0.3+w3	1.00	0.00	0.97	0.05
CG0.3+w10	0.96	0.07	0.96	0.04
CG0.3+w30	1.00	1.00	0.97	0.03
CG0.3+w100	1.00	0.00	0.98	0.03
CG0.2+w1	0.96	0.07	0.96	0.04
CG0.2+w3	1.00	0.00	0.97	0.05
CG0.2+w10	0.96	0.07	0.96	0.04
CG0.2+w30	1.00	1.00	0.97	0.03
CG0.2+w100	1.00	0.00	0.98	0.03

Table 2. Prediction results for 35 subgroups measured by cross-validation for the brand YO

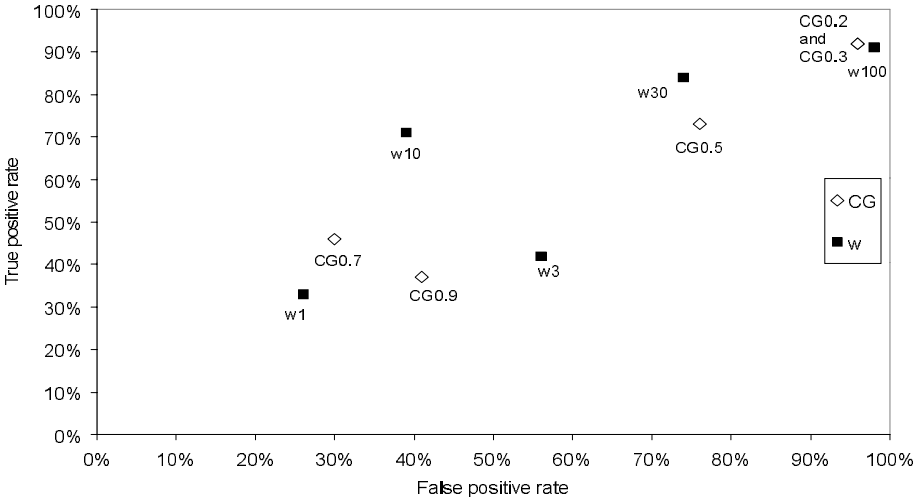

Fig. 7. Properties of 5 rules induced by CN2 system with cost-gain ratio values 0.9, 0.7, 0.5, 0.3, and 0.2 and 5 rules induced as confirmation rules with generalization levels 1, 3, 10, 30, and 100 .

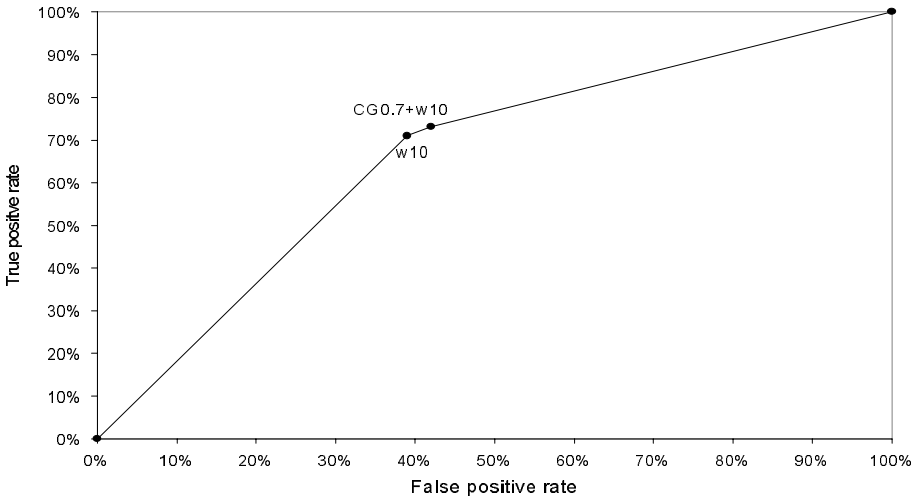

Fig. 8. ROC curve for brand YO with two points between (0,0) and (1,1). These points correspond to subgroups w10 and CG0.7+w10 from Table 2.

5 Discussion and suggestions for further work

This paper presented a possible application of the mailing marketing model to a questionnaire dataset. To this end, we introduced the novel concepts intended slope and normalised profit. These concepts allowed us to adapt the ROC methodology to the

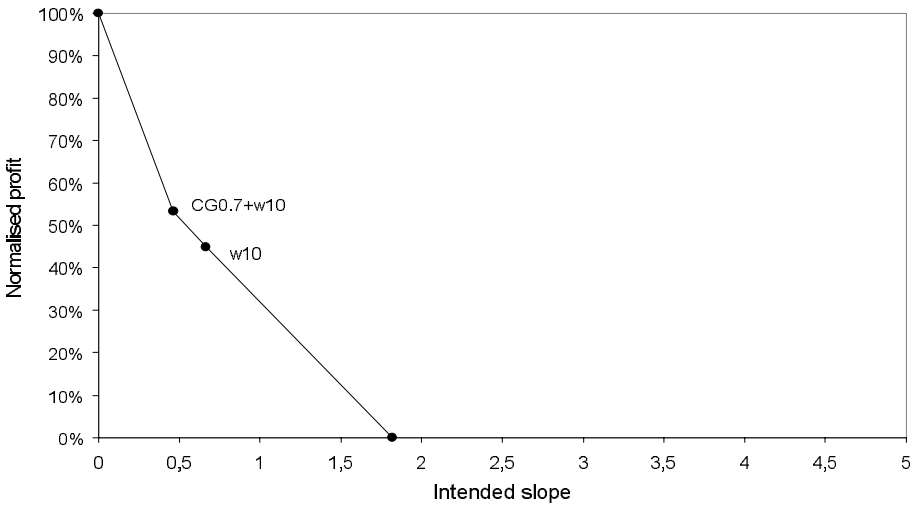

Fig. 9. Normalised profit curve for the brand YO.

marketing model, providing answers about optimal mailing strategy and the value of expected relative profit. Selection among multiple models establishes a decision support problem, to which the normalised profit curve is particularly well suited, since it does not only indicate the optimal subgroup in a given context, but also what the resulting profitable proportion of true positives is estimated to be. The approach assumes constant marginal cost of mailing and average profit per new customer – it could however be easily adapted to incorporate different values of these parameters per induced rule.

When compared with other performance measures used for identifying likely buyers in direct marketing problems based on *lift*, *lift index*, and the *area under the lift and ROC curves*, which have been presented in the work by Ling and Li, [2], the importance of our approach is in the fact that it is applicable also for pure classification rules without confidence measure. In this way a broader range of algorithms may be used for subgroup detection. Additionally, we have adapted the methodology developed by Provost and Fawcett [3] which enabled us not only to compare classification methods but to combine them into one solution so that it is obvious which classification method is optimal for the given situation. Although the work includes rules induced by different machine learning systems, the intention was not to compare them with respect to the obtained results. But, it may be assumed that inclusion of other systems, like MIDOS [6], and the possibility to choose among more classification methods and detected subgroups, might lead to better final profit gains.

For practical illustration three brands with significantly different target distributions were selected. For the ACI brand with even target distribution we demonstrated that significant improvements were obtainable over the two default strategies. The other two cases demonstrated that with very uneven class distributions, the default strategies are hard to beat. In the case of FANTA there are a small number of people who do not

recognize the brand and they all represent outliers for the data mining process. In this situation no reliable rule could be induced. In the case of YO there are many positive cases (people who might become customers once they know about the brand) but the prediction quality of induced rules is again low because of the lack of negative cases. However, two subgroups were identified that represented a significant improvement over the default strategies. Note that, in the absence of an intended context, we could express the overall improvement achieved by data mining by measuring the area under the ROC curve but above the diagonal.

We plan to carry out further work in collaboration with marketing experts. In this phase, practical applicability of the induced rules should be evaluated.

Acknowledgements

This work is supported by the Esprit V project IST-1999-11495 *Data Mining and Decision Support for Business Competitiveness: Solomon Virtual Enterprise*. Thanks are due to our partners in the project, in particular Nada Lavrač and Bojan Cestnik for providing us with the data. We also thank the anonymous reviewers for their insightful comments and suggestions.

References

1. Gamberger, D. and Lavrač, N. (2000) Confirmation rule sets. In D.A. Zighed, J. Komorowski, and J. Żytkow, editors, *4th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD2000)*, pp. 34-43. Springer-Verlag.
2. Ling, C. and Li, C. (1998) Data mining for direct marketing: problems and solutions . In R. Agrawal and P. Stolorz, editors, *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD98)*, pp. 73-79. AAAI Press.
3. Provost, F. and Fawcett, T. (2001) Robust Classification for Imprecise Environments. *Machine Learning* 42(3), pp. 203-231. http://www.stern.nyu.edu/ fprovost/Papers/rocch-mlj.pdf
4. Somoza, E., Soutullo-Esperon, L., and Mossman, D. (1989) Evaluation and optimization of diagnostic tests using ROC analysis and information theory. *International Journal of Biomedical Computing* 24, pp. 153-189.
5. Todorovski, L., Flach, P., and Lavrač, N. (2000) Predictive Performance of Weighted Relative Accuracy. In D.A. Zighed, J. Komorowski, and J. Żytkow, editors, *4th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD2000)*, pp. 255-264. Springer-Verlag.
6. Wrobel, S. (1997) An algorithm for multi-relational discovery of subgroups. *Proceedings of the First European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD'97)*, pp. 78-87. Springer-Verlag.