

Comparing Complete and Partial Classification for Identifying Latently Dissatisfied Customers

Tom Brijs, Gilbert Swinnen, Koen Vanhoof, and Geert Wets

Limburg University Centre, Universitaire Campus, gebouw D
B-3590 Diepenbeek, Belgium
tom.brijs@luc.ac.be

Abstract. This paper evaluates complete versus partial classification for the problem of identifying latently dissatisfied customers. Briefly, latently dissatisfied customers are defined as customers reporting overall satisfaction but who possess typical characteristics of dissatisfied customers. Unfortunately, identifying latently dissatisfied customers, based on patterns of dissatisfaction, is difficult since in customer satisfaction surveys, typically only a small minority of customers reports to be overall dissatisfied and this is exactly the group we want to focus learning on. Therefore, it has been claimed that since traditional (complete) classification techniques have difficulties dealing with highly skewed class distributions, the adoption of partial classification techniques could be more appropriate. We evaluate three different complete and partial classification techniques and compare their performance on a ROC convex hull graph. Results on real world data show that, under the circumstances described above, partial classification is indeed a serious competitor for complete classification. Moreover, external validation on holdout data shows that partial classification is able to identify latently dissatisfied customers correctly.

1 Introduction

Latently dissatisfied customers are customers who report overall satisfaction but who possess typical characteristics of customers reporting overall *dissatisfaction*. In this sense, latently dissatisfied customers constitute an important - but hidden - group that should not be ignored by the management. Indeed, because of their possession of characteristics of overall dissatisfied customers, latently dissatisfied customers have a high probability of becoming overall dissatisfied in the near future and, as a result, they are potential defectors. Therefore, we argue that the identification of latently dissatisfied customers may act as an early warning signal, providing the opportunity to correct a problem before customers decide to defect.

There are mainly two methodological approaches to solve this problem. The first one entails the construction of a classical *complete* classification model (such as decision trees) which has the objective of discriminating between overall satisfaction (negative class) and dissatisfaction (positive class). In this setting, latently dissatisfied instances are considered as false positive (FP) instances, i.e. instances reporting overall satisfaction but who are missclassified by the model as overall dissatisfied. The second approach is based on the construction of a *partial*

classification model (such as an association rules ruleset). The motivation is that previous researchers [3] have argued that, under specific circumstances, the use of classical classification models may be inappropriate and partial classification systems should be used instead. More specifically, and especially relevant in our study, the presence of a very skewed class distribution and, at the same time, the intention to concentrate learning on the low-frequency class (overall dissatisfied customers) advocates the use of a partial classification technique.

The paper is organised as follows. Firstly, we will elaborate on the different methodological approaches to the problem of identifying latently dissatisfied customers. In the second part, an empirical comparison of different techniques on real-world data will be carried out. The objective is to make a comparison in terms of a common performance criterion, such as the ROC convex hull graph [10]. In addition, validation will be carried out on separate testing data. The final section will be reserved for conclusions.

2 Alternative Methodological Approaches

2.1 Approach 1: Complete Classification

The complete classification approach assumes that a classification model can be built that discriminates between overall dissatisfied (positive instances) and overall satisfied (negative instances) customers in the dataset. The term *complete* classification stems from the fact that the model covers *all* instances and *all* classes in the data. Consequently, from the methodological point of view, latently dissatisfied instances can then be defined as false positive (FP) classifications.

In the past, most of the attention in research has been devoted to these kind of classification techniques. In this study, we will concentrate on two well-known complete decision tree classification techniques, i.e. C4.5 [11] and CART [5].

2.2 Approach 2: Partial Classification

The term *partial classification* refers to the discovery of models that show characteristics of the data classes, but may not necessarily cover all classes and all instances of a given class. In fact, the aim of partial classification techniques is to learn rules that are *individually* accurate and, thus not to predict future values, but rather to discover the necessary or most prevalent characteristics of some of the data classes [3]. Especially in domains where the class distributions are very skewed and the user is especially interested in understanding the low-frequency class, partial classification can be preferred over complete classification.

Consequently, in the case of partial classification, latently dissatisfied customers are identified somewhat differently. Firstly, characteristics are generated (in terms of frequently co-occurring attribute-value combinations) that are prevalent within the group of overall dissatisfied customers. Given these frequently co-occurring attribute-value combinations, customers in the other group (overall satisfied) that have similar characteristics, are selected. We call the latter group latently dissatisfied

because customers in this group report overall satisfaction although they possess characteristics that are prevalent to overall dissatisfied customers.

In this paper, we will highlight one specific partial classification technique, i.e. association rules.

2.2.1. Association Rules

Association rules [2] were first introduced as a technique to discover hidden purchase patterns in large sales transaction databases, also known as market basket analysis. In such a context, a typical association rule might look like *beer* \Rightarrow *diapers*, indicating that customers who buy beer also tend to buy diapers with it. Recently, however, other applications of association rules have been put forward [3, 4].

Finding association rules in large databases typically involves two phases. In the first phase, frequent itemsets are discovered, i.e. all combinations of items that are sufficiently supported by the transactions (i.e. exceed some predefined minimum support threshold). In the second phase, frequent itemsets are used to generate association rules that exceed a user-defined confidence threshold. The general idea is that if, say, *ABCD* and *AB* are frequent itemsets, then it can be determined if the rule *AB* \Rightarrow *CD* holds by calculating the ratio $r = \text{support}(ABCD) / \text{support}(AB)$. Detailed information on how to perform each of the two phases can be found in [1, 2].

2.3 Comparison of Both Approaches

Conceptually, the difference between complete and partial classification models can be illustrated as shown in figure 1.

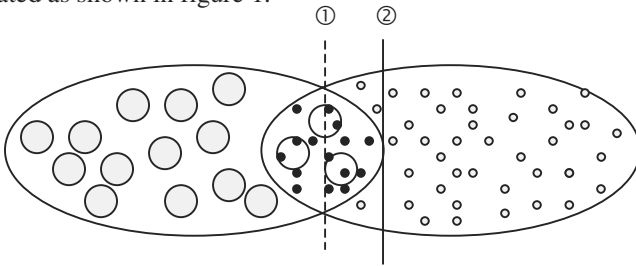


Fig. 1. Manifestly dissatisfied customers are represented as big white dots, whereas small white dots represent manifestly satisfied instances. Latently dissatisfied customers are represented as small black dots

Figure 1 illustrates that it is the objective of *complete* classification to *discriminate* between positive (dissatisfied) and negative (satisfied) instances. In the case of decision trees, this involves the discovery of several (quasi pure) multi-dimensional cubes in the multi-dimensional instance space. However, in the case of *partial* classification, the objective is to find the necessary or most prevalent characteristics of the target class, i.e. to find a description of the target group which is as complete as possible (i.e. which covers as many positive instances as possible).

As pointed out in the introduction, the distinction between complete and partial classification also entails a different way of identifying latently dissatisfied customers. Indeed, in the case of *complete* classification, latently dissatisfied customers are defined as customers who are missclassified as being ‘overall dissatisfied’ which corresponds to instances in the intersection situated on the left-hand side of line 1. In contrast, in the case of *partial* classification, latently dissatisfied customers are those black dots situated on the left-hand side of line 2. This is obvious since the objective of partial classification is to characterise the positive group as completely as possible, causing some of the descriptions (rules) to cover instances of the negative group as well.

Finally, there exists a relationship between line 1 and 2. Namely, in the case of complete classification, increasing the cost of false negative (FN) errors will cause line 1 to shift into the direction of line 2. Indeed, increasing the cost of NF errors will increase the true positive (TP) and false positive (FP) rate, and thus line 1 will shift to the right. In analogy, line 2 will shift into the direction of line 1 by lowering the support (coverage) threshold of the association rules.

3 Empirical Evaluation

3.1 Data

The data being used in this study comes from a large-scale anonymous customer satisfaction survey carried out by a major Belgian bank in 1996. Data were obtained for a random sample of 7264 customers.

3.1.1 Satisfaction Opinions of Different Service Items

Customers were asked about their satisfaction with 16 service items of the bank including questions related to the empathy of the staff (e.g. friendliness), information and communication (e.g. investment advice), and finally the practical organisation of the bank office (e.g. waiting time). Each question (i.e. each attribute in this study) was measured on a 5-point rating scale as illustrated by the following example:

	Never	Seldom	Sometimes	Often	Always	No-opinion
<i>I have to queue for a long time</i>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

One specific question probed for the overall level of (dis)satisfaction of the customer. This question was used to allocate customers into two groups: overall satisfied or overall dissatisfied, and it is the target variable in our study.

3.1.2 Complaints Behaviour

Finally, information was collected with regard to the number and type of complaints that a customer had placed during 1996.

3.2 Data Recoding

Figure 2 on the next page illustrates the distribution of the different attribute values, both for the independent variables and the target attribute (i.e. the question probing for the overall level of satisfaction).

Figure 2 shows that, for all attributes, there exists only a very small tendency to be (manifestly) dissatisfied. This has important implications with respect to the construction of appropriate classification models. Indeed, for many attributes the number of customers responding (manifest) dissatisfaction is too low to guarantee statistically reliable models. For instance, in the case of decision trees, already in the very beginning of the growing process of the tree, there will exist nodes with very few instances producing terminal nodes that will contain very few observations, and as a consequence, their classification label will be very doubtful.

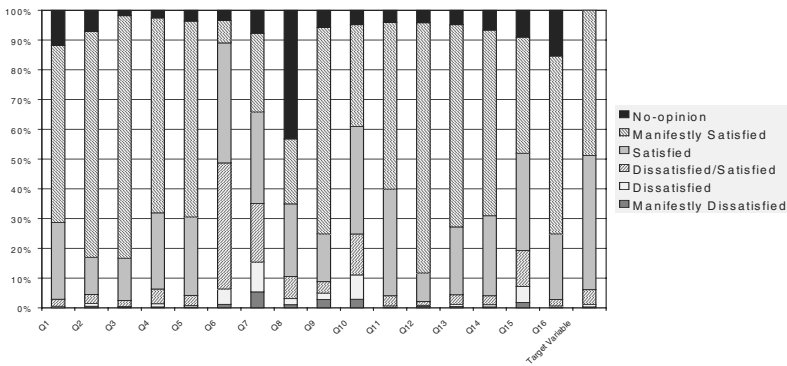


Fig. 2. Distribution of answer patterns on different questions in the survey

To overcome these problems, it is suitable to group certain attribute values to obtain more observations per grouped attribute value, of course with the drawback of losing some detailed information. More specifically, 5 attribute values were recoded into 3 new values, i.e. the answers 'manifestly dissatisfied', 'dissatisfied' and 'dissatisfied/satisfied' were grouped into one new attribute value, with the other attribute values unchanged. Moreover, the target variable was converted into a binary attribute. More specifically, an aggregate value 'overall dissatisfaction' was constructed (still only containing 6.1% of all instances), grouping the attribute values 'manifestly dissatisfied', 'dissatisfied' and 'dissatisfied/satisfied', and an aggregate attribute value 'overall satisfaction' (containing 93.9% of all instances) containing the attribute values 'satisfied' and 'manifestly satisfied'.

3.3 Empirical Evaluation of the Different Techniques

3.3.1 Experimental Design

More specifically, for the complete classification approach, C4.5 was carried out with and without misclassification costs, and with and without grouping of symbolic

values in the tree (C4.5 GSV). The use of misclassification costs is justified because of the skewed class distributions in the data and the grouping of symbolic values in the tree is enabled to obtain a fair comparison with CART which produces *binary* splits. CART was also carried out with and without misclassification costs to adjust the prior probabilities of the target classes.

For the partial classification approach, different association rule models were induced as well. Firstly, we generated all frequent combinations of attribute values for instances of the target class (overall dissatisfied) with a minimum support threshold of 20%. The outcome is the set of all combinations of attribute values that appeared together in the target class with frequency exceeding the minimum support threshold. The support threshold is used to guarantee the discovery of *prevalent* patterns of dissatisfaction. In total, 97 rules for dissatisfaction were obtained.

Secondly, different models of association rules were obtained by modifying the number of rules retained according to some measure of interestingness [4]. This is necessary because the discovered characteristics may also be characteristics of the complete dataset as they represent the necessary but not the sufficient condition for the membership of the positive example set. The following measure was used [4]:

$$Interest_R = S_{Target} - S_{Total} / \max\{S_{Target}, S_{Total}\} \quad (1)$$

where, S_{Target} (resp. S_{Total}) is the support of the rule in the target class (resp. total database). The denominator is introduced to normalise the interestingness between $[-1, +1]$.

3.3.2 The ROC Convex Hull Method

In order to compare the performance of different classification methods on a common basis, we choose the ROC convex hull method [10] because it is robust to imprecise class distributions and misclassification costs. The method decouples classifier performance from specific class and cost distributions, and may be used to specify the subset of methods that are potentially optimal under any cost and class distribution assumptions.

On the ROC convex hull graph (see figure 3), the TP rate is plotted on the Y-axis and the FP rate on the X-axis. One point in the ROC graph (representing one classifier with given parameters) is better than another if it is to the northwest (TP is higher, FP is lower, or both) of the graph.

One can observe different CART, C4.5 and C4.5 GSV models with increasing false negative costs (from cost 1 to 9). This means that for each decision tree technique different models have been induced, each time increasing the penalisation of the FN errors which in turn results in higher TP and FP rates.

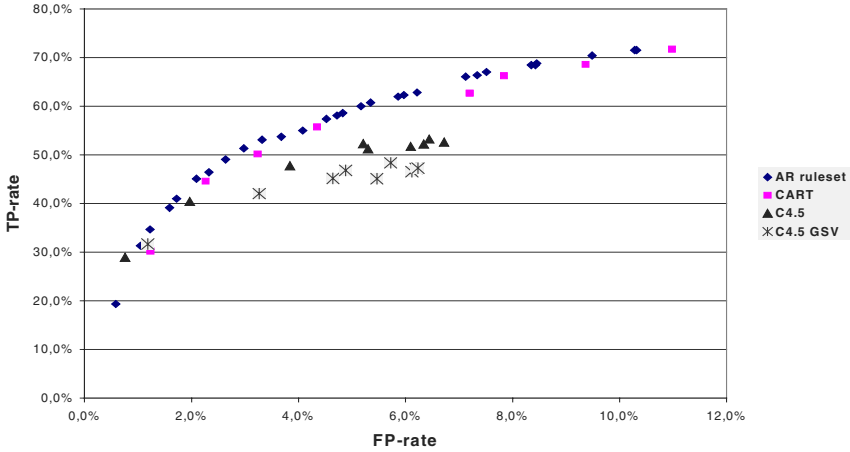


Fig. 3. ROC convex hull graph: performance of different classifiers on separate testing data from 1997

Figure 3 also shows the performance of the different association rule rulesets (AR ruleset). The number of rules in the ruleset determines the performance of the model, i.e. the AR ruleset in the bottom left corner of the graph contains only the single most interesting rule that was generated in section 3.3.1, whereas the top right AR ruleset model contains the 29 most interesting rules as determined by the interestingness measure presented earlier in this paper. The latter ruleset covers more overall dissatisfied instances (higher TP-rate) but because of adding less and less ‘interesting’ rules the FP-rate will increase as well. Decision tree models with higher FN error costs and AR rulesets containing more rules are not plotted on figure 5 since they did not increase the TP-rate significantly but only further increased the FP-rate.

According to [10], only when one classifier dominates another over the entire performance space, it can be declared better. From figure 3, it can be observed that this is the case with the collection of AR ruleset classifiers since for each FP-rate (i.e. for each group size of latently dissatisfied customers), the TP-rate of the AR ruleset is higher than for any of the other types of classifiers (CART, C4.5 and C4.5 GSV) considered in this study. Moreover, for this study, CART is clearly superior to C4.5 with (see C4.5 GSV) and without grouping of symbolic values (see C4.5).

4 External Validation

External validation on holdout data showed that association rulesets are able to capture the idea of latent dissatisfaction since it was discovered that, in the line with marketing theory [6, 7, 9], complaints behaviour and the rate of defection was significantly higher for latently dissatisfied customers than for manifestly satisfied customers. Unfortunately, due to space limitations, we cannot elaborate on this. However, the authors can be contacted for additional information on this issue.

5 Conclusion

In this study, we have compared two different methodological approaches to the identification of latently dissatisfied customers, i.e. complete versus partial classification. More specifically, C4.5, CART and association rule rulesets were evaluated and we compared their performance on a ROC convex hull graph. The reason is that the ROC convex hull graph enables comparison of different types of classification techniques under different misclassification costs and class distributions. We found confirmation for the fact that partial classification would be more appropriate when the data are characterised by very skewed class distributions. Furthermore, external validation results indicated that latently dissatisfied customers put more complaints than manifestly satisfied customers and also have a higher tendency to defect.

References

1. Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A.I.: Fast Discovery of Association Rules. In: *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, chapter 12, (1996) 307-328.
2. Agrawal, R., Imielinski, T., Swami, A.: Mining Association Rules between Sets of Items in Large Databases. In: *Proceedings of the ACM SIGMOD Conference on Management of Data*, Washington D.C., (1993) 207-216.
3. Ali, K., Manganaris, S., Srikant, R.: Partial Classification using Association Rules. In: *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, The AAAI Press, (1997) 115-118.
4. Anand, S.S., Hughes, J.G., Bell, D.A., Patrick, A.R.: Tackling the Cross-sales Problem using Data Mining. In: *Proceedings of the first Pacific-Asia Conference on Knowledge Discovery and Data Mining*, (1997) 331-343.
5. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: *Classification and Regression Trees*, The Wadsworth Statistics/Probability Series, California (1984).
6. Day, R.L.: Modelling choices among alternative responses to dissatisfaction. In: *Advances in Consumer Research* 11, (1984) 496-499.
7. Fornell, C., Wernerfelt, B.: Defensive Marketing strategy by complaint management: a theoretical analysis. In: *Journal of Marketing Research* 24, (1987) 337-346.
8. Greene, W.H.: *Econometric Analysis*. Prentice Hall, chapter 19, (1999).
9. Heskett, J.L., Sasser, W.E., Schlesinger, L.A.: The service profit chain: How leading companies link profit and growth to loyalty, satisfaction and value. In: *The Free Press*, New York, (1997).
10. Provost, F., Fawcett, T.: Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions. In: *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, Newport Beach, California (USA), August 14-17, (1997) 43-46.
11. Quinlan, J.R.: *C4.5: Programs for Machine Learning*. Morgan Kaufmann (1993).