Predicting Multi-class Customer Profiles Based on Transactions: a Case Study in Food Sales

Edward Apeh, Indrė Žliobaitė, Mykola Pechenizkiy and Bogdan Gabrys

Abstract Predicting the class of a customer profile is a key task in marketing, which enables businesses to approach the right customer with the right product at the right time through the right channel to satisfy the customer's evolving needs. However, due to costs, privacy and/or data protection, only the business' owned transactional data is typically available for constructing customer profiles. Predicting the class of customer profiles based on such data is challenging, as the data tends to be very large, heavily sparse and highly skewed. We present a new approach that is designed to efficiently and accurately handle the multi-class classification of customer profiles built using sparse and skewed transactional data. Our approach first bins the customer profiles on the basis of the number of items transacted. The discovered bins are then partitioned and prototypes within each of the discovered bins selected to build the multi-class classifier models. The results obtained from using four multi-class classifiers on real-world transactional data from the food sales domain consistently show the critical numbers of items at which the predictive performance of customer profiles can be substantially improved.

1 Introduction

Advances in data warehousing and management technologies provide great opportunities for businesses to enhance long-term relationships with their customers. The ultimate goal is to effectively utilize the resulting data for transforming relationships

Edward Apeh, Indrė Žliobaitė and Bogdan Gabrys
Smart Technology Research Centre, Bournemouth University, UK
e-mail: (eapeh,izliobaite,bgabrys)@bournemouth.ac.uk

Mykola Pechenizkiy
Eindhoven University of Technology, Eindhoven, the Netherlands
e-mail: m.pechenizkiy@tue.nl

into greater profitability thorough improved customer product targeting, increased customer loyalty and purchase probability [19].

To effectively identify, understand and satisfy the needs of their customers, the businesses need to develop the right interventions for the right customer at the right time through the right channel.

Customer profiles encapsulate the detailed knowledge of the customer behaviour. Predictive models can help to classify customer profiles so as to effectively recognize the status and preference of each individual customer. Such predictive models can be incorporated into the company's market segmentation, customer targeting and channelling decisions with the goal of maximizing the total customer lifetime profit. Transactional data is a valuable resource for building such predictive models. Transactional data can be electronically collected and readily made available for data mining in plenty quantity at minimum extra costs. Transactional data is however, inherently sparse and skewed which adversely affects the performance of ad-hoc classifier models built using transactional data based customer profiles.

We introduce a new approach for customer profile prediction that addresses the problem of multi-class classification of sparse and skewed transactional data. We address the challenge of transactional data sparsity and skewness by modelling customer profiles which have been locally specialized by first binning them into homogeneous groups, instead of building a global model of all diverse customer profiles. Prototypes of customer profiles are then extracted from the discovered bins and multi-class classifier models are built using those prototypes. The learned models can then be used to predict the class of customer profiles (e.g. restaurants, schools, supermarkets, etc.) based on their purchases. The approach is validated on the case study encompassing a food retail and food service company operating in the Dutch food and beverages market.

This study presents two major contributions. First, we propose a new algorithm for predicting the class of customer profiles based on their transactions that can handle sparse multi-class datasets, as exhibited by real life applications. Second, we present an extensive case study in food sales domain illustrating a set of practical challenges in customer profile prediction and validating the proposed approach. The paper is organized as follows. Section 2 discusses background and related work. Section 3 presents the proposed approach. Section 4 presents our experimental analysis using a real-world case study. Section 5 summarizes the findings and implications to a customer-centric business.

2 Background and Related Work

2.1 Customer Profiles

A customer profile is a description of customers using available information, which help in understanding their background and behaviour. Well developed customer profiles are essential in market analysis as they aid businesses in saving time and money by highlighting the real potential customers whose needs are to be met rather than concentrating on too wide a range of individuals.

Customer profiles can be *factual* or *behavioural*. A factual customer profile consists of a set of characteristics (e.g. demographic information such as name, gender, birth date) while a behavioural customer profile consists of what the customer is actually doing and is usually derived from transactional data [1].

Behavioural customer profiles can be much stronger predictors of the future actions of a customer as they encapsulate the changing behaviour of the customer more than demographically based customer profiles which are more or less static. Furthermore, the information that make up demographically based customer profiles are expensive to acquire while the information for behavioural customer profiles can be gleaned each time a customer makes a purchase.

Formally, a *behavioural profile* is defined as follows. Given, a transactional dataset $\mathbf{T}_{N\times d}$, containing N transactions categorized into d product items of $\mathbf{M}$ customers, we define a set of customer profiles $P_{M\times d}$ as $p_{i,j} = \sum_{k=i_1}^{n_i} t_{k,j}$, $\forall_{i=1,\dots,M} \forall_{j=1,\dots,d}$, where $k \in \{i_1, i_2, \dots, i_{n_i}\}$ is a set of indexes referring to the n transactions of the i-th customer profile in the set of transactions $\mathbf{T}$. In other words, we define a behavioural customer profile as a vector containing the individual sum of the d product items transacted by a customer over a period of time.

2.2 Challenges in Mining Transactional Data

Transactional data are time-stamped data collected over time at no particular frequency. Examples of transactional data may include: point of sales (POS) data, retail data, inventory data, call centre data, trading data.

Transactional data tend to be inherently skewed and sparse, due mainly to the underlying process from which they are generated which provide relatively little information per available attribute. For example, in retail, customers usually purchase only a small number of products out of thousands of products in the retailers inventory. Such a transaction when represented in an attribute-vector representation has many zeroes and only a few informative numbers, (i.e. the data is sparse). In our case study the Sligro data is inherently very sparse with sparsity factor[1] 0.3%.

In addition, transactional datasets are typically very large in volume and dimensionality. Although processing power has continued to increase, predictive modelling on the entire dataset can be prohibitive in terms of computational time and costs. Conventional data reduction techniques may not work well [12], because the data is sparse. There is also a high risk of information loss, as different rows may have very different sparsity factors leading to each sampled data instance conveying little or no information for accurate inference. Typically data is concentrated around a point where a vast majority of customers buy only few items over long periods of

[1] The faction of informative (non-zero) elements to all the elements in the data

time. Thus, very little information is available for distinguishing between customer profiles, which makes profile prediction very challenging.

2.3 Related work

Over the last decade and beyond data mining has been widely used for user profiling. In this context three research streams can be distinguished: (i) recommender approaches (personalized recommendations) given a customer aim to predict what she/he will buy (e.g. [1]), (ii) analytical approaches (constructing profiles) that given customers and their categories analyze which profiles buy what (e.g. [15]), (iii) predictive approaches (predict profiles) that given a customer aim to predict who (which profile) she/he is (e.g. [3]).

From technical perspective the proposed binning approach relates to contextual learning (e.g. [2, 22]), where specialized predictive models are used in making predictions based on the current context during operation. Typically contexts are determined arbitrary using domain knowledge. Our approach combines the domain knowledge and computational analysis for finding good bins.

Profile prediction from transactional data closely resembles user modelling for textual data [23]. A customer may have many transactions, likewise a user may read many documents. Items in transactional data resemble words in textual data. In this study we take a domain driven approach expecting the predictive power to be concentrated in purchase quantities. We reserve exploration of alternative feature representations from the textual data analysis domain for our future work.

3 Predicting Customer Profiles from Transactional Data

Our approach builds upon the algorithm proposed in [3]. We introduce two principal extensions to be able to handle realistic sparse multi-class data:

1. the algorithm is redesigned from binary to a multi-class setting that is required by realistic application tasks;
2. we introduce k-means prototyping of customer profiles to be able to efficiently build multi-class classifier models using large sparse datasets.

This section first presents the setting and the new approach for customer profile prediction and then discusses specific algorithmic aspects and design choices of the proposed approach in more detail.

3.1 Problem Setting

Formally, given a set of M customer profiles $P = \{\mathbf{p_1}, \mathbf{p_2}, \ldots, \mathbf{p_M}\}$, with each customer profile $\mathbf{p_i}$ having its aggregated d-dimensional transaction as defined in Section 2.1, our goal is to efficiently build multi-class classifier models that *accurately* assign the i-th customer $\mathbf{p_i}$ to the j-th class label $y_j \in \omega_A, \ldots, \omega_K$. We assess the predictive accuracy as the area under the Receiver Operating Characteristic (ROC) curve (AUC) for classifying unseen customers.

3.2 The Proposed Approach for Predicting Customer Profiles

The intuition behind the new approach for predicting customer profiles from large sparse transactional data is as follows. Since customer profiles may vary in magnitude and normalizing the profiles may occlude some predictive features, we would like to partition our customers into homogeneous groups first. Thus, we bin the customer profiles based on the number of items purchased and select the most typical instances as prototypes within each bin to form our training sets. Then we train multi-class classifiers on these prototypes. Algorithm 1 formally describes the procedure for training the predictors.

Algorithm 1: Train predictors

```
Input: Training set $P$
Output: Predictors $C_b$, cluster centers $\mu_k^b$
Initialize: $S = 0.6$
Bin $P$ into $B$ bins based on the number of items per transaction.;
for $b \leftarrow 1$ to $B$ do
    Cluster into $K_{Classes}$ groups using K-means, where $K_{Classes}$ is the number of classes in $T$;
    Record cluster centers $\mu_k^b$;
    for each instance $x_i$ in bin $b$ do
        for each group $k \in K_{classes}$ do
            Compute the Silhouette Statistics $sw_i^k$;
            if $sw_i^k > S$ then
                include $x_i$ into the set of prototypes $P_{kb}^\star$
    Train a predictor $C_b$ on instances in $P_{kb}^\star$;
```

The class of new customer profiles are predicted as follows. First we determine the closest bin of the new customer profile, based on the number of transactions in the historical data. Then we use the classifier trained within that bin to predict the class of the new customer profile as formally described in Algorithm 2.

In the next sections we give more details and justification of the design choices for the proposed approach.

Algorithm 2: Predict class of new customer profile

```
Input: new instance $x_i$, predictors $C_b$, cluster centers $\mu_k^b$
Output: predicted class $k_i$
Assign $x_i$ to the closest bin $B_p$ : $p = \arg\min \left\| x - \mu_k^b \right\|$;
classify $x_i$ using the predictor $C_b$;
```

3.3 Data Binning

Data binning is an unsupervised discretization technique in which the data is grouped into either *Equal Interval Width* or *Equal Frequency Intervals*.

The equal-width data binning algorithm work by determining the minimum and maximum values of the attribute of interest and then divides the range into a user-defined number of equal width bin intervals. This approach to data binning is however vulnerable to outliers that may drastically skew the range [6].

The equal-frequency data binning algorithm, on the other hand, determines the minimum and maximum values of the attribute of interest, sorts all values in ascending order, and divides the range into a user-defined number of intervals so that every interval contains the same number of sorted values.

Irrespective of the grouping method, a typical data binning process broadly consists of four steps:

1. sorting the continuous values of the feature to be binned,
2. evaluating a cut-point for splitting or adjacent intervals for merging,
3. splitting or merging intervals of continuous value according to the chosen criterion.

One key parameter of concern in the data binning process is determining the best "cut-point" to split a range of continuous values or the best pair of adjacent intervals to merge. Entropy based-and/or-statistical based evaluation function have been used to determine an appropriate "cut-point" with varying results [17].

In order to overcome the problem of skewness that is inherent in sparse transactional data as well as to ensure sufficient statistical power for inference, it is particularly important that the cut-point for splitting the range of the number of items bought be such that each bin contains a proportional representation of the number of customer profiles for each class.

To meet this requirement, our approach uses the equal-frequency data binning algorithm outlined below to partition/group customer transactions by the number of items purchased.

Given a set of M customer profiles,

$$P = [\mathbf{p_1}, \mathbf{p_2}, \ldots, \mathbf{p}_M]$$

with each customer profile $\mathbf{p_i}$ having its aggregated d-dimensional transaction as defined in Section 2.1, re-ordered to obtain:

$$\widehat{\mathbf{P}} = \begin{bmatrix} \widehat{\mathbf{p_1}} \\ \vdots \\ \widehat{\mathbf{p_M}} \end{bmatrix}, \text{ where } \widehat{\mathbf{p_i}} = \left[\widehat{p}_{i,1}, \ldots, \widehat{p}_{i,d}\right]$$

based on the sum total of the number of items transacted by each of the customers.

The corresponding vector $\widehat{\mathbf{s}}$, consisting of the total number of items bought by each of the M customers is:

$$\widehat{\mathbf{s}} = \begin{bmatrix} \sum_{i=1}^{d} \widehat{p}_{1,i} \\ \vdots \\ \sum_{i=1}^{d} \widehat{p}_{M,i} \end{bmatrix} = \begin{bmatrix} \widehat{s}_1 \\ \vdots \\ \widehat{s}_M \end{bmatrix}, \text{ where } \widehat{s}_1 \leq \widehat{s}_2 \leq \ldots \leq \widehat{s}_M$$

Given that $\widehat{\mathbf{P}}$ and $\widehat{\mathbf{s}}$ are sorted in ascending order, the bins can be easily determined. That is, for a given bin size of Q (in our case $Q = 40000$ instances) there will be $\lceil M/Q$ bins.

The training sets in the respective *b-1* bins will be:

$$B_i = \begin{bmatrix} \widehat{\mathbf{p_{1+(i-1)Q}}} \\ \vdots \\ \widehat{\mathbf{p_{iQ}}} \end{bmatrix}, \text{ for } i = 1, \ldots, b-1$$

with the minimum and maximum number of items transacted within each bin given by:

$$\left[\widehat{s}_{1+(i-1)Q}\ \widehat{s}_{iQ}\right]$$

For the (b-th) bin, the number of items can be smaller than Q; and the respective training set and range of total items bought is given by:

$$B_i = \begin{bmatrix} \widehat{\mathbf{p_{1+(i-1)Q}}} \\ \vdots \\ \widehat{\mathbf{p_M}} \end{bmatrix}, \text{ for } i = b$$

and $\left[\widehat{s}_{1+(i-1)Q}\ \widehat{s}_M\right]$.

3.4 Prototype Selection

The binning process whilst abating the problem of skewness in transactional data can result in transactional data groups whose sparsity makes sampling them for classifier modelling unwieldy with a resultant poor performance of the classifier model. An alternate approach to sampling the data in each bin is to carefully select prototypes that most represent each bin.

Many methods have been developed for prototype selection. Some of them are aimed at minimizing space and time needed for the classification of a dataset; while others attempt to improve accuracy. Typical examples of the former include Edited Nearest Neighbour (ENN) [24], Multi-edit [11], Relative Neighbourhood Graph Edition (RNGE) [20], etc.; while the Decremental Reduction Optimization Procedure Family (DROP3) [25], Prototype Selection by Relative Certainty Gain (PSRCG) [18] and Model Class Selection (MoCS) [5] have been proposed as prototype selection for accuracy improvement.

Although different researchers have addressed the issue of prototype selection, there is no research that suggests an automatic procedure for instance selection, which can be employed for any given classification algorithm and in a computationally efficient way, for sparse transactional data. This paper presents an algorithm for prototype selection, which exploits the K-means clustering algorithm. It is aimed at reducing the error rate compared to that obtained to a simple sampling of the transactional data. The proposed approach in this paper is analogous to fuzzy clustering based approach proposed in [4] in which the centroids are selected as prototypes.

We choose the basic K-means for prototype selection that uses the K-means algorithm for partitioning the data within each bin. Then we use the silhouette statistic to select the prototypes, i.e. the instances that are the closest to the centre (measured by the average silhouette width) of the partitioned bin.

First, the K-means algorithm is used to partition the transactional data in each bin into k groups such that the within-group sum-of-squares is minimized. The K-means algorithm works by defining the within-bin scatter matrix given by:

$$S_W = \frac{1}{n} \sum_{j=1}^{g} \sum_{i=1}^{n} I_{ij} \left(X_i - \overline{X_j} \right) \left(X_i - \overline{X_j} \right)^T , \tag{1}$$

where I_{ij} is one if X_i belongs to group j and zero otherwise, and g is the number of groups. The criterion that is minimized by the k-means algorithm is given by the sum of the diagonal elements of S_W, i.e., the trace of the matrix, as follows

$$Tr(S_W) = \sum S_{W_{ii}} \tag{2}$$

Minimizing the trace, is equivalent to minimizing the total within-group sum of squares about the group means [10].

In order to proceed with the decomposition of the unlabelled data, Kmeans requires the number of subsets, or in our case, groups, existing in the data. The Kmeans algorithm requires this parameter as input, and is affected by its value.

Various heuristics attempt to find an optimal number of groups most of them refer to intercluster distance or intracluster similarity. Nevertheless, in this case, as we know the actual class of each instance, we use the number of classes in the transactional data and employ silhouette statistic to determine and select the instances closes to the centre of each class as prototypes for classification.

[16] present the silhouette statistic as a way of estimating the number of groups in a data set. Given observation i, denote the average dissimilarity to all other points in its own cluster as a_i. For any other cluster c, let $\overline{d}\left(i,c\right)$ represent the average dissimilarity of i to all objects in cluster c. Finally, let b_i denote the minimum of these average dissimilarities. The silhouette width for the i-th observation is

$$sw_i = \frac{(b_i - a_i)}{max\left(a_i, b_i\right)} \tag{3}$$

We can find the average silhouette width by averaging sw_i over all observations:

$$\overline{sw} = \frac{1}{n}\sum_{i=1}^{n} sw_i \tag{4}$$

Observations with a large silhouette width are well clustered, but those with small values tend to be ones that are scattered between clusters. The silhouette width sw_i in Equation 3 ranges from -1 to 1. If an observation has a value close to 1, then the data point is closer to its own cluster than a neighbouring one. If it has a silhouette width close to -1, then it is not very well-clustered. A silhouette width close to zero indicates that the observation could just as well belong to its current cluster or one that is near to it. We use the silhouette statistics [16] of the customer profiles in each bin to select prototypes that are most representative of a category to train the classifier models.

3.5 Solution for Multi-class Classification

Multi-class classification involves assigning one of many class labels to an input instance. Formally, given a training dataset of the form (x_i, y_i), where $x_i \in R_n$ is the ith example and $y_i \in \omega_A, \ldots, \omega_K$ is the ith class label, multi-class classification algorithms aim to learn a model H such that $H\left(x_i\right) = y_i$ for new unseen instances.

Classifiers such as k-nearest neighbours and multi-layered perceptrons can directly deal with multi-class problems. However, for complex classification problems involving a large number of classes, it has been often observed in [21], that obtaining a classifier that discriminates between two classes outperforms the one that simultaneously distinguishes among all classes. The proposed approach uses the one-vs-all [21] method with error correcting encoding [8].

4 Case Study in Food Sales Domain

To validate and support the proposed approach this section presents an experimental case study using a real-world transactional data from the food sales domain. We first introduce the case study, next we outline methodology and goals of this experimental analysis and then discuss the results and implications.

4.1 SLIGRO Data

The data was provided by Sligro Food Group N.V., a food retail and food service company operating in the Dutch food and beverages market. The provided data consist of 408625 aggregated Sligro customer transactions collected over three consecutive years. Each aggregated customer transaction record contains information about the customer number, the item number, the number of items purchased and the customer category as stipulated by Sligro.

In total 148601 products were transacted by 65 customer categories. 148 top selling products were used in this study. For this experimental analysis customer profiles with over 3000 transactions were selected. Figure 1 shows the number of distinct top selling items purchased per customer transaction. We can see that the majority of transactions (note the logarithmic scale) are concentrated around purchasing a few items.

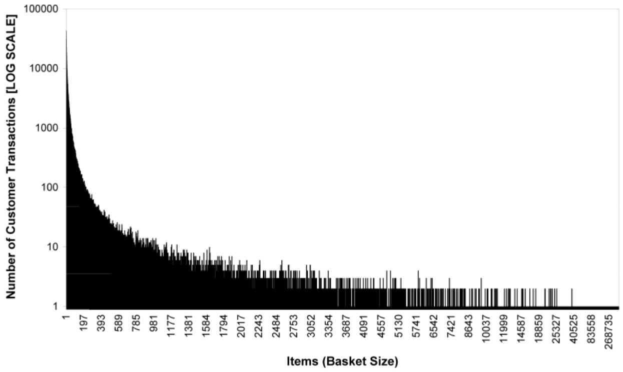

Fig. 1: The distribution of Items per Transaction (Basket size).

4.2 Experiments Protocol

The main goal of our experiments was to validate the proposed approach on a real-world case study. More specifically, we were interested in determining the effect of the number of items bought by each category on the classification performance of four multi-class classifiers. We compared the performance of our approach with the random sampling baseline, which randomly sub-samples customer profiles for classification.

Experimental comparison were performed on four base classifiers (Logistic regression, Decision Tree (J48), Naive Bayes and Support Vector Machine) in WEKA's [13] built-in OVA and ECOC. For each of the selected prototypes 10-fold cross-validation was repeated 10 times.

4.3 Evaluation using the AUC score

The performance of predictive data mining algorithms is typically evaluated using predictive accuracy in which each instance is classified into the class which has the largest estimated probability of class membership over all classes. However, this is not appropriate when the data is imbalanced [7], as the large difference in representation between the classes can lead to a bias in which even a simple default strategy of guessing would give a high predictive accuracy to the majority class.

The area under the Receiver Operating Characteristic (ROC) curve (AUC) is a standard technique for summarizing classifier performance over a range of trade-offs between true positive and false positive error measures [9]. It measures whether the estimated probability of an instance belonging to class c_i is larger than the estimated probability of belonging to class c_j. The AUC is used in this work for comparisons because it is independent of any threshold used by the learning algorithm. It is not influenced by decision biases and prior probabilities, and it places the performance of diverse systems on a common, easily interpreted scale [14].

To compare classifier performance on the entire multi-class transactional dataset, we use the weighted average AUC, where each target class c_i is weighted according to its prevalence [14]:

$$AUC_{weighted} = \sum_{\forall c_i \in C} AUC\left(c_i\right) \times p\left(c_i\right) \tag{5}$$

Weighting the AUC prevents target classes with smaller instance counts from adversely affecting the results.

Table 1: Identified Data Bins.

Class	Bins				
	1..5	6..20	21..127	128+	Sum
100	551	470	699	1439	3159
190	4443	3957	2677	300	11377
230	9284	7963	6358	1125	24730
300	1454	1765	2697	2676	8592
310	972	1117	2069	3083	7241
331	970	1052	1713	3235	6970
360	898	957	1064	948	3867
380	714	768	1116	1037	3635
390	974	1071	1814	1593	5452
391	789	935	1140	586	3450
590	1088	1075	1334	900	4397
620	891	1104	1457	1189	4641
800	8718	8749	7904	1701	27072
820	1518	1431	1105	134	4188
840	4051	4396	4089	609	13145
890	1443	1513	1364	217	4537
900	2941	1928	1420	206	6495
Total	41699	40251	40020	20978	142948

Table 2: Selected Prototypes.

Class	Bins				
	1..5	6..20	21..127	128+	Sum
100	23	17	12	115	167
190	105	185	47	31	368
230	246	381	94	112	833
300	80	72	82	268	502
310	55	36	7	305	403
331	39	39	23	320	421
360	30	43	12	93	178
380	24	22	13	103	162
390	36	42	21	158	257
391	41	42	16	59	158
590	42	56	25	90	213
620	18	60	21	109	208
800	287	328	101	171	887
820	38	51	11	14	114
840	121	199	43	61	424
890	25	69	21	22	137
900	72	68	17	21	178
Total	1282	1710	566	2052	5610

4.4 Data Analysis

Following the proposed approach outlined in Section 3, the customer profiles were first binned, using the equal-frequency binning algorithm described in Section 3.3, to obtain the homogeneous groups for each of the categories (i.e. classes) as shown in Table 1. It can be seen that employing the equal-frequency binning algorithm enables the proportions of the categories (i.e. classes) to be maintained across the bins. The K-means prototype selecting algorithm as described in Section 3.4 was then applied to the discovered bins to obtain the prototypes for each class in each bin as shown in Table 2.

4.5 Analysis of the Predictive Performance

Figure 2 shows the plots of the predictive performance results. It can be seen from all four plots that there is a critical number of items purchased $o^\star$ at which the overall AUC classification performance is higher than that obtained from the baseline approach of random sampling customer profiles for classification.

Identifying the critical point $o^\star$ validates the contribution to predicting multi-class customer profiles based on transactional data, in that, not only does it help in overcoming the challenge of transactional data sparseness and skewness but it also enables practitioners to build specialized classifier models that can more accurately classify customers based on the number of items purchased.

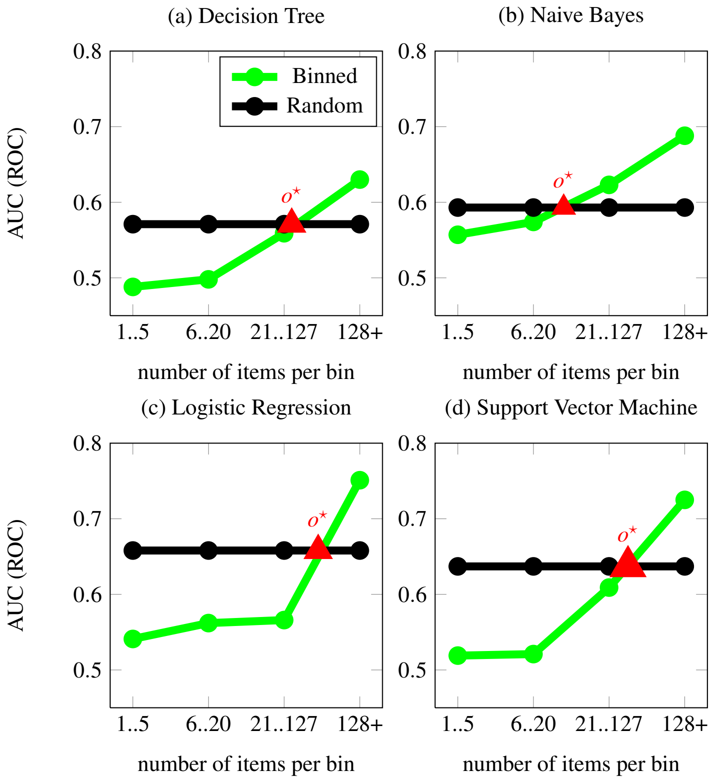

Fig. 2: Performance on the selected prototypes of customer profiles.

Customer profiles based on transactional data with their number of items purchased similar to that at the critical point $o^\star$ can be more confidently distinguished by the specialized multi-class classifier, than would be by the global multi-class classifier modelled on all or on a sample of the diverse customer profiles. The performance of the global multi-class classifier tends to be adversely affected by the inherent dominance of the large number of customers with fewer number of items purchased. Finding the critical point $o^\star$ at which customers can be more accurately classified enables meaningful analysis to be undertaken that can lead to a better customer relationship management.

4.6 Analysis from the Business Perspective

From a business perspective, the customers with profiles whose classification fall above the critical point can be prime candidates for direct interactive/one-to-one marketing campaigns while customers whose profiles fall below the critical point can be candidates for general market campaigns. In addition, the differences in classification performance on individual categories across the bins provide insight that can be valuable for developing better relationship with the customers.

As an illustration, consider the top 10 items for the category codes 100 (small supermarket) and 310 (cafeteria) across the four bins in Tables 3 and 4 show that the highlighted product codes[2] have a strong influence on the classification of the customer profiles based on transactional data in that category.

These products could be included in product promotions and customer targeting programmes as incentives for product growth and customer retention. The decoded examples verify the profiles, as the products intuitively match the types of business (e.g. small supermarkets/ corner shops buy a lot of cigarettes, cafeterias/bars buy Coca-Cola).

Table 3: Top 10 Products within the *Small supermarket* category.

Bin1		Bin2		Bin3		Bin4	
Product Code	Total	Product Code	Total	Product Code	Total	Product Code	Total
93456	42	190855	56	190855	193	190855	1710
184532	22	93456	42	882627	180	882627	1657
81491	11	936251	36	936251	156	190499	765
663523	10	24224	12	93456	151	190960	696
637263	2	652378	12	900360	42	235346	677
284099	1	591881	9	397805	8	192653	618
882627	1	64787	8	736177	6	900360	601
190855	0	637263	3	226054	6	438554	557
432257	0	432257	2	433245	4	432257	537
900360	0	226054	2	81491	4	81491	500

Table 4: Top 10 products of the *Cafeteria* category.

Bin1		Bin2		Bin3		Bin4	
Product Code	Total	Product Code	Total	Product Code	Total	Product Code	Total
184532	111	184532	119	297296	146	192653	5062
93456	60	93456	70	93456	92	184532	4782
81491	8	64787	65	882627	63	190902	3689
394548	2	652378	36	936251	60	882627	2329
637263	1	190855	26	24224	54	265943	2130
192653	1	936251	24	736177	13	432257	1868
269117	1	591881	20	591881	11	255710	1578
476997	1	87353	9	282241	8	516140	1569
736177	1	432257	4	900360	7	231708	1483
882627	0	882627	3	192653	6	401963	1450

[2] 184532 drink, 192653 Coca-Cola, 882627 bread, 81491 bread, 882627 cigarettes, 190855 beer, 432257 energy drink, 900360 cigarettes

5 Conclusion

We presented an investigation into the classification of multi-class customer profiles using real-world transactional data in the food sales domain. We proposed and validated a new approach for predicting customer profiling that can deal with sparse realistic transactional data using the ‘divide-and-conquer’ principle. The proposed approach first partitions data into homogeneous bins, then crystallizes each bin by extracting the most representative prototypes to be used for training the predictive models.

The experimental case study validates the approach on a difficult real world problem. The predictive performance is consistent across different base classifiers. The overall accuracy improves with the number of items purchased. The analysis demonstrates that it is possible to find a critical number of items to be purchased to ensure accurate classification. Knowing this point allows for the filtering of customers and for focused marketing activities to be undertaken on the ones where better predictive accuracy can be expected. Our case study also illustrated that the proposed approach can be used not only for the prediction of new customer profile classes, but also for business analysis, as closer insights can be gleaned from the predicted customer profiles thereby enabling better understanding of the customers of the business.

Acknowledgements We thank Sligro Food Group N.V. for providing the data and the domain knowledge for the case study.

References

1. Adomavicius, G., Tuzhilin, A.: Using data mining methods to build customer profiles. Computer **34**(2), 74–82 (2001)
2. Adomavicius, G., Tuzhilin, A.: Context-aware recommender systems. In: F. Ricci, L. Rokach, B. Shapira, P.B. Kantor (eds.) Context-Aware Recommender Systems, pp. 217–253. Springer (2011)
3. Apeh, E., Gabrys, B., Schierz, A.: Customer profile classification using transactional data. In: Proceedings of the Third World Congress on Nature and Biologically Inspired Computing (NaBIC2011) (2011)
4. Bezdek, J.C., Kuncheva, L.I.: Nearest prototype classifier designs: An experimental study. International Journal of Intelligent Systems **16**(12), 1445–1473 (2001)
5. Brodley, C.E.: Addressing the selective superiority problem: Automatic algorithm/model class selection. In: Proc 10th Machine Learning Conf., pp. 17–24 (1993)
6. Catlett, J.: On changing continuous attributes into ordered discrete attributes. In: Proceedings of the European working session on learning on Machine learning, pp. 164–178. Springer-Verlag New York, Inc., New York, NY, USA (1991)
7. Chawla, N.: Data Mining for Imbalanced Datasets: An Overview, pp. 853–867. Springer US (2005)
8. Dietterich, T.G., Bakiri, G.: Solving multiclass learning problems via error-correcting output codes. Journal of Artificial Intelligence Research **2**, 263–286 (1995)
9. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification (2nd Edition). Wiley-Interscience (2000)

10. Everitt, B.S., Landau, S., Leese, M., Stahl, D.: Cluster Analysis, 5th edition edn. Hodder Arnold (2011)
11. Ferri, F., Albert, J., Vidal, E.: Considerations about sample-size sensitivity of a family of edited nearest-neighbor rules. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on **29**(5), 667 –672 (1999)
12. Gemulla, R., Lehner, W.: Sampling time-based sliding windows in bounded space. In: Proceedings of the 2008 ACM SIGMOD international conference on Management of data, SIGMOD '08, pp. 379–392. ACM, New York, NY, USA (2008)
13. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. SIGKDD Explor. Newsl. **11**, 10–18 (2009)
14. Hempstalk, K., Frank, E.: Discriminating against new classes: One-class versus multi-class classification. In: Proceedings of the 21st Australasian Joint Conference on Artificial Intelligence: Advances in Artificial Intelligence, AI '08, pp. 325–336. Springer-Verlag, Berlin, Heidelberg (2008)
15. Hsieh, N.C., Chu, K.C.: Enhancing consumer behavior analysis by data mining techniques. International Journal of Information and Management Sciences **20**, 39–53 (2009)
16. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis (Wiley Series in Probability and Statistics). Wiley-Interscience (2005)
17. Kohavi, R., Sahami, M.: Error-based and entropy-based discretization of continuous features. In: Proc. 13th Int'l Joint Conf. Artificial Intelligence, pp. 1022–1027 (1996)
18. Olvera-Lpez, J., Carrasco-Ochoa, J., Martnez-Trinidad, J.: A new fast prototype selection method based on clustering. Pattern Analysis & Applications **13**, 131–141 (2010)
19. Payne, A., Frow, P.: A strategic framework for customer relationship management. Journal of Marketing **69**(4), 167–176 (2005)
20. Sánchez, J.S., Pla, F., Ferri, F.J.: Prototype selection for the nearest neighbour rule through proximity graphs. Pattern Recogn. Lett. **18**(6), 507–513 (1997)
21. Vapnik, V.N.: The nature of statistical learning theory. Springer-Verlag New York, Inc., New York, NY, USA (1995)
22. Žliobaitė, I., Bakker, J., Pechenizkiy, M.: Beating the baseline prediction in food sales: How intelligent an intelligent predictor is? Expert Syst. Appl. **39**(1), 806–815 (2012)
23. Webb, G.I., Pazzani, M.J., Billsus, D.: Machine learning for user modeling. User Modeling and User-Adapted Interaction **11**(1-2), 19–29 (2001)
24. Wilson, D.L.: Asymptotic properties of nearest neighbor rules using edited data. Systems, Man and Cybernetics, IEEE Transactions on **2**(3), 408 –421 (1972)
25. Wilson, D.R., Martinez, T.R.: Instance pruning techniques. In: D. Fisher (ed.) Proc 14th International Conference on Machine Learning, pp. 403–411. Morgan Kaufmann (1997)