Who/where are my new customers?

Salvatore Rinzivillo and Salvatore Ruggieri

Pisa KDD Laboratory
ISTI-CNR & Università di Pisa
rinzivillo@isti.cnr.it ruggieri@di.unipi.it

Abstract. We present a knowledge discovery case study on customer classification having the objective of mining the distinctive characteristics of new customers of a service of tax return. Two general approaches are described. The first one, a *symbolic approach*, is based on extracting and ranking classification rules on the basis of significativeness measures defined on the 4-fold contingency table of a rule. The second one, a *spatial approach*, is based on extracting geographic areas with predominant presence of new customers.

Keywords: classification rules, interestingness measures, spatial classification, spatial visualization.

1 Introduction

The research problems and solutions presented in this paper have been motivated by a case study in the context of fiscal services. The business problem consists of providing distinctive characteristics of new customers in order to plan mass marketing campaigns. We report two knowledge discovery approaches, a symbolic one and a spatial one. The symbolic approach adopts classification rules for unveiling contexts, in terms of itemsets, where specific attribute values differentiate new from old customers. The spatial approach adopts spatial partitioning and classification for unveiling contexts, in terms of geographic areas, where specific neighborhoods differentiate new from old customers. Summarizing, both approaches aim at discovering distinctive characteristics of new customers, either by attributes values or by geographic areas. This paper is organized as follows. Section 2 presents the business context and the definition of new customers. The symbolic approach is discussed in Section 3, while the spatial approach is presented in Section 4. Finally, we summarize the contribution in Section 5.

2 The Business Problem

Filling income tax forms can be a demanding and time-consuming task for everybody, especially when fiscal rules are cumbersome as in the Italian legislation. A service in filling forms or in checking already filled forms is offered by business consultants, by trade union associations, or by social assistance organizations.

They also provide services in other fiscal and social assistance matters concerning wages and pensions, local taxes, household duties, caregivers contracts, and so on. These services are highly qualified and they lead to co-responsibility of the servant in the truthfulness of the filled forms. As such, the services are paid, either by the declarant or by the national government. As a consequence, there is a competitive market nationwide among the various consultants and organizations for the acquisition of new customers and the reduction of customer churn. This market has its own specificities:

- income tax return is due once per year, other services may be more frequent (wage calculation for caregivers is once per month), occasional (house selling) or restricted to certain categories (house ownership is taxed only for non-residents in the house);
- customer transactions are exclusive among the competitors, e.g., a person cannot fill income tax forms twice a year. As a consequence, the market is perfectly partitioned among competitors.

This is a radically different scenario from other well-studied markets, such as retailing, banking and telecommunications, where transactions occur at a fine-grained scale, typically one or more times per week, and where competitors may share customers. In this context, we have conducted a joint research-industrial project in the years 2008-2010 with a leading organization offering income tax return consultancy services at dependant workers and retired persons in Italy nationwide. The Customer Relationship Management (CRM) managers of the organization had knowledge that the percentage of new customers each year was in the range 15%-20%. This high fluctuation of customers seemed to be constant both in time and among competitors. Thus, the managers were routinely interested in the marketing problem of attracting new customers. The media adopted for the advertising include newspapers and magazines ads, TV and radio spots, leaflet distribution and roadside posters. All of them are mass media, i.e., no one-to-one marketing is possible due to the constraint that customers are “seen” only once a year. Therefore, the marketing problem reduces to the following:

Business problem: *which attributes best characterize new customers?*

The answer to this question can drive the design of marketing campaigns along several directions including which media to prefer, where or when to post advertising, which messages to deliver in ads. In the project, we followed a CRISP methodology[1] to arrive at data mining models providing two types of answers, a *symbolic* and a *spatial* one, which will be presented in the next sections. The analyses were conducted on four local branches of the organization over a total of one-hundred branches. The rationale for selection was to consider branches with a market share penetration around the average, that are located at the four corners of Italy, and with branch managers acquainted with statistical analysis. The average number of customers in the selected branches ranges from 14.500 to 21.500 per year. Data were collected for fiscal years from 2004 to 2009. After the (notoriously long and demanding) pre-processing of collected data, the dataset

[1] `http://www.crisp-dm.org`

Personal data: age, gender, civil status, residence, isNewResident, no householders, no childs, wasChildOfCustomer, wasConsortCustomer.
Income: job or pension type, employment sector, total wages, houses owned, lands owned, capital gains.
Expenses: health care, mortgages, tuition payments, caregiver wages, donations, insurances.
Derived data: isCommuting, isRetitedThisYear, expenses/income, tax/income, wages/income
Class related data: newCustomer, consortNewCust, occasional, churn.

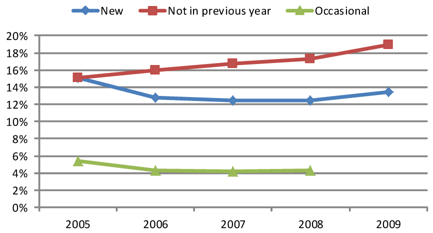

Fig. 1. *Left:* dataset attributes. *Right:* types of new customers.

for the analysis comprised 69 attributes (including binary, nominal, ratio and absolute values) and 375.000 rows. In Fig. 1 *(left)*, we list a few attributes about personal data, income, expenses and some derived attributes.

Before arriving at extracting data mining models, we had to clarify with the marketing managers the notion of "new customer" of the income tax return service, whose definition can be fairly complex. Intuitively, a new customer in a given year is a customer that has not been "seen" in previous year(s). By "seen", one can mean that: (i) she has not been a customer; (ii) she has not been married with a customer; (iii) she has not been a child of a customer. The points (ii, iii) consider the case of in-family persons of a customer that become themselves customers, e.g., because they finish university and start working. Also, the definition of new customer could be restricted not to consider occasional customers, i.e., customers seen only once, or extended to consider as "seen" customers of other services offered by the organization. At the end of the business understanding phase, the definition of new customer at year N has been set to *a customer of the income tax return service in year N that has not been himself a customer of that same service in years prior to N*. It is worth noting, however, that the various facets of the notion can have a considerable impact on data analysis. In Fig. 1 *(right)*, we report for one of the selected branches the percentage of new customers, of customers that were not seen the previous year, and of occasional customers, namely new customers seen only in a year. The number of new customers and customers not seen in previous year coincide for the first year, since the historic data available trace back to 2004 only.

3 A Symbolic Approach

Strictly speaking, the business problem at hand cannot be translated into a classification problem. The business requirement is not to provide a characterization of who are the new customers and who are not. Rather, it asks for attribute values that are distinctive characteristics of new customers, maybe not in the overall dataset but in a subset of it. For instance, the CRM managers were interested in extracting knowledge such as "with reference to the city X, the rate of new customers among young women is much higher than the average", which means

Contingency Table

Classification rule: $c = \mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$

$\mathbf{B}$	$\mathbf{C}$	$\neg\mathbf{C}$	
$\mathbf{A}$	a	b	n_1
$\neg\mathbf{A}$	c	d	n_2
	m_1	m_2	n

$$p_1 = a/n_1 = conf(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C})$$
$$p_2 = c/n_2 = conf(\neg\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C})$$
$$p = m_1/n = conf(\mathbf{B} \rightarrow \mathbf{C})$$

$$elift(c) = \frac{p_1}{p} \qquad slift(c) = \frac{p_1}{p_2}$$

$$olift(c) = \frac{p_1(1-p_2)}{p_2(1-p_1)} = \frac{ad}{bc}$$

$$contrast(c) = \frac{a}{m_1} - \frac{b}{m_2}$$

$$jaccard(c) = \frac{a}{a+b+c}$$

```
ExtractClassificationRules()
N =|$\mathcal{D}$|, $\mathcal{C}$ = { class items }, $\mathcal{L} = \emptyset$
ForEach $k$ s.t. there exist $k$-frequent itemsets
  $\mathcal{F}_k$ = { $k$-frequent itemsets }
  delete from $\mathcal{L}$ unmarked elements
      and unmark all the marked ones
  ForEach $\mathbf{R} \in \mathcal{F}_k$ with $\mathbf{R} \cap \mathcal{C} \neq \emptyset$
    $\mathbf{C} = \mathbf{R} \cap \mathcal{C}$, $\mathbf{X} = \mathbf{R} \setminus \mathbf{C}$
    $a_1 = supp(\mathbf{R})$
    $n_1 = supp(\mathbf{X})$    // $\mathbf{X}$ found in $\mathcal{F}_{k-1}$
    add marked $\mathbf{X} \rightarrow \mathbf{C}$ to $\mathcal{L}$
      with $supp = a_1$ and $cov = n_1$
    ForEach $\mathbf{A} \subseteq \mathbf{X}$ with $0 < |\mathbf{A}| \leq amax$
      $\mathbf{B} = \mathbf{X} \setminus \mathbf{A}$
      $a_2 = supp(\mathbf{B} \rightarrow \mathbf{C}) - a_1$
        // $\mathbf{B} \rightarrow \mathbf{C}$ found in $\mathcal{L}$
      $n_2 = cov(\mathbf{B} \rightarrow \mathbf{C}) - n_1$
      output $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ with
      contingency table $\begin{pmatrix} a_1 N & (n_1 - a_1)N \\ a_2 N & (n_2 - a_2)N \end{pmatrix}$
      mark $\mathbf{B} \rightarrow \mathbf{C}$ in $\mathcal{L}$
    EndForEach
  EndForEach
EndForEach
```

Fig. 2. *Left:* contingency table and measures. *Right:* extraction of classification rules.

that young women are being successfully attracted as new customers in city X. We provide statements of that form by resorting to classification rules. Let us first recall some notation.

Classification Rules. Classification rules from a relation $\mathcal{R}$ are built from a finite set of items $\mathcal{I}$ of the form $a = v$, where a is an attribute of $\mathcal{R}$ and v belongs to the domain of values of a. An itemset $\mathbf{X} \subseteq \mathcal{I}$ is a set of items. As usual in the literature, we write $\mathbf{X}, \mathbf{Y}$ for the itemset $\mathbf{X} \cup \mathbf{Y}$. The (relative) support of $\mathbf{X}$ is the ratio of tuples supporting $\mathbf{X}$ over the total number of tuples in $\mathcal{R}$: $supp(\mathbf{X}) =|\{ t \in \mathcal{R} \mid t \models \mathbf{X} \}|/|\mathcal{R}|$, where $|\ |$is the cardinality operator and $t \models \mathbf{X}$ holds iff for every $a = v$ in $\mathbf{X}$, $t[a] = v$.

An association rule is an expression $\mathbf{X} \rightarrow \mathbf{Y}$, where $\mathbf{X}$ and $\mathbf{Y}$ are itemsets, with $\mathbf{X} \cap \mathbf{Y} = \emptyset$. $\mathbf{X}$ is called the *antecendent* and $\mathbf{Y}$ is called the *consequent* of the association rule. We say that $\mathbf{X} \rightarrow \mathbf{Y}$ is a *classification rule* if $\mathbf{Y}$ is a singleton $a = v$, where a is a specific attribute in $\mathcal{R}$ called the class attribute. The support of $\mathbf{X} \rightarrow \mathbf{Y}$ is: $supp(\mathbf{X} \rightarrow \mathbf{Y}) = supp(\mathbf{X}, \mathbf{Y})$, and its confidence is: $conf(\mathbf{X} \rightarrow \mathbf{Y}) = supp(\mathbf{X}, \mathbf{Y})/supp(\mathbf{X})$. Support and confidence range over $[0, 1]$. We refer the reader to [5] for a survey on mining *frequent* itemsets and association rules, i.e., itemsets and rules with a specified minimum support.

Classification Rules for the Business Problem. Consider a classification rule $c = \mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ where the antecedent is partitioned into two itemsets: $\mathbf{A}$ denotes distinctive characteristics of new customers, and $\mathbf{B}$ denotes a context condition. With reference to our previous example, $\mathbf{A}$ is `age=young, gender=female`, $\mathbf{B}$ is `city=X` and $\mathbf{C}$ is `newCustomer=yes`. How do we let interesting rules emerge from all classification rules? First, the classification rule c must have a support higher than a minimum threshold, in order to cover a significant subset of new customers. Second, significant rules should be highlighted by means of some interestingness measure. Unfortunately, a direct use of the numerous measures in the literature (see e.g., [4]) is not possible, due to the fact that they must be relativized to the partition of the antecedent into **A** and **B**. Consider the contingency table of c shown in Fig. 2 *(left)*. The following measures have been investigated:

Extended lift: $\mathit{elift}(c) = \mathit{conf}(\mathbf{A}, \mathbf{B} \to \mathbf{C}) / \mathit{conf}(\mathbf{B} \to \mathbf{C})$ is an extension of the well-known lift measure of an association rule. It measures how much attribute values **A** increase the chance of being a new customer over the average case for people in a context **B**. $\mathit{elift}(c) = 3$ means that, among people satisfying **B**, the ratio of new customers satisfying **A** is 3 times higher than the average;

Selection lift: $\mathit{slift}(c) = \mathit{conf}(\mathbf{A}, \mathbf{B} \to \mathbf{C}) / \mathit{conf}(\neg\mathbf{A}, \mathbf{B} \to \mathbf{C})$ measures how much attribute values **A** increase the chance of being a new customer compared to people not satisfying **A** in the context **B**;

Odds lift: $\mathit{olift}(c)$ is the ratio between the odds of being a new customer for people satisfying **A** over the odds of being a new customer for people not satisfying **A** in the context **B**. The odds of a rule is the ratio $p/(1-p)$ where p is the rule confidence. In the gambling terminology, the odds 2/3 mean that for every 2 cases an event may occur there are 3 cases the event may not occur;

Contrasting degree: $\mathit{contrast}(c) = \mathit{conf}(\mathbf{B}, \mathbf{C} \to \mathbf{A}) - \mathit{conf}(\mathbf{B}, \neg\mathbf{C} \to \mathbf{A})$ is an extension to generic contexts **B** of the degree of contrasting between new and old customers imposed by the condition **A** (known as *contrast set* [11]). A contrasting degree of 0.4 means that condition **A** is satisfied among new customers by a differential percentage of 40% compared to old customers in the context **B**;

Jaccard coefficient: $\mathit{jaccard}(c)$ measures the asymmetric similarity between the set of people satisfying **A** and the set of new customers, among all people in the context **B**. A Jaccard coefficient of 0.8 means that, with reference to people in **B**, for a person that is a new customer or that satisfies **A** there is 80% of chance to be both a new customer and to satisfy **A**.

The notation for extended, selection and odds lifts is from [7], where they are used in the context of discrimination discovery from historical decision records. Here, the goal is to find contexts **B** where minority groups **A** (e.g., blacks, women, olders) suffered a disproportionate burden in obtaining a benefit **C** (e.g., a loan, a job). With different names, however, those measures have been studied by [2] in the context of medical data analysis. Here, the goal is, given a context **B** (e.g., coronary artery bypass grafting) to find attributes values **A** (e.g., clamp time range) for which the result **C** of a medical treatment (e.g., recovery) had a falling success rate.

Our methodology for analysis consisted then in the following steps: (1) extract frequent classification rules with consequent `newCustomer=yes`; (2) rank rules on the basis of one of the above mentioned measure; (3) eliminate redundant rules; (4) validate significance of the top rules with branch managers.

As for (1), we performed, together with CRM managers, a preliminary discretization of continuous attributes into ranges, to arrive at a dataset of discrete attributes only. For rule extraction, we considered the option of using existing tools such as 4-ft Miner [8], based on the GUHA method. Unfortunately, it

Support = 0.05 No of rules = 5063

	slift	*olift*	*contrast*	*jaccard*
elift	0.827	0.837	0.899	0.343
slift		0.993	0.907	0.450
olift			0.895	0.418
contrast				0.501

Support = 0.1 No of rules = 19

	slift	*olift*	*contrast*	*jaccard*
elift	0.894	0.899	0.991	0.803
slift		0.999	0.932	0.861
olift			0.935	0.822
contrast				0.866

Table 1. Pearson coefficient (r) for different minimum support of classification rules.

does not scale to large datasets and it does not allow to extend the collection of measures it implements. Therefore, we designed the **ExtractClassificationRules()** procedure shown in Fig. 2 *(right)*, as a post-processing step of frequent itemset mining, for which highly optimized tools exist[2]. To limit the exponential growth of extracted rules, the procedure parameter *amax* sets the maximal size of itemsets **A** in a rule. *amax* was set to 2 in the project, due to the fact that short characterizations are actionable (e.g., by designing an ad-hoc marketing campaign), while long ones are not. The procedure outputs classification rules and their contingency tables, so that (2) can be easily implemented by sorting the rules on the basis of the reference measure. As for (3), we tackled rule overlapping due to density of the dataset by a preliminary clustering of items based on their Jaccard distance (formally, the distance between $a_1 = v_1$ and $a_2 = v_2$ is $1 - \mathit{jaccard}(a_1 = v_1 \rightarrow a_2 = v_2)$), and by choosing a representative for each cluster. Notice that the mentioned 4-ft Miner tool offers the possibility to specify a representative of a set of items, but this is a user-specified action, with no automatic support. Finally, (4) was conducted by a domain expert selection of the most interesting rules.

Example 1. Tax return forms can be filled individually or, for married couples that share family properties, jointly. Although the grain of the dataset was at individual level, the following binary attributes are available: `jointFill` to record that the form was filled jointly with the consort; and `consortNewCust` to record that the consort of the declarant is a new customer. The following rule clearly emerged regardless of the measure adopted:

`consortNewCust=yes, jointFill=yes` → `newCustomer=yes`

where **A** is `consortNewCust=yes`, **B** is `jointFill=yes`, and **C** is `newCustomer=yes`. The rule can be interpreted as the fact that "customers that fill joint forms are both new or both old customers", namely, a couple tends not to split over two competitors. For the rule c above, it turns out that, among the customers filling a joint declaration: $\mathit{contrast}(c) = 0.58$, i.e., being the consort a new customer is 58% more frequent in new customers than in old customers; $\mathit{elift}(c) = 6.24$, i.e., new customers are 6 times more frequent among the consorts of new customers than in the average; $\mathit{slift}(c) = 15.2$, i.e., new customers are 15 times more frequent among the consorts of new customers than among the consorts of old customers; $\mathit{olift}(c) = 39.66$, i.e., betting to find new customers among the

[2] `http://fimi.cs.helsinki.fi`

consorts of new customers should be paid 39.66 times less than betting to find new customers among the consorts of old customers; and $jaccard(c) = 0.46$, i.e., new customers and the consorts of new customers share 46% of their members.

An issue at the beginning of the project was concerned with which measure had to be selected for ranking the classification rules extracted. This is a well-known problem in association rule mining [10]. From the literature, we know that the various lift and contrasting measures are related, in the sense that $\mathit{elift}(c) \geq 1$ iff $\mathit{slift}(c) \geq 1$ iff $\mathit{olift}(c) \geq 1$, as shown in [7, Lemma 4.1], and this holds iff $\mathit{contrast}(c) \neq 0$, as shown (for empty contexts) in [11, Section 3.1]. However, this was not enough for drawing a conclusion, and we left the choice of the measure open by computing values for all of them. Table 1 shows the Pearson linear correlation coefficient for extracted rules with minimum support of 10% and 5%. Interestingly, *elift*, $slift$, $olift$ and *contrast* are highly correlated, which means they let emerge the same ranking. This is a stronger conclusion than the previously recalled results. Correlation of $jaccard$ with the other measures is rather high for the minimum support of 10%, but it degrades for the minimum support of 5%.

4 A Spatial Approach

Let us provide a second answer to our business problem by characterizing the distribution of new customers over the territory, with the intent to drive advertising campaigns, e.g., planning leaflet distributions and allocating roadside posters, and to support local business actions, e.g., locating the best place for a new branch office. First, we state an assumption and a pre-processing step.

Partitioning the territory into cells. We assume a spatial grid on the territory. In particular, we adopt a tessellation of the space into statistical sectors[3], which has several advantages: *(1)* the spatial extension of a cell is not too large, usually it comprises a few city blocks in a urban context; *(2)* its fence coincides with the road networks, allowing to plan effective leaflet distribution and roadside poster allocation; *(3)* each cell is associated with publicly available statistical data including the resident population, the grade of education, etc.; *(4)* the cells are organized in a linear hierarchy with levels `[cell, city, province, region, country]`. We have also considered a finer-grained data-driven partition which is based on city blocks. In particular, given the road network of an area, we determine a complete partition of the territory by using the roads as edges for the polygons in the partition. The road partition is obtained by computing a complete noding of the linestrings representing the roads and then polygonizing the resulting rings.

Mapping customers to points. We associate each customer in the dataset under analysis to a spatial location by means of a reverse geocoding function of the customer's address. On the basis of the accuracy of the geocoding function, the mapping operation may produce results with different level of precision. In

[3] Publicly available from the Italian Institute of Statistics `http://www.istat.it`.

our experiments, the mapping was able to determine the location approximated to the street granularity, which is finer than the statistical sector, for almost all the addresses. Outlier points have been removed from the subsequent analyses.

Intuitively, a customer is related to a cell g, if the location x obtained by mapping her address is contained in the cell – in symbols if the spatial predicate $\mathit{contains}(x, g)$ holds. We introduce the new attribute *cell* in the dataset under analysis by setting a spatial item *cell=g* for a customer (a row in the dataset) if $\mathit{contains}(x, g)$ holds for x being the customer's address location. Analogously, attributes *city*, *province*, *region* and *country* are added for the various levels of the spatial hierarchy. With these new attributes, we can now resort to the symbolic approach and extract classification rules $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$, where spatial items may occur both in $\mathbf{A}$ and $\mathbf{B}$. For example the rule:

$$\texttt{gender=female, cell=102} \rightarrow \texttt{newCustomer=yes}$$

concerns whether being a woman is a distinctive characteristic of new customers among the customers living in cell 102. For the rest of this section, however, we will be interested in rules of the form $c = (\vee_i \mathbf{A}_i), \mathbf{B} \rightarrow \mathbf{C}$, where $\vee_i \mathbf{A}_i$ is a disjunction of items $\mathbf{A}_i$ of *spatially adjacent* cells, and $\mathbf{B}$ is the city item at the higher level of the spatial hierarchy. For instance, the following:

$$\texttt{(cell=102} \vee \texttt{cell=103), city=Rome} \rightarrow \texttt{newCustomer=yes}$$

concerns whether the union of spatially adjacent cells (here, cells 102 and 103) is a distinctive characteristics of new customers among those living in the parent cell at the city level (here, the city of Rome). With reference to the *elift* measure, high values of $\mathit{elift}(c)$ identify adjacent geographic areas on the territory $(\vee_i \mathbf{A}_i)$ of a city $(\mathbf{B})$ where new customers occur more frequently than the average of the whole city. Similar interpretations can be given for the other measures from Fig. 2. Let $f()$ be the reference measure from now on. Since $\mathbf{C}$ is fixed (we are interested in `newCustomer=yes`), and $\mathbf{B}$ is fixed as well to the city containing the cells $\mathbf{A}_i$'s, we can write $f(\vee_i \mathbf{A}_i)$ as a shorthand for $f(c)$ and, for a single cell, $f(g)$ instead of $f(\mathit{cell} = g)$. In other words, the measure $f()$ can be extended from (symbolic) classification rules to (spatial) cells. As an example, $\mathit{elift}(g)$ is the ratio of the percentage of new customers in cell g over the percentage of new customers in the city containing g. Let us now depart from the symbolic approach in favor of a spatial approach, for two main reasons. First, visualising cells g over a map on a colored scale on the basis of $f(g)$ makes it easy and immediate for an analyst to locate interesting cells. Second, extracting rules of the form above can be interpreted as the problem of grouping spatially adjacent cells. Let us explain in more details these two issues.

Cell coloring. As a general rule, geo-marketing analysis, such as site assessment and penetration analysis, can be made very effective by a visual analytics approach (see e.g., [3]). The visualization of the spatial location of data enables the analyst to catch instantly the distribution of the different types of customers on the territory. However, the thematic visualization of points may be too chaotic to be fruitfully interpreted. By exploiting the measure $f()$, we associate a scale

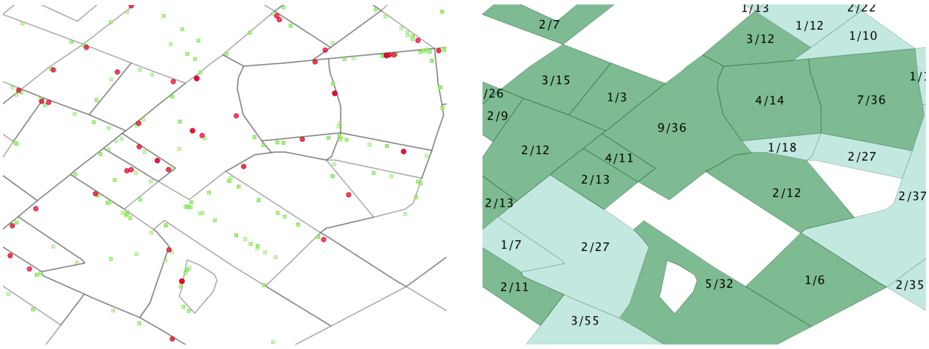

Fig. 3. Example of classification of cells according to a confidence threshold. *Left:* the points are colored according to the attribute `newCustomer` where a red point denotes a new customer and a green one denotes an old customer. *Right:* regions are themed according to minimum threshold confidence of 15%. The threshold is determined by the average percentage of new customers from Fig. 1 *(right)*. Each cell is labeled with the ratio of the number of new customers over the total number of customers in the cell. Cells containing no new customer are omitted.

of colors to a cell g on the basis of $f(g)$. Fig. 3 shows an example: the visualization of points (map on the left) do not let emerge any useful pattern, while the visualization of cells satisfying a minimum confidence threshold (map on the right) enable us to distinguish areas with a high proportion of new customers.

Example 2. The various interestingness measures may have different distributions over cells. For example, consider confidence and *elift* for the example in Fig. 4. Confidence highlights cells with a high percentage of new customers, whereas *elift* highlights cells with a percentage of new customers higher than the average of the whole city. Here, we set a minimum threshold of 15% for confidence and of 2 for *elift*. Cells with a measure value above the minimum threshold are colored with a darker color, and the others with a brighter color. The administrative borders of cities are rendered through thicker lines. The comparison of the two maps shows how the selection of interesting cells may vary with the measure adopted. In particular, the two cells highlighted in red in Fig. 4 *(right)* are “downgraded” from a dark color to a light color when moving from confidence to *elift*. This means that the proportion of new customers in those cells is high (precisely, higher than 15%) as an absolute value, but not that high when compared to the average proportion of the city they belong to.

Cell grouping. It is desirable to have a compact representation of the groups of adjacent cells that satisfy a minimum threshold value *minf* for a given measure $f()$. This allows for providing the business user with a high level description of those cells, for example by enumerating the roads that overlap the fence of the group region or by providing census statistics for (a few) groups rather than for (many) cells. Approaches for spatial rule extraction, such as [1,6,9], are not directly applicable for cell grouping. In fact, since they exploit the *apriori*

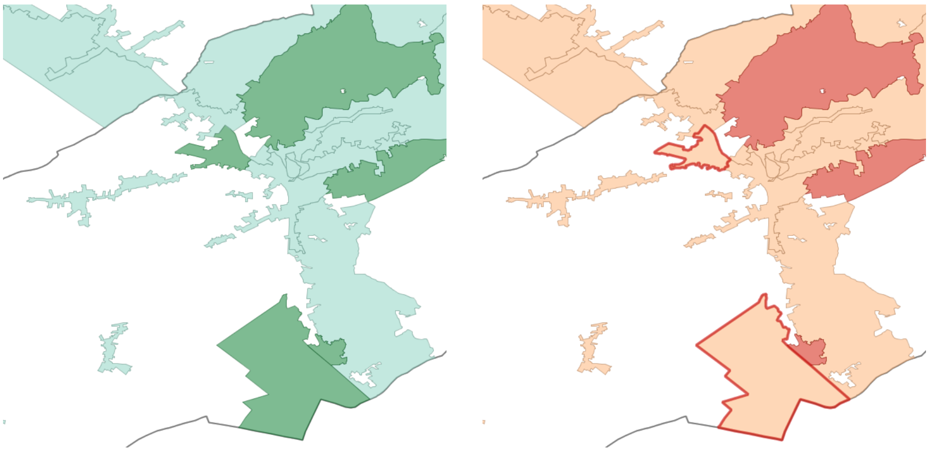

Fig. 4. Sample cell classification by measure treshold. *Left:* confidence. *Right: elift.*

principle, the extracted patterns are limited to conjunctions of spatial predicates – while groups consist of disjunctions of adjacent cells. Also, extracting multilevel spatial classification rules would collapse all cells in a city to a single group, and this is too coarse-grained. We follow an approach that does not assume any *a priori* generalization. Our method consists of the following steps:

(1) join all adjacent cells g with $f(g) \geq minf$. Since for the various measures in Fig. 2 it holds that $f(union(g_1, g_2)) \geq min\{f(g_1), f(g_2)\}$, we have that the resulting groups of merged cells still satisfy the minimum threshold value;

(2) expand groups g_1 found at step *(1)* by connecting adjacent cells g_2 (for which $f(g_2) < minf$) that do not cause the merged area to violate the minimum threshold requirement, i.e., such that $f(union(g_1, g_2)) \geq minf$ still holds.

The approach is implemented in the **CellGrouping** algorithm reported in Fig. 5 together with a sample output. Notice that step (2) is stated as a non-deterministic choice of a pair of adjacent cells g_1 and g_2. In actual implementation, we adopted the heuristics of ordering candidate pairs on the basis of $f(union(g_1, g_2))$. The pair with the highest value, but still not lower than $minf$, is chosen. The algorithm terminates when for all pairs $f(union(g_1, g_2)) < minf$.

5 Conclusions

We have presented two complementary approaches for mining the distinctive characteristics of new customers in terms of their attribute values and geographic location. The two approaches, named *symbolic* and *spatial*, have been investigated in the case study of fiscal services, with the CRM objective of planning mass-marketing advertising campaigns. Nevertheless, we presented the approaches in general terms, both in the algorithms and in the interestingness

```
**CellGrouping**($f$, $minf$)
$\mathcal{G} = \{g | \ f(g) \geq minf\}$
$\mathcal{G}' = \{g | \ f(g) < minf\}$
// step (1): join adjacent cells in $\mathcal{G}$
While exists $g_1$ and $g_2$ in $\mathcal{G}$
  s.t. $touches(g_1, g_2)$
    remove $g_1$ and $g_2$ from $\mathcal{G}$
    add $union(g_1, g_2)$ to $\mathcal{G}$
EndWhile
$\mathcal{G}'' = \mathcal{G} + \mathcal{G}'$
// step (2): join adjacent cells in $\mathcal{G}''$
While exists $g_1$ and $g_2$ in $\mathcal{G}''$
  s.t. $touches(g_1, g_2)$
  and $f(union(g_1, g_2)) \geq minf$
    remove $g_1$ and $g_2$ from $\mathcal{G}''$
    add $union(g_1, g_2)$ to $\mathcal{G}''$
EndWhile
output $\mathcal{G}''$
```

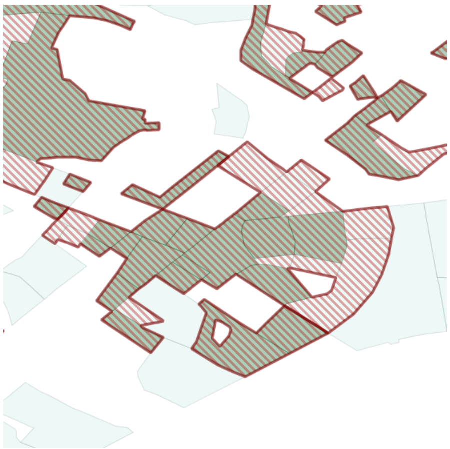

Fig. 5. *Left:* cell grouping algorithm. *Right:* sample output on cells from Fig. 3 *(right)*.

measures adopted. We are confident they can be reused in other customer classification problems, where the distinctive characteristics of a class value (here, new customers) is the target concept to describe either symbolically or spatially.

References

1. A. Appice, M. Ceci, A. Lanza, F. A. Lisi, and D. Malerba. Discovery of spatial association rules in geo-referenced census data: A relational mining approach. *Intelligent Data Analysis*, 7(6):541–566, 2003.
2. P. Eklund, J. Karlsson, J. Rauch, and M. Simunek. On the logic of medical decision support. In *Theory and Applications of Relational Structures as Knowledge Instruments*, volume 4342 of *LNCS*, pages 50–59. Springer, 2006.
3. V. H. Ernst, A. Voss, and F. Berghoff. Visual analytic services for geomarketing in spatial data infrastructures. In *Proc. of GIS 2007*, article 42. ACM, 2007.
4. L. Geng and H. J. Hamilton. Interestingness measures for data mining: A survey. *ACM Computing Surveys*, 38(3), 2006.
5. J. Han, H. Cheng, D. Xin, and X. Yan. Frequent pattern mining: Current status and future directions. *Data Mining and Knowledge Discovery*, 15(1):55–86, 2007.
6. K. Koperski and J. Han. Discovery of spatial association rules in geographic information databases. In *Proc. of SSD 1995*, vol. 951 of *LNCS*, pages 47–66, 1995.
7. D. Pedreschi, S. Ruggieri, and F. Turini. Measuring discrimination in socially-sensitive decision records. In *Proc. of SDM 2009*, pages 581–592. SIAM, 2009.
8. J. Rauch and M. Simunek. 4-ft Miner Procedure. `http://lispminer.vse.cz`, 2011.
9. S. Rinzivillo and F. Turini. Extracting spatial association rules from spatial transactions. In *Proc. of GIS 2005*, pages 79–86. ACM, 2005.
10. P.-N. Tan, V. Kumar, and J. Srivastava. Selecting the right objective measure for association analysis. *Information Systems*, 29(4):293–313, 2004.
11. G. I. Webb, S. M. Butler, and D. A. Newlands. On detecting differences between groups. In *Proc. of KDD 2003*, pages 256–265. ACM, 2003.