Cost-based Quality Measures in Subgroup Discovery

Rob M. Konijn, Wouter Duivesteijn, Marvin Meeng, Arno Knobbe

LIACS, Leiden University, the Netherlands
{konijn,wouterd,meeng,knobbe}@liacs.nl

Abstract. In this paper we consider data where examples are not only labeled in the classical sense (positive or negative), but also have costs associated with them. In this sense, each example has two target attributes, and we aim to find clearly defined subsets of the data where the values of these two targets have an unusual distribution. In other words, we are focusing on a Subgroup Discovery task over somewhat unusual data, and investigate possible quality measures that take into account both the binary as well as the cost target. In defining such quality measures, we aim to produce interpretable valuation of subgroups, such that data analysts can directly value the findings, and relate these to monetary gains or losses. Our work is particularly relevant in the domain of health care fraud detection. In this data, the binary target identifies the patients of a specific medical practitioner under investigation, whereas the cost target specifies how much money is spent on each patient. When looking for clear specifications of differences in claim behavior, we clearly need to take into account both the 'positive' examples (patients of the practitioner) and 'negative' examples (other patients), as well as information about costs of all patients. A typical subgroup will now list a number of treatments, and how the patients of our practitioner differ in both the prevalence of the treatments as well as the associated costs. An additional angle considered in this paper is the recently proposed Local Subgroup Discovery, where subgroups are judged according to the difference with a local reference group, rather than the entire dataset. We show how the cost-based analysis of data specifically fits this local focus.

1 Introduction

This paper is about data that involves a binary label for each example, as well as a cost. The motivation comes from a real-life problem where we are interested in benchmarking (comparing) the claim behavior of medical practitioners. When a patient visits a medical practitioner, the practitioner charges an amount of money, corresponding to the treatment the patient received, to a health insurance company. Several parties are involved in this treatment, each with its own set of knowledge. The patient knows which treatments are performed, but he is unaware of the communication between the practitioner and the insurance company. The insurance company knows which treatments are claimed by the practitioner, but it is unaware of what exactly happened when the patient visited the practitioner's office. The practitioner is the only party that has both sets of information: he knows what treatments he performed with this patient, and he knows what treatments he claimed at the insurance company. Because of this information advantage, a malevolent practitioner is in a unique position that gives leeway to inefficient claim behavior or even fraud.

Detecting fraud on this level is very interesting to the insurance company – much more so than fraud on the level of individual patients – since the commercial implications are substantial. Hence there is a market for a data mining solution to identify unusual claiming patterns that have a substantial economical impact. The problem of identifying interesting patterns in claim behavior is essentially an unsupervised learning problem. We have no claims that are labeled as interesting beforehand. The approach we take is to single out a practitioner and compare his claim behavior with the claim behavior of other practitioners. The data we consider describes patients and practitioners. A single record summarizes the care a patient received during a certain period. We are interested in finding patient groups (patterns), that describe the difference between a single medical practitioner and its peers. In other words, we would like to develop a data mining algorithm, of which the output would be: patients that are in subgroup S occur much more frequent for this medical practitioner, and indicate a difference with other practitioners. The task of identifying such interesting subgroups is known as Subgroup Discovery [5], and also as Emerging Pattern mining [2], and Contrast Set mining [1].

We are interested in patterns that distinguish one practitioner from the others. In order to find such patterns, we need quality measures to describe how 'interesting' a pattern is. We would like the quality measure to capture the distributional difference between one practitioner and the others: the higher the distributional difference, the more interesting a pattern is. Secondly, we are interested in including costs into the quality measure. The main motivation is that in our application, subgroups involving more money are more interesting. Also, a monetary-valued quality value for each subgroup greatly improves the interpretability of a subgroup, because the measure itself has a monetary value. In this paper we will describe how to take costs into account when calculating quality measures. Each patient is 'labeled' by a monetary value – in our application this is the total costs spent on treatments during a specific period – and the quality measures we develop use these monetary values. As a result, the subgroups we find should be easier to interpret by domain experts, since the groups have an associated value in a commodity the experts understand.

2 Preliminaries

Throughout this paper we assume this dataset D with N examples (typically patients). Each row can be seen as a $(h+2)$-dimensional vector of the form $x = \{a_1, .., a_h, t, c\}$. Hence, we can view our dataset as an $N \times (h+2)$ matrix, where each example is stored as a row $x^i \in D$. We call $a^i = \{a^i_1, .., a^i_h\}$ the *attributes* of the i^{th} example x^i. The attributes are taken from an unspecified domain $\mathcal{A}$. The last two elements of each row are the targets. The first target, t, is binary. Its values are set by singling out a medical practitioner. This t-vector then indicates if a patient visited a medical practitioner (a positive example), or not (a negative one). The other target, c, indicates a monetary value. In our application this monetary value indicates the total costs spent on treatments, per year. For other applications c could indicate the profit or the per-customer value. Just like for the attributes, we will refer to the target values of a specific record by superscript: t^i and c^i are the targets of example x^i.

The goal of our approach is to find differences between the singled-out medical practitioner (positive examples) and the rest. Simultaneously, the difference should constitute a considerable amount of money; the more money involved, the better. For this purpose the second target vector c is used. These differences are described by subgroups. A subgroup can be seen as a bag of examples, it can be any subset S of the dataset $S \subseteq D$. We describe how interesting a subgroup is with the use of a quality measure. A quality measure $q : 2^D \to \mathbb{R}$ is a function assigning a numeric value to any subgroup. Generally, the larger the subgroup is, the better (very small subgroups are usually not preferred). Also, the bigger the distributional difference, the better. In our case, because we have two target vectors, this distributional difference can be measured in terms of the binary target vector t (the higher the frequency of $t = 1$ in the subgroup, the better), and the distributional difference can also be measured in the monetary-valued target vector c (the higher the values for c within the subgroup, the better). The quality measure combines these properties of interestingness in a single numeric value.

In traditional Subgroup Discovery, there is only one binary target attribute t. We denote the set of examples for which t is true (the *positives*) by T, and the set of examples for which t is false (the *negatives*) by F. When we consider a particular subgroup S, we denote its complement by $\neg S$. In this setting we denote the true/false positives/negatives in the traditional way: $TP = T \cap S$, $FP = F \cap S$, $FN = T \cap \neg S$, and $TN = F \cap \neg S$. For any subset of examples $X \subseteq D$, we let $\bar{c}_X$ denote the mean cost of the examples in X: $\bar{c}_X = \sum_{x^i \in X} c^i / |X|$, where $|X|$ is the cardinality of the set X.

2.1 The local subgroup discovery task

To deal with locality, in a previous publication we introduced the Local Subgroup Discovery (LSD) task [3]. The idea is to “zoom in” on a part of the data set, and detect interesting subgroups locally. In our application, we can think of the patient population as if they are distributed among different patient groups (for example one group could be patients having a type of cancer). Such a coherent group of patients on which we zoom in is called a *reference group.* LSD is a distance-based approach to find subgroups and reference groups, based on prototypes. A *prototype* can be any point in attribute space $x \in \mathcal{A}$. The *distance-based subgroup* S_σ based on x for parameter $\sigma \in \mathbb{N}$, consists of the σ nearest

Table 1. The *counts* cross table and the *costs* cross table

	T	F
S	TP	FP
$\neg S$	FN	TN

	T	F
S	$\overline{c}_{S\cap T}$	$\overline{c}_{S\cap F}$
$\neg S$	$\overline{c}_{\neg S\cap T}$	$\overline{c}_{\neg S\cap F}$

neighbors of x in D. The *reference group* R_ρ based on the same x for parameter $\rho \in \mathbb{N}$ s.t. $\rho \geq \sigma$, consists of the ρ nearest neighbors of x in D.

The goal of LSD is to find subgroups $S_\sigma \subseteq R_\rho$ for which the target distribution is different from the target distribution in the reference group. The reason for zooming in on a reference group is twofold. On the one hand, this allows us to provide information about the neighborhood of a found subgroup. On the other hand, it accounts for inhomogeneities in the dataset. The idea behind that is that R_ρ forms a region in input space where the target distribution is different from that distribution over the whole dataset. Subgroups that are interesting to report are not these reference groups: they are simply groups of patients sharing a disease that is relatively expensive to treat. The interesting subgroups from a fraud detection point of view, are those subgroups that represent a deviation in target distribution *relative to their peers*: we want to find subgroups $S_\sigma \subseteq R_\rho$ in which the target distribution is different from the distribution in the reference group.

We write $S(x, \sigma, \rho)$ for the subgroup S_σ in a reference group R_ρ, which we call a *reference-based subgroup*. The prototype can be seen as the center of this subgroup, and as the center of the reference group encompassing the subgroup. A quality measure calculated for a reference-based subgroup considers only examples inside the reference group.

3 Quality measures

The quality measures we consider are defined in terms of two cross tables, both depicted in Table 1. The cross table on the left is common in traditional Subgroup Discovery. The cross table on the right is concerned with the mean costs for each of the categories. Our quality measures should satisfy the following criteria:

- in the first cross table, the higher the numbers on the diagonal (TP+TN) are, the more interesting the subgroup is;
- in the second cross table, the higher the mean cost value in the true positive cell $\overline{c}_{S\cap T}$ is, relative to those values in the other cells, the more interesting a subgroup is.

Furthermore, it would be desirable if the value of the quality measure has a direct interpretation in terms of money. In Sections 3.1-3.3 we introduce quality measures satisfying these criteria, but before that we will shortly discuss why a straightforward Subgroup Discovery approach does not suffice.

$$\begin{array}{rl} t = & \{\ +\ ,\ -\ ,\ -\ ,\ -\ ,\ +\ ,\ -\ ,\ +\ ,\ +\ ,\ +\ ,\ -\ ,\ -\ ,\ -\ \} \\ c = & \{\ 1000, 2000, 2000, 1250, 2000, 3000, 200, 200, 200, 200, 200, 200\ \} \\ c * t = & \{\ 1000,\ 0\ ,\ 0\ ,\ 0\ , 2000, \underset{\overset{\uparrow}{\sigma}}{0}\ , 200, 200, 200,\ 0\ ,\ 0\ ,\ 0\ \} \end{array}$$

Fig. 1. A dataset of twelve examples with a subgroup of six examples indicated.

A naive way of dealing with the two targets in our dataset, is to multiply t by c for each observation, and use these new values as one numeric target variable of a traditional Subgroup Discovery run. By taking the difference in means between the subgroup and the mean of the data, the quality measure has a monetary value. Consider the dataset in Figure 1. The first 6 examples (up to σ) belong to the subgroup, and the other examples do not. Computing the difference in means for $c \cdot t$, the value of this naive quality measure would be ${}^{3000}/_{6} - {}^{3600}/_{12} = 200$ (where we are comparing the subgroup mean with the mean of the total $c \cdot t$ column).

The disadvantage of this measure, is that the value of this monetary value of 200 does not have a direct meaningful interpretation: it does not directly relate to the amount of money that is present 'more' in the subgroup, or that could be recovered. But more importantly, the positive quality value for this subgroup is misleading. When we are looking for high average values for our target, this suggests that there is somehow more money involved than expected, but when we take a look at this subgroup, there is not more money involved for the positive examples than for the negative examples. In our application, suppose the reference group (12 examples) would indicate diabetes patients. The subgroup indicates patients receiving expensive diabetes treatments, and the rest of the reference group indicates patients receiving inexpensive diabetes treatments. A low number of true positives (lower than expected), would mean that there are less diabetes patients present at this practitioner than at other practitioners. Also the practitioner claims less money for diabetes patients than other practitioners do (the costs of the true positive patients is less than the false positives). Since overall the practitioner is claiming less money than expected, the quality measure should indicate this, but for this measure the positive quality value of 200 suggests more money than expected is claimed. The measure based on the average value of $c \cdot t$ is too much biased towards regions with high values for c only, and its quality value is not interpretable at all.

3.1 Measures weighting counts by costs

When the emphasis of the measure should still be on the deviation in observed counts (rather than costs), the following measures can be used. The idea is to weight the deviation in counts in the true positive cell of the counts cross table. Such a positive deviation (so observing more true positives than expected), will be more interesting if more money is involved in those true positives. The measures we propose weight this deviation by costs.

$$CWTPD(S) = \left(TP - \frac{1}{N}(TP + FP)(TP + FN)\right) \cdot \overline{c}_{S \cap T} \tag{1}$$

This measure is called the Cost-Weighted True Positive Deviation (CWTPD). The first of the two factors in this equation is the deviation (in counts) within the subgroup from the expected value. This part is similar to the WRAcc measure [4] for binary targets. Only here the deviation is measured in number of observations instead of fraction of the whole dataset, as the WRAcc measure does. This deviation is then multiplied by the average costs of true positives. Hence the measure can be interpreted as: difference in counts × costs involved per count = total costs involved. The big advantage of this definition is that it has a direct interpretation in terms of money. The disadvantage of this measure, especially for the local subgroup discovery task, is that it does not take the costs outside the subgroup into account. It could be that the costs in the reference group outside the subgroup are also high.

The measure following equation 2 eschews this disadvantage, by compensating in the second factor with the costs outside the subgroup:

$$\textit{Relative } CWTPD(S) = \left(TP - \frac{1}{N}(TP + FP)(TP + FN)\right) \cdot (\overline{c}_{S \cap T} - \overline{c}_{\neg S}) \tag{2}$$

This quality measure is called the Relative Cost-Weighted True Positive Deviation (Relative CWTPD). In our application, the measure can be interpreted as the amount of money that would be claimed less if the cross table of counts would be homogeneous. This can be viewed as moving examples from the TP cell into the FP cell until the expected costs cross table is obtained, where costs of non-subgroup examples are estimated by $\overline{c}_{\neg S}$.

The measure in Equation (2) is very suitable for local subgroup discovery because it searches for difference in counts and difference in costs between the subgroup and the examples outside the subgroup simultaneously.

3.2 Measures based on cost difference

To find subgroups for which the mean costs of the target are different from the negative examples, the Total Mean Cost difference between Classes (TMCC) (3) can be used.

$$TMCC(S) = TP \cdot (\overline{c}_{S \cap T} - \overline{c}_{S \cap F}) \tag{3}$$

This measure compares the mean costs of the positive examples with those of the negative examples. Subgroups for which this difference is high are the most interesting. To obtain a total amount (as a monetary value), the difference in means is multiplied by the number of true positives.

$$\begin{array}{l}
t = \{ \;\; + \;, \;\; - \;, \;\; - \;, \;\; - \;, \;\; + \;, \;\; - \;\;, \ldots \} \\
c = \{ \; 2000, \; 1000, \; 1000, \; 1000, \; 2000, \; \underset{\underset{\sigma}{\uparrow}}{1000}, \; \ldots \; \}
\end{array}$$

Fig. 2. A dataset with indicated subgroup of six examples. The mean value for c is higher for the positive examples than for the negative examples in the subgroup. This leads to quality $(2000 - 1000) \cdot 2 = 2000$, computed with Equation (3).

When the number of false positives is very small, the estimate of $\overline{c}_{S \cap F}$ can be based on too few examples. A more robust measure is obtained by using $\overline{c}_S$ instead:

$$TMC(S) = TP \cdot (\overline{c}_{S \cap T} - \overline{c}_S) \tag{4}$$

This measure is called the Total Mean Cost difference (TMC).

The advantage of the TMCC and TMC quality measures is the interpretability: the quality value corresponds directly to the amount of money that is involved. The disadvantage for local subgroup discovery is that these measures do not take the reference group into account at all. Figure 2 shows the calculation of the quality measure. This subgroup will not be found with quality measure (1) or (2) if the probability of the target being true is higher outside the subgroup than inside the subgroup.

3.3 Measures based on the proportion of costs

The previous measures were detecting differences in one target vector, and weighing this distance with the other target vector. The following measure is based on another approach: it considers the difference in distribution of total costs. We can define a cross table of observed costs. This table can be obtained by simply multiplying cells of the two basic cross tables about counts and costs:

	T	F
S	$\sum_{x^i \in S \cap t} c^i$	$\sum_{x^i \in S \cap F} c^i$
$\neg S$	$\sum_{x^i \in \neg S \cap t} c^i$	$\sum_{x^i \in \neg S \cap F} c^i$

This quality measure operates on this cross table of observed total costs. It is based on the proportion of costs per cell relative to the total costs within the whole dataset. It does not take the size of the subgroup into account. We can calculate the $Costs - WRAcc$ measure (called WRAcc due to its similarity to the WRAcc measure for a single binary target vector t). When we denote $c_T = \sum_{i=1}^{N} c^i$ as the total costs in the dataset, the Proportional Costs Deviation (PCD) measure can be calculated as follows:

$$PDC(S) = \frac{1}{c_T} \sum_{x^i \in S \cap t} c^i \quad - \quad \frac{1}{{c_T}^2} \sum_{x^i \in S} c^i \sum_{x^i \in T} c^i \tag{5}$$

This measure can be interpreted as the fraction of costs that is observed beyond expectation in the true positive cell, relative to the total costs in the whole dataset.

In our application, suppose the subgroup indicates cancer patients. A value of 0.1 would mean that in the true positive cell, the fraction of costs compared to the whole data set is 10 % higher than expected. Because this interpretation as a fraction of the total costs in the data set is rather difficult, a much more intuitive measure is the one that has a monetary value. This can be obtained by multiplying equation (5) with the total costs:

$$MVPDC(S) = \sum_{x^i \in S \cap t} c^i \quad - \quad \frac{1}{c_T} \sum_{x^i \in S} c^i \sum_{x^i \in T} c^i \tag{6}$$

This quality measure is called the Monetary Valued Proportional Costs Deviation (MVPCD). This monetary value can be interpreted as the amount of money that is observed beyond expectation in the true positive cell of the total costs cross table, if the total costs distribution would be the same for the positives and negatives. The higher the value of the measure, the more interesting a subgroup is. In our example of cancer patients, a value of $100,000$ would mean that the total amount spent on cancer patients by the target hospital is $100,000$ more than expected. The advantage of this measure is that this measure can detect both deviations in average costs in the subgroup as well as deviations in counts. A disadvantage can be that the calculation of the expected value depends on the total costs distribution of points in T. In our example of the subgroup of cancer patients, it can be that the cancer patients are not more present in this hospital, and also cancer patients are not more expensive than cancer patients at other hospitals, but due to the presence of 'cheap' patients (with a relative low value for c) outside the subgroup, the proportion of observed costs spent on cancer patients can still be higher than expected. The fact that the costs outside the subgroup also play a role in calculating the expected value can cause misinterpretation.

4 Experiments and results

In this section we show how the quality measures are used to detect interesting local subgroups in a real-world application. Our health care application concerns fraud amongst dentists. Each patient is represented by a binary vector of treatments that the patient received during a year. The dataset contains 980,350 patients and 542 treatment codes. As a distance measure between patients we use the Hamming distance between the treatments they received. Note that because of the discrete nature of the data, there are many duplicate examples (many patients with an identical combination of treatments). Additionally, the distance

Table 2. The *observed counts* cross table and the *observed costs* cross table, for the subgroup found with the weighted costs measure

	T	F
S	8	77
$\neg S$	0	101

	T	F
S	$1,619$	697
$\neg S$	0	686

of a point to different neighbors may be identical, which limits the number of subgroups that need to be tested.

We select a dentist with a markedly high claiming profile, and define the target vector t accordingly. The dentist is visited by $5,567$ patients (0.57% of the total data set). The costs vector c is calculated by summing the costs spent on the treatments that the patient received during the year.

4.1 Results with the Relative Cost-Weighted True Positive Deviation

We start with results for quality measure (2). Because the Relative CWTPD measure in equation 2 is very suitable for local subgroup discovery, where the CWTPD measure from equation 1 is better suitable for normal subgroup discovery. Within local subgroup discovery we can alter the reference group size, and 'zoom in' to different resolutions. The following subgroup is found at reference group size 186, with a subgroup size of 85 patients. The prototype patient is using the treatments:

$$\{221153, \text{C11}, \text{C12}, \text{D22}, \text{D24}, \text{D32}, \text{D33}, \text{D42}, \text{M20}, \text{M50}\}$$

Treatment C11 and C12 are regular consults, treatment M20 and M50 are dental cleaning treatments, and treatment 22153, D22, D24, D33, and D42 are orthodontist treatments performed by a dentist. Table 2 (left) shows the counts within this part of the data set. When we observe the cross table with counts we see that there are 8 patients that visit the target dentist, and none in the reference group (outside the subgroup). Patients within the subgroup have treatments similar to the prototype, where patients outside the subgroup, but in the reference group, also use similar treatments, but use less, and use other treatments as well. The table on the right shows the corresponding mean costs. From the two tables we can see that the number of true positives is higher than expected, and the mean costs for observations in the true positive cell are also higher than the mean costs within the other cells. The expected value for the number of true positives is 3.66. This leads to a quality of $(8 - 3.66) \cdot (1,619 - 686) = 4,051$ euros.

To further investigate the observations within the subgroup, and compare them to the rest of the reference group, we observe Table 3, where the support and costs of all frequent treatments in the reference group are compared. Only treatments that have a support ≥ 0.1 are in the table, which means that the

Table 3. Prototypes and their support in the subgroup, and their support in the reference group excluding the subgroup. The codes indicate treatments that were charged for a patient, the supports indicate the fraction of patients receiving those treatments respectively. The costs indicate the mean costs spent on each treatment.

Subgroup	**Prototype, Supports, and Costs**									
S_1 prototype	221153	C11	C12	D22	D24	D32	D33	D42	M20	M50
$S_1 \cap T$	1.00	1.00	0.63	0.18	0.75	0.25	0.50	0.63	0.88	0.25
S_1	0.78	1.00	0.44	0.05	0.31	0.22	0.31	0.85	0.72	0.26
$R_1 \setminus S_1$	0.75	1.00	0.57	0.38	0.20	0.13	0.13	0.88	0.56	0.26
$\bar{c}_{S_1 \cap T}$	169	37	17	175	197	90	131	233	46	5
$\bar{c}_{S_1}$	85	32	12	80	42	73	49	224	29	7
$\bar{c}_{R_1 \setminus S_1}$	73	34	16	22	18	41	19	233	24	5

costs spent on the treatment are bigger than zero for more than 10 percent of the patients in the reference group.

The first line in Table 3 corresponds to these treatments. The next lines correspond to the supports in the set $S_1 \cap T$ (the true positives), the support in S_1 and the support in $R_1 \setminus S_1$ (the patients outside the subgroup, but in the reference group). For each treatment, we can also calculate the mean costs in these sets. These numbers are in the last three lines of Table 3. For this subgroup we can conclude that more orthodontist treatments are claimed (codes D22, D24, D32, D33) within the subgroup compared to the rest of the reference group. From the mean costs numbers, we can conclude that the D22 and D24 treatments are interesting for this subgroup, because of the high costs of those treatments for the true positives.

4.2 Detecting outliers

When restricting the reference group to very small sizes it is possible to find very small groups of outliers, or even find individual outliers as a subgroup. For example, with a reference group size ρ of 13, the best subgroup found for this value of ρ has only one observation with costs of 3779 euros, compared to the mean costs of 1073 euros for its nearest neighbors. The quality value for this individual outlier (again using the Relative CWTPD measure in Equation 2) is $(1 - {}^{1}\!/_{13})(3779 - 1073) = 2,498$ euros.

4.3 Measure based on cost difference

With the TMCC quality measure, using Equation (3), subgroup S_2 was found for the following prototype:

$$\{\text{A10, C11, C12, C13, E01, E13, E40, H30, M50, M55, R25, R31, R74, V11, V12, V13, V14, V20, V21, V40, V60, V80, X10, X21}\},$$

Table 4. The *observed count* cross table and the *observed costs* cross table, for the subgroup found with the costs wracc measure

	T	F
S	87	1236
$\neg S$	110	4390

	T	F
S	427	347
$\neg S$	411	307

Table 5. Prototypes and their support in the subgroup, and their support in the reference group excluding the subgroup. The codes indicate treatments that were charged for a patient, the supports indicate the fraction of patients receiving those treatments respectively. The costs indicate the mean costs spent on each treatment.

Subgroup	**Prototype, Supports, and Costs**										
S_3 prototype	C11	C12	M55	V12	V13	V14	V21	V40	V60	X10	X21
$S_3 \cap T$	1.00	0.93	0.97	0.85	0.31	0.49	0.99	0.47	0.84	0.89	0.74
S_3	1.00	0.90	0.99	0.85	0.21	0.54	0.99	0.49	0.56	0.91	0.29
$R_3 \setminus S_3$	1.00	0.86	0.96	0.82	0.21	0.35	0.97	0.28	0.29	0.87	0.17
$\bar{c}_{S_3 \cap T}$	37	29	58	51	19	44	69	4	24	21	36
$\bar{c}_{S_3}$	36	25	59	47	16	46	52	5	13	22	14
$\bar{c}_{R_3 \setminus S_3}$	35	24	55	45	16	28	46	3	7	21	9

which is a single patient using a combination of many treatments. Patients within the subgroup have a maximum distance of 7 treatments to this prototype. The mean costs of the true positives, $\bar{c}_{\{S_2 \cap T\}}$, is 983 euros, and the mean costs of patients for which the target is false is 773 euros. In the set $S \cap T$ there are 89 patients, while in the set $S \cap F$ there are 592 patients. This leads to a quality value of $18,665$ euros. When we investigate the subgroup, the main difference in costs are due to the treatments R25 (a metal crown with porcelain on top), for which the difference between the target and non-target points is 66 euros, V21 (polishing a filling) with a difference of 31 euros, and V60 (a pulpa-coverage), and X21 (X-ray) each with a difference of 21 euros. With this measure we were also able to mine individual outliers: this comes down to a k-nearest neighbor outlier detection algorithm for which each target point is compared to the mean value of its k nearest neighbors.

4.4 Measure based on the proportion of costs

We calculate the MVPCD measure (6). The best subgroup for a maximum reference group size ρ of 6,000, has a quality of $16,476$. The optimal quality value is found for a σ of 1,323 and a ρ of $5,823$. Table 5 shows the difference in treatments and treatment costs, to get an idea what is the difference between the subgroup and the rest of the reference group.

Patients in this reference group are using the following treatments: $C11$ and $C12$ are regular consults, $M55$ is a dental cleaning, $V12$, $V13$, and $V14$ stand for

2-hedral, 3-hedral, and 4-hedral fillings. $V40$ is for polishing amalgam fillings, $V60$ for a pulpa-covering, and $X10$ and $X21$ for an inexpensive and expensive X-ray respectively. From Table 5, we can see that the main difference between the subgroup and the reference group are the treatments V21 (costs for polishing a filling), and X21 (costs for an expensive X-ray picture). We can conclude that for patients using standard consults, a dental cleaning, and a few fillings, the treatments V21 and X21 are claimed much more often at this dentist. In total, for this patient group, an amount of $16,476$ euros is claimed more than expected.

5 Conclusion

In this paper, we have presented several suggestions for quality measures that involve both binary labels and costs. We demonstrated their effectiveness in producing interesting and actionable patterns in a fraud detection application. As is common in Subgroup Discovery, more than one definition of interestingness can be conceived, and it is up to the end user to determine which measure best fits the specific analysis. We have proposed several measures, and explained the specific benefits of each.

References

1. S. Bay and M. Pazzani. Detecting group differences: Mining contrast sets. *Data Mining and Knowledge Discovery*, 5(3):213–246, 2001.
2. G. Dong and J. Li. Efficient mining of emerging patterns: discovering trends and differences. In *Proceedings of KDD '99*, pages 43–52, New York, NY, USA, 1999.
3. R. M. Konijn, W. Duivesteijn, W. Kowalczyk, and A. Knobbe. Discovering local subgroups, with an application to fraud detection. In *PAKDD 2013*, 2013.
4. N. Lavrac, P. Flach, and B. Zupan. Rule evaluation measures: A unifying view. In S. Dzeroski and P. Flach, editors, *Inductive Logic Programming*, volume 1634 of *Lecture Notes in Computer Science*, pages 174–185. Springer Berlin / Heidelberg, 1999.
5. S. Wrobel. An algorithm for multi-relational discovery of subgroups. *Proceedings of PKDD*, pages 78–87, 1997.