J Intell Inf Syst (2015) 45:337–355
DOI 10.1007/s10844-014-0313-8

Cost-based quality measures in subgroup discovery

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

Received: 21 July 2013 / Revised: 17 December 2013 / Accepted: 13 February 2014 /
Published online: 26 March 2014

Abstract 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 with a somewhat unusual target concept, and investigate quality measures that take into account both the binary and the cost target. In defining such quality measures, we aim to produce an interpretable valuation of a subgroup, such that data analysts can directly appreciate 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 domain, the binary target identifies the patients of a specific medical practitioner under investigation, and the cost target specifies the money spent on each patient. When looking for differences in claim behavior, we 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 list a number of treatments, and the target

This paper is an extended version of the paper with the same title [12] which appeared in *New Frontiers in Applied Data Mining—PAKDD 2013 International Workshops*, 2013.

R. M. Konijn (✉) · W. Duivesteijn · M. Meeng · A. Knobbe
LIACS, Leiden University, Leiden, The Netherlands
e-mail: konijn@liacs.nl

W. Duivesteijn
e-mail: wouter.duivesteijn@tu-dortmund.de

M. Meeng
e-mail: meeng@liacs.nl

A. Knobbe
e-mail: knobbe@liacs.nl

R. M. Konijn
Achmea Health Insurance, Zeist, The Netherlands

practitioner's patients behavioral difference in both treatment prevalence and associated costs. An additional angle is the Local Subgroup Discovery task, where subgroups are judged according to the difference with a local reference group instead of the entire dataset. We show how the cost-based analysis of data specifically fits this local focus.

Keywords Subgroup discovery · Quality measures

1 Introduction

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 which treatments he performed on this patient, and he knows what treatments he claimed with 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 an 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 example summarizes the care a patient received over a certain period. We are interested in finding patient groups (subgroups), that describe the difference between a single medical practitioner and his 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. Tasks identifying such groups are Subgroup Discovery (Wrobel 1997), Emerging Pattern Mining (Dong and Li 1999), or Contrast Set Mining (Bay and Pazzani 2001).

In order to find such deviating subgroups, we need quality measures to describe how 'interesting' a subgroup is. We would like the quality measure to capture the distributional difference between one practitioner and the others: the higher the difference, the more interesting a subgroup is. From an economical point of view, it makes sense to involve costs in the quality measure; subgroups involving more money are more interesting. In this paper, we will design quality measures taking costs into account. Each patient is 'labeled' with a commodity—in our main application this is the total money spent on treatments during a specific period—and the quality measures we develop use these commodity 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.

This article is an extended version of the paper published at the QIMIE workshop (Konijn et al. 2013b). In that paper, only the Local Subgroup Discovery (LSD, to be explained later) task was considered. This paper discusses how the quality measures work in the 'Descriptive' Subgroup Discovery setting, and shows examples thereof.

2 Preliminaries

Throughout this paper we assume a dataset D with N examples (typically patients). Each row can be seen as a $(h+2)$-dimensional vector of the form $x = (a_1, \ldots, a_h, t, c)$. 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 $a^i = \left(a_1^i, \ldots, a_h^i\right)$ the *attributes* of the ith 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 whether a patient visited the practitioner. The other target, c, indicates a commodity. In our primary application, this commodity indicates the total costs of treatments, per year. For other applications, c could indicate e.g. a profit or a per-customer value. We will refer to the target values of a specific example by superscript: t^i and c^i are the targets of example x^i.

Our goal is bivariate: find differences between the singled-out medical practitioner and the rest, that simultaneously constitute a considerable amount of costs; the more costs involved, the better. These differences are expressed by subgroups. A subgroup can be seen as a bag of examples: it can be any subset $S \subseteq D$ of the dataset. The interestingness of a subgroup is quantified by a quality measure $q : 2^D \to \mathbb{R}$, which is a function assigning a numeric value to any subgroup. Such a function should prefer subgroups that have a substantial size, and a substantial distributional difference over the targets when compare to the overall target distribution. This difference is ideally gauged in terms of both the binary target t and the commodity target c simultaneously.

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 $t = -$. 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: $TP = (t = +) \cap S$, $FP = (t = -) \cap S$, $FN = (t = +) \cap \neg S$, and $TN = (t = -) \cap \neg S$. 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 Subgroup Discovery

There are multiple ways to describe a subgroup S. The most common way in Subgroup Discovery is to describe a subgroup as a conjunction of attribute-value combinations. We will call this *Descriptive* Subgroup Discovery (DSD). We will explain the search for these subgroups in Section 2.1.1. Another way to describe a subgroup S is by means of a prototype and a distance or a number of nearest neighbors. We will call this *Local* Subgroup Discovery (LSD). This will be further explained in Section 2.1.2.

2.1.1 Descriptive Subgroup Discovery

The most common way to describe subgroups in Subgroup Discovery is by using attribute-value descriptions. Attributes are allowed to have a binary, nominal, or numeric domain. A subgroup is defined by a *description*, which is a conjunction of conditions on the attributes, for instance $Age \leq 18 \wedge Gender = Female$.

In DSD, quality is assessed in terms of the distributional difference of only one binary target attribute t. To find high-quality subgroups, the usual choice is a top-down search strategy. First, we start with small subgroup descriptions, defined by only one condition on one attribute. These descriptions are extended with other attribute-value conditions, thus

creating more specific subgroups. We say we *refine* the generic subgroups into more specific subgroups. Usually two search parameters are used: a minimum support threshold *minsup* describes the minimum size a subgroup is required to have, and a maximum depth parameter d^{max} sets a maximum on the number of attribute-value conditions in the subgroup description.

If the dataset is small enough, exhaustive search can be used to search for subgroups with the highest quality. To speed up the search, pruning strategies can be used that take optimistic estimates (Grosskreutz et al. 2008) into account. An optimistic estimate is the maximum quality that any refinement of a subgroup can have. If this optimistic estimate is lower than a threshold (for example the kth best subgroup found so far), refinements of the corresponding subgroups are disregarded.

For big datasets, containing many high-cardinality attributes, exhaustive search is infeasible due to the large number of attribute-value conditions. The commonly-used alternative is a search strategy called *beam search*. It uses a levelwise top-down strategy, only exploring promising parts of the search space. At each search level, refinements are obtained of only the top-w subgroups (called the beam) from the preceding level. The parameter w is called the beam *width*. The search stops when the maximum depth d^{max} is reached.

2.1.2 Local Subgroup Discovery

The idea of the Local Subgroup Discovery (LSD) task (Konijn et al. 2013a) is to 'zoom in' on a part of the dataset, and detect interesting subgroups locally. In our application, the patient population is distributed among different patient groups (e.g., one group could be patients having a type of cancer). A patient group 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 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

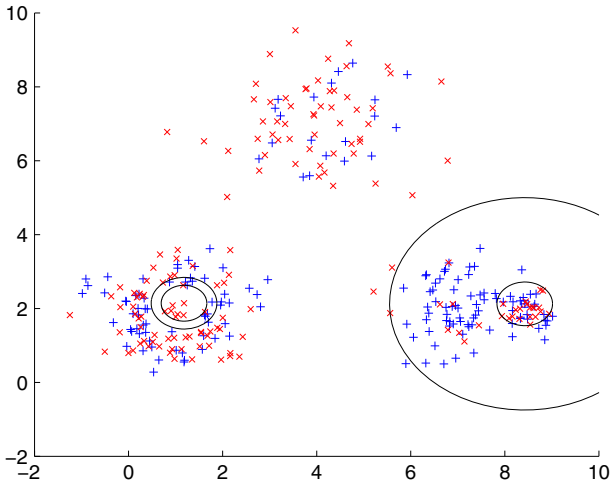


Fig. 1 Two local subgroups that are hard to find with traditional techniques

$$ranking(x) = \{ +, +, -, +, -, +, +, +, -, -, -, +, -, -, \dots \}$$
$$\uparrow_{\sigma} \qquad\qquad \uparrow_{\rho}$$

Fig. 2 Target vector ranking for a prototype x. The target vector is sorted left-to-right according to increasing distance to x. All examples up to σ are in the subgroup. All examples up to ρ are in the reference group

the ρ nearest neighbors of x in D. In LSD, quality is assessed in terms of the distributional difference between the distance-based subgroup and its reference group, disregarding data outside of the reference group.

The LSD task is visualized in Fig. 1. In DSD, we seek groups with relatively many positives compared to the whole dataset. In LSD, instead, we seek groups with relatively many positives compared to the reference group. Two local subgroups are shown as the smaller circles, surrounded by their reference groups.

Distance-based subgroups and reference groups are illustrated in Fig. 2. The goal of LSD is to find subgroups $S_\sigma \subset R_\rho$ for which the target distribution is different from that distribution in the reference group. For a single prototype x, this search is quadratic (since optimal values for both σ and ρ are sought) in the number of neighbors that are considered for x (which is at most the total dataset). To assure substantially-sized results, a minimum support threshold should be used. Alternatively, one can search for the most significant quality values (Konijn et al. 2013a).

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 this 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 representing a distributional deviation *relative to their peers*: we want to find subgroups $S_\sigma \subset R_\rho$ in which the target is distributed differently from the reference group.

3 Related work

Including costs in data mining and machine learning methods is not new. However, for the subtask of Subgroup Discovery there are no known methods that include costs in the quality measure. For other data mining techniques, costs inclusion is well-established; we will argue that they have different goals. In an extension of Frequent Pattern Mining, costs are assigned to each item. This number is usually called the *utility* of an item. High-Utility Itemset Mining algorithms (Chan et al. 2003; Liu et al. 2005) mine itemsets that have a utility exceeding some threshold. For example, a retail business may be interested in identifying its most valuable customers by mining these high utility itemsets (where the utility is the profit on an item). The difference between our setting and the high utility approach is that our setting is supervised: we are interested in differences between positive and negative examples, rather than high support/utility in the dataset only. Also, in our approach the items themselves do not have a cost assigned to them, but the customers do.

Cost Sensitive Learning (Elkan 2001; Hernández-Orallo et al. 2011) assigns costs to predictions. A cost matrix specifies the costs of labeling an example for the four different cases: True/False Positive/Negative. In our setting, we also have a contingency table describing the

subgroup and the target. However, there are no 'misclassification costs' for False Positives and False Negatives, so these methods do not readily apply.

In Subgroup Discovery with a single continuous target (Atzmueller and Lemmerich 2009; Jorge et al. 2006; Pieters et al. 2010), quality measures are either differences between the probability density function of the continuous target (Jorge et al. 2006), differences in statistical significance (Pieters et al. 2010), or ratios or differences of the mean value of the continuous target (Atzmueller and Lemmerich 2009). The setting with a binary as well as a continuous target is not discussed in these papers. Subgroup Discovery with a cost target on health care data has been performed in Grosskreutz (2010). The authors search for a set of subgroups that best describe costs of patients, by taking the costs as the (single) target variable. The article focuses on the regression model-building part, rather than the quality description and the detection of subgroups. An extension to Grosskreutz (2010) is described in the paper Konijn and Kowalczyk (2012), also applied on health care data. Here also, a model is built with costs as target variable (in this case the model is a Random Forest). After the model is created, descriptions are generated per practitioner to describe their difference in claim behavior from the model. This method detects claims or patients that are outliers. Usually the outlier score does not have a monetary value, and therefore the interpretability of the quality measure can be difficult. When the outlier score does have a monetary value, the quality measures described in this paper can be used to gauge the interestingness of subgroups.

4 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 DSD. The cross table on the right is concerned with the mean costs for each of the categories. Our quality measures should assign preference to the following concepts:

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

Furthermore, having a direct interpretation in terms of the commodity expressed by c makes a quality measure easier to interpret. In Sections 4.1.1–4.1.3, we introduce quality measures satisfying these criteria. Before that, we shortly discuss why a naive Subgroup Discovery approach does not suffice.

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 reference group in Fig. 3. The first 6 examples (up to σ) belong to the subgroup, and the other examples

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

	$t=+$	$t=-$
S	TP	FP
$\neg S$	FN	TN

	$t=+$	$t=-$
S	$\overline{c}_{S \cap (t=+)}$	$\overline{c}_{S \cap (t=-)}$
$\neg S$	$\overline{c}_{\neg S \cap (t=+)}$	$\overline{c}_{\neg S \cap (t=-)}$

$$
\begin{array}{rl}
t = & \{\ +\ ,\ -\ ,\ -\ ,\ -\ ,\ +\ ,\ -\ ,\ +\ ,\ +\ ,\ +\ ,\ -\ ,\ -\ ,\ -\ \} \\
c = & \{\ 1000,\ 2000,\ 2000,\ 1250,\ 2000,\ 3000,\ 200,\ 200,\ 200,\ 200,\ 200,\ 200\ \} \\
c \cdot t = & \{\ 1000,\ \ 0\ \ ,\ \ 0\ \ ,\ \ 0\ \ ,\ 2000,\ \ 0\ \ ,\ 200,\ 200,\ 200,\ \ 0\ \ ,\ \ 0\ \ ,\ \ 0\ \ \} \\
 & \qquad\qquad\qquad\qquad\qquad\qquad\quad \uparrow \\
 & \qquad\qquad\qquad\qquad\qquad\qquad\quad \sigma
\end{array}
$$

Fig. 3 A reference group with a subgroup of six examples indicated

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 resulting value of 200 does not have a meaningful interpretation in terms of money: 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 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 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. The relatively low number of true positives indicates that there are fewer 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 positives is less than the false positives. Since overall, the practitioner is claiming less money than expected, a quality measure should indicate this, but for this measure the positive quality of 200 suggests that more money than expected is claimed. The measure based on the average value of $c \cdot t$ is overly biased towards regions with high values for c only, and assigned qualities are hard to interpret.

4.1 The new cost-based quality measures

We introduce six quality measures, roughly falling into three categories. Measures from the first category focus on the distributional difference in the binary target: the CWTPD assigns to a subgroup the total costs involved in the subgroup, and the Relative CWTPD measures the amount of money that would be recovered if the patients were distributed homogeneously within and outside of the subgroup. Measures from the second category focus on the distributional difference in the real-valued target: the MCC finds subgroups that maximizes the distance between average real-valued target values, and the TMCC computes the amount of money involved in the distance. Measures from the third category take a hybrid approach to the two targets: the PCD gauges the fraction of costs that is observed beyond expectation in the subgroup for the practitioner relative to the total costs in the whole dataset, and the MVPCD gauges the amount of money observed beyond expectation within the subgroup for the practitioner.

Depending on domain expert demands, different quality measures may be preferable. CWTPD has a very clear interpretation, but does not take into account that a practitioner may also serve relatively expensive patients *outside* of the subgroup, which Relative CWTPD does. MCC finds very specialized subgroups having very expensive members, but the subgroups are usually not very large; TMCC is designed to find larger subgroups. PCD computes a fraction, which is dimensionless, allowing comparison of PCD-measured quality of results over multiple datasets, but it is not trivial to interpret; MVPCD caters for the

domain expert in that respect. The focus of (Relative) CWTPD on the membership target results in subgroups that are markedly different from results of measures from the other two categories, which focus at least partially on the commodity target.

4.1.1 Measures weighting counts by costs

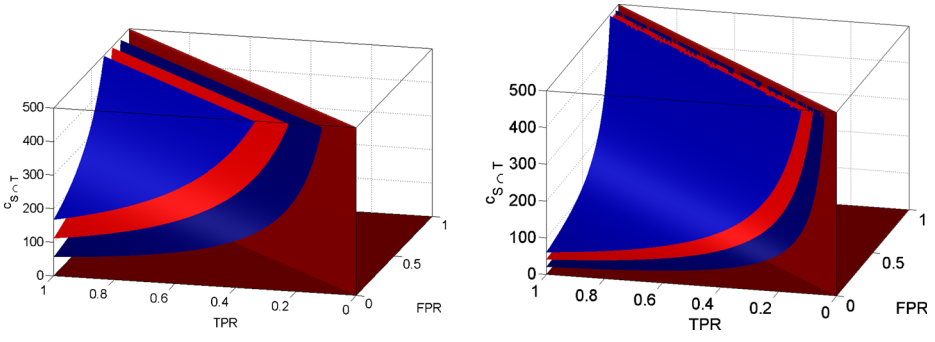
When the emphasis of the measure should still be on the deviation in observed counts (rather than costs), the following measures can be used. The idea is to weight the deviation in counts in the true positive cell of the counts cross table. Such a positive deviation (i.e. observing more true positives than expected), is particularly relevant if a substantial sum of money is represented by 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 (1) is the deviation (in counts) within the subgroup from the expected value. This factor is equivalent to the WRAcc measure (Lavrač et al. 1999) for binary targets, except that we measure the deviation in the absolute number of observations instead of relative. 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 value of CWTPD depends on the number of true positives and false positives, and the average costs ($\overline{c}_{S \cap (t=+)}$) within the subgroup. The other quantities (FN and TN) are determined by TP and FP. To investigate the behavior of the CWTPD measure, Fig. 4a shows several isometrics for this measure. Each surface in the figure represents the same CWTPD-value for different values of $\overline{c}_{S \cap (t=+)}$, TP and FP. Comparison with Fig. 4b illustrates the effect of increasing the share of positives in the dataset on the isometrics. The CWTPD-value for the same subgroup in this isometric plot increases when the size of the dataset increases; when the dataset is bigger, more costs will be involved.

The big advantage of CWTPD is that it has a direct interpretation in terms of money. Its disadvantage, especially for LSD, is that it does not consider the costs outside of the



(a) $|t = +|/|D| = 10\%$ (b) $|t = +|/|D| = 50\%$

Fig. 4 CWTPD isometrics

subgroup. The costs in the reference group outside the subgroup could also be high. The following measure counters this disadvantage, by compensating in the second factor for the costs outside the subgroup:

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

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 FN cell until the expected costs cross table is obtained, where costs of non-subgroup examples are estimated by $\overline{c}_{\neg S}$.

Relative CWTPD is very suitable for LSD because it simultaneously searches for difference in counts, and difference in costs between the subgroup and the examples outside the subgroup.

4.1.2 Measures based on cost difference

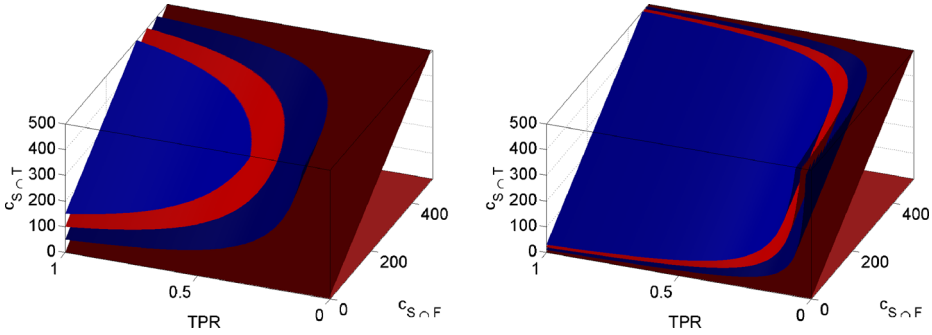
CWTPD emphasizes deviation in observed counts. Hence, CWTPD is unable to detect subgroups for which the only deviation is in the costs. The measures in this section target such subgroups. To find subgroups for which the mean costs of the target are different from the negatives, the Mean Cost difference between Classes (MCC) measure can be used. The MCC detects subgroups for which the continuous target has a higher value for the positives:

$$MCC(S) = \overline{c}_{S\cap(t=+)} - \overline{c}_{S\cap(t=-)} \quad (3)$$

MCC generally finds small subgroups with a big difference in costs. Preferring larger subgroups, the Total Mean Cost difference between Classes (TMCC) computes the total amount of money that is involved in this difference:

$$TMCC(S) = TP \cdot \left(\overline{c}_{S\cap(t=+)} - \overline{c}_{S\cap(t=-)}\right) \quad (4)$$

This measure compares the mean costs of the positives with those of the negatives. 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. TMCC isometrics are shown in Fig. 5a and b, with $\overline{c}_{S\cap(t=+)} \in [0, 500]$ for readability.



(a) $|t = +|/|D| = 10\%$ (b) $|t = +|/|D| = 50\%$

Fig. 5 TMCC isometrics

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

Fig. 6 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 (4)

The horizontal asymptote lies at $\overline{c}_{S\cap(t=-)}$ and the vertical asymptote at $TP = 0$. MCC isometrics are not plotted: its isometrics only depend on $\overline{c}_{S\cap(t=-)}$, and do not vary with the size of the subgroup or the number of true positives. The number of true positives and false positives in a subgroup can be very small, in which case estimates for $\overline{c}_{S\cap(t=+)}$ and $\overline{c}_{S\cap(t=-)}$ are based on very few examples. To ensure substantiated results, a minimum support threshold should be used with this measure. The advantage of MCC and TMCC is their 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 consider the reference group.

Figure 6 illustrates the computation of the quality measure. The indicated subgroup will not be found with (Relative) CWTPD if the probability of the target being true is higher outside the subgroup than inside the subgroup.

4.1.3 Measures based on the proportion of costs

The previously introduced measures detect differences in one target vector, and weight this difference with the other target vector. The following measures are based on another approach: they consider the difference in distribution of total costs. We consider the cross table of observed costs:

	$t = +$	$t = -$
S	$\sum_{x^i \in S\cap(t=+)} c^i$	$\sum_{x^i \in S\cap(t=-)} c^i$
$\neg S$	$\sum_{x^i \in \neg S\cap(t=+)} c^i$	$\sum_{x^i \in \neg S\cap(t=-)} c^i$

Operating on this cross table, the Proportional Costs Deviation (PCD) measure gauges the proportion of costs per cell relative to the total costs within the whole dataset, without considering the size of the subgroup. Denoting the total costs in the dataset by $c_D = \sum_{i=1}^{N} c^i$, the PCD measure can be calculated as follows:

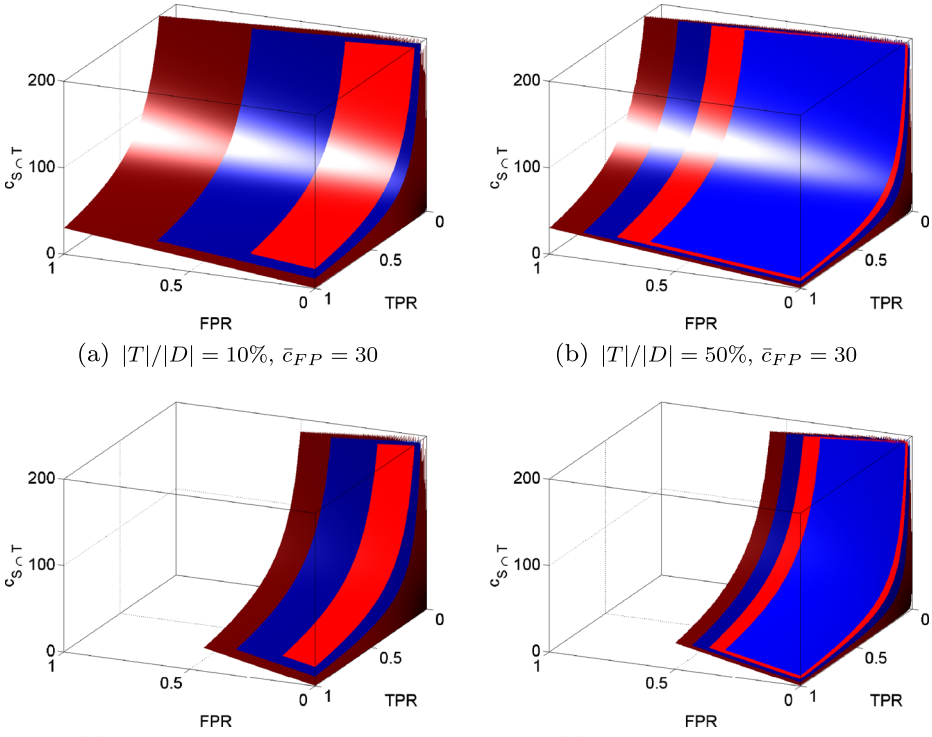
$$PCD(S) = \frac{1}{c_D} \sum_{x^i \in S\cap(t=+)} c^i \ - \ \frac{1}{{c_D}^2} \sum_{x^i \in S} c^i \sum_{x^i \in (t=+)} c^i \tag{5}$$

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 dataset is 10 % higher than expected. From a domain expert point of view, this concept is more intuitively expressed as

an absolute monetary value instead of a proportion. Multiplying (5) by the total costs, we obtain:

$$MVPCD(S) = \sum_{x^i \in S \cap (t=+)} c^i - \frac{1}{c_D} \sum_{x^i \in S} c^i \sum_{x^i \in (t=+)} c^i \tag{6}$$

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. 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. MVPCD can detect both deviations in average costs in the subgroup and deviations in counts. Less fortunate might be that the calculation of the expected value depends on the total costs distribution of examples in $(t = +)$. In our example of the subgroup of cancer patients, it can be that the patients are both not more present in this hospital and not more expensive than cancer patients at other hospitals, but due to the presence of 'cheap' patients (with a relatively 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. Figure 7 displays the MVPCD isometrics for an artificial dataset of 1000 examples, with $\overline{c}_D = \overline{c}_{(t=+)} = \overline{c}_{(t=-)} = 30$. Furthermore, we assume $c > 0$.



(a) $|T|/|D| = 10\%$, $\bar{c}_{FP} = 30$ (b) $|T|/|D| = 50\%$, $\bar{c}_{FP} = 30$

(c) $|T|/|D| = 10\%$, $\bar{c}_{FP} = 60$ (d) $|T|/|D| = 50\%$, $\bar{c}_{FP} = 60$

Fig. 7 PCD isometrics

5 Experiments and results

In this section, we illustrate the quality measures by detecting interesting subgroups in our real-life health care application concerning fraud amongst dentists. We will use the Local Subgroup Discovery (LSD) approach. As a second application, we show the results of Descriptive Subgroup Discovery (DSD) on a dataset containing claims from pharmacies. To show the working of quality measures outside the health care domain, we present DSD results on a dataset about the scheduling and outcome of adult criminal court appearances.

5.1 LSD results weighting counts by costs

We illustrate the quality measures by detecting interesting subgroups in our real-life health care application concerning 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 (patients with an identical combination of treatments). Additionally, the distance of a point to different neighbors may be identical, which limits the number of subgroups to be tested. The discrete nature of the data thus limits the possible values for σ and ρ.

We select a dentist with a markedly high claiming profile, and define the target vector t (which is supposed to single out a practitioner, as indicated in Section 2) accordingly. The dentist is visited by 5,567 patients (0.57 % of the total dataset). The costs vector c is calculated by summing the costs spent on the treatments that the patient received during the year.

Since the LSD task is more beneficial in the health care fraud setting than the DSD task, we start with results for quality measure (2), since Relative CWTPD is more suitable for LSD, while the CWTPD measure from (1) is more suitable for DSD. Recall that in LSD, we can alter the reference group size, and 'zoom in' to different resolutions. We search for optimal values of σ and ρ, setting the maximum reference group size $\rho = 1000$. The following subgroup containing 85 patients is found at reference group size 186. The prototype patient is using the treatments:

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

Treatments C11 and C12 are regular consults, M20 and M50 are dental cleaning treatments, and 22153, D22, D24, D33, and D42 are orthodontist treatments performed by a dentist. Table 2 displays the counts and costs within this part of the dataset. The cross table on the left contains the counts. Eight patients in the subgroup 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, use similar but not exactly the same treatments. From the two tables, we can see that the number of true positives is higher than expected (the expected value under

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

	$t=+$	$t=-$
S	8	77
$\neg S$	0	101

	$t=+$	$t=-$
S	1,619	697
$\neg S$	0	686

Table 3 Fractions of patients receiving treatments, and associated costs

S_1	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
$\overline{c}_{S_1 \cap (t=+)}$	169	37	17	175	197	90	131	233	46	5
$\overline{c}_{S_1}$	85	32	12	80	42	73	49	224	29	7
$\overline{c}_{R_1 \setminus S_1}$	73	34	16	22	18	41	19	233	24	5

the assumption that P(S) and P(t $=$ $+$) are independent), 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, comparing support and costs of all frequent treatments in the reference group. The table features only those treatments that have support ≥ 0.1, i.e. the costs spent on the treatment surpass zero for more than 10 % of the patients in the reference group.

The first three (non-header) rows in Table 3 detail the supports of the treatments in several sets: the true positives, the subgroup, and the reference group minus the subgroup, respectively. The last three lines feature the mean costs in these sets. 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 come with suspiciously high costs in the true positives.

5.2 LSD results based on cost difference

With the TMCC quality measure, we search for optimal values of σ, setting the maximum subgroup size $\sigma = 1000$ (note that the TMCC measure is insensitive to values of ρ so we could equally well set the maximum value of $\rho = 1000$). We find subgroup S_2 with prototype x:

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

which is a single patient using a combination of many treatments. Patients within S_2 have a maximum distance of 7 treatments to x. The mean costs of the true positives is 983 euros, and the mean costs of patients for which the target is false is 773 euros. The set $S \cap (t = +)$ contains 89 patients, while 592 patients make up the set $S \cap (t = -)$. This leads to a quality of 18,665 euros. The main difference in costs is 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.

Table 4 Fractions of patients receiving treatments, and associated costs

S_3	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
$\overline{c}_{S_3 \cap (t=+)}$	37	29	58	51	19	44	69	4	24	21	36
$\overline{c}_{S_3}$	36	25	59	47	16	46	52	5	13	22	14
$\overline{c}_{R_3 \setminus S_3}$	35	24	55	45	16	28	46	3	7	21	9

5.3 LSD results based on the proportion of costs

The best subgroup S_3 found with MVPCD for a maximum reference group size ρ of 6,000, has a quality of 16,476. This optimal quality is found for a σ of 1,323 and a ρ of 5,823. Table 4 shows the difference in treatments and costs.

Patients in this reference group are subjected to regular consults (C11, C12), dental cleaning (M55), 2-, 3-, and 4-hedral fillings (V12, V13, V14), polishing amalgam fillings (V40), pulpa-covering (V60), and inexpensive and expensive X-rays (X10, X21). From Table 4, we see that the main difference between the subgroup and the reference group are treatments V21 (costs for polishing a filling), and X21 (the expensive X-ray), when we compare the average costs of the true positives and the average costs in the reference group. Hence, for reference-group patients, treatments V21 and X21 are claimed unusually often at this dentist. In total, for this patient group, 16,476 euros is claimed more than expected.

5.4 DSD results on individual transactions

To illustrate the quality measures for Descriptive SD, we use the raw transactional data of pharmacies (instead of an aggregation on the level of patients). Each individual example describes one line of a prescription that is delivered to a patient. The example contains information about the patient (age, gender, birthdate), prescription date (day of the week, date), and medication (name, therapeutic classification, number of items). The cost vector c is the amount of money that is charged for the medication. As a binary target, we consider again a single practitioner (the target pharmacy), and we compare its transactions with 99 pharmacies from its peer group. The binary target is true for 9,832 examples (0.46 % of the data). To gauge qualities, we use the CWTPD measure to mine the top 50 subgroups, using the Cortana system (Meeng and Knobbe 2011). We use a beam search strategy with a search width of 100, a minimum support threshold of 1 % of the data, and a maximum search depth of 2. The SD run takes 3 minutes and 55 s. Since the top-50 subgroups contains redundancies, we post-process the list by searching for the set of 3 least redundant (gauged by joint entropy; see Knobbe and Ho (2006)) subgroups, detailed in Table 5.

Subgroup S_4 is defined by the pharmacy being visited by many old patients receiving many items (usually pills). It has the description $Birthyear \leq 1942 \wedge \#Pieces \geq 14.0$. There are 4,167 true positives, where we would expect 3,472, so the difference is 694 transactions. These transactions have an average cost of 45.29 euros, leading to a quality of 31,465.49 euros. Subgroup S_5 indicates that relatively much non-generic medication is delivered at this pharmacy. The description of the subgroup is: $MedicationType =$

Table 5 Descriptions of the three least redundant descriptive subgroups from the top-200 found on the pharmacy dataset

S	$\lvert S \rvert$	$\lvert S \cap (t = +) \rvert$	$\overline{c}_{S \cap (t=+)}$	$\overline{c}_{S \cap (t=-)}$	CWTPD
S_4	708,141	4,167	45.29	33.08	31,465.49
S_5	1,055,077	5,671	39.42	34.54	30,388.79
S_6	651,447	3,638	47.65	40.89	29,182.60

Subgroup S_4 is defined by conditions $Birthyear \leq 1942 \wedge \#Pieces \geq 14.0$, subgroup S_5 by conditions $MedicationType = non-generic \wedge \#Pieces \geq 30.0$, and subgroup S_6 by conditions $\#Pieces \geq 90.0$

$non-generic \wedge \#Pieces \geq 30.0$ Non-generic medication is produced by pharmaceutical companies that hold the (sometimes already expired) corresponding patent. It is more expensive than the generic equivalent, so this is an interesting subgroup from a cost perspective. Subgroup S_6 indicates that relatively many items are delivered at once. It has the description $\#Pieces \geq 90.0$. When we investigate the distribution of number of items, we observe indeed that this pharmacy frequently delivers boxes of 90 or 180 items. Also, the mean costs of the true positives is heavily influenced by a few outliers. Eight examples of the target pharmacy correspond to one patient that is using Lenalidomide, an immunosuppressant to treat myeloma. Lenalidomide is an expensive type of medication (the costs per patient are on average 6,453.80 euros per month). Using domain knowledge, this is an interesting finding because it may indicate inefficient claim behavior, since the medication called thalidomide (also an immunosuppressant to treat myeloma) costs on average 82 euros per patient per month. The other two outliers are interesting from a cost/fraud perspective because of deliveries of 3,300 and 6,600 pills (very high amounts) of an anti-parkinson drug.

To compare the computation time of the quality measures on this dataset, we use Matlab to calculate the quality of the top 100 subgroups, 100 times for each subgroup, and average the time that is needed. The average computation time per subgroup for CWTPD is 0.0603 s, 0.0557 s for MCC, and 0.0795 s for PCD. We compare these runtimes with those of traditional DSD quality measures. The average computation time of the popular WRAcc measure (Lavrač et al. 1999) is 0.0258 s. For Klösgen's mean test (Pieters et al. 2010) the average time is 0.0324 s. We can see that the cost-based quality measures take approximately twice the computation time of quality measures that operate on a single column.

5.5 DSD results on CourBC data

In this section, we illustrate using the quality measures when the continuous target c does not have a monetary value. The CourBC dataset (Reid et al. 2013) details scheduling and outcome of all adult criminal court appearances (in total: 82,027 examples), heard in a Provincial Court within the Canadian province of British Columbia (BC) between June 1, 2007 and May 31, 2011. The dataset is one component of a data warehouse maintained at the Institute for Canadian Urban Research Studies (ICURS) at Simon Fraser University (SFU), and was collected from publicly available data published by BC Court Services.

We analyze a subset of the CourBC dataset, concerning only scheduling information. We strive to detect differences between courts, singling out one court as our binary target, randomly picking the Richmond Provincial Court. In total 2,221 court decisions in our dataset are from the Richmond Provincial Court (the number of positive examples is 2,221, or 2.71 %). The relevant commodity in the CourBC dataset is the number of days it takes to reach a verdict. Hence we strive to find subgroups of cases for which the court works

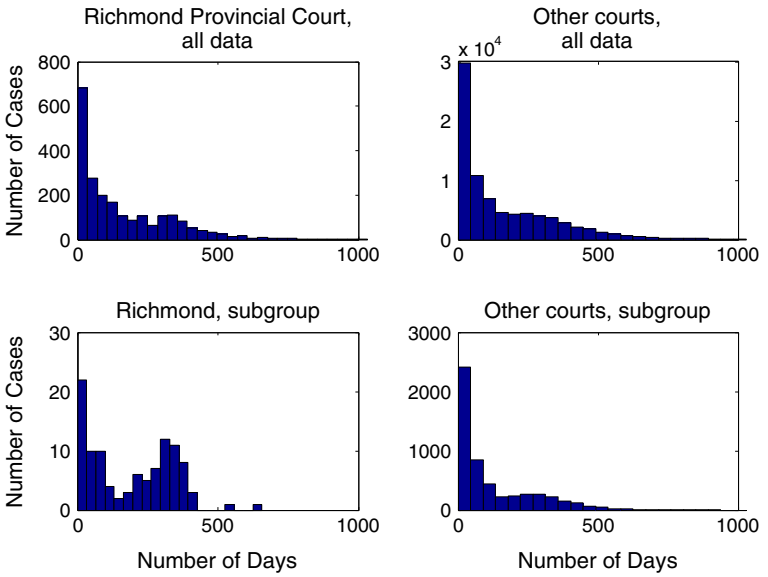


Fig. 8 Distribution of the number of days, for positive and negative examples, on the whole dataset and restricted to subgroup S_7

relatively inefficiently in terms of time spent. Beyond that, the dataset details for each case the *number of people* charged, *number of charges*, average *severity* of the charges, and the *number of appearances* before court. The histograms in Fig. 8 show the distribution of the continuous target c for several subsets of the dataset. The 180 cases (0.22 %) lasting over 1000 days are left out of the figure, to increase legibility.

We run the DSD task with several quality measures. Firstly, with the MCC measure we find the subgroups presented in Table 6. The best subgroup found concerns those cases for which the average severity of the charges lies between 2.5 and 3.5, and that appeared before court less than 8 times. In the CourBC dataset, charges are assigned severity 3, if they correspond to offences that impede the process of justice or result from an earlier offence. Hence, such administrative offences with a relatively low number of appearances take up a disproportionally large amount of days of the Richmond Provincial Court.

Table 6 CourBC descriptive subgroups found with the MCC measure

S	$\lvert S\rvert$	$\lvert S \cap (t=+)\rvert$	$\overline{c}_{S\cap(t=+)}$	$\overline{c}_{S\cap(t=-)}$	MCC
DB	82,027	2221	161.14	161.05	
S_7	5480	105	194.94	121.04	73.91
S_8	19,560	473	231.30	169.21	62.09
S_9	7014	144	235.63	177.85	57.79
S_{10}	21,039	479	195.19	139.22	55.97

The best subgroup found is $S_7 : Severity \leq 3.5 \wedge Severity \geq 2.5 \wedge \#Appearances \leq 8$. The other subgroups in the table are the three least redundant subgroups in the top-50; $S_8 : \#Charges \geq 2.0 \wedge \#Appearances \leq 12 \wedge Severity \leq 4.0$, $S_9 : Severity \leq 3.5 \wedge Severity \geq 2.5$, and $S_{10} : \#Charges = 2 \wedge \#Appearances \leq 9$

Table 7 CourBC descriptive subgroups found with the PCD and MVPCD measure

S	$\lvert S\rvert$	$\lvert S \cap (t=+)\rvert$	$\overline{c}_{S\cap(t=+)}$	$\overline{c}_{S\cap(t=-)}$	PCD	MVPCD
S_{11}	33,531	1093	152.88	135.94	0.00326	43,109.14
S_{12}	39,194	1226	157.69	142.65	0.003131	41,363.16
S_{13}	19,528	661	267.10	257.73	0.003030	40,033.84
S_{14}	28,187	948	210.39	210.2	0.002945	38,904.46

The best subgroup found is S_{11} : $\#Charges \leq 2.0 \wedge \#Appearances \leq 11 \wedge Severity \leq 4.0$. The other subgroups in the table are the three least redundant subgroups in the top-50; S_{12} : $\#Appearances \leq 12 \wedge Severity \leq 5.0$, S_{13} : $Severity \leq 5.0 \wedge \#Appearances \geq 5.0 \wedge \#Appearances \leq 13$, and S_{14} : $\#Charges \leq 2 \wedge Severity \leq 4.0 \wedge \#Appearances \geq 3$

Secondly, with the PCD measure, we find top subgroups that are very similar to each other. We present the 3 least redundant subgroups from the top-200 found subgroups in Table 7. They are typically much larger than the ones found with MCC. From the total number of 357,899 days spent by the Richmond Provincial Court we would expect 123,997 days to be assigned to cases in the top subgroup. Instead, 167,106 days were spend on those cases, so 43,109 days could be gained.

Thirdly, the results obtained with TMCC are quite similar to those with PCD: subgroups that contain $Severity \leq 5$ in their description, and a low number of appearances also score high according to those measures, since for these subgroups both the mean costs and the amount of true positives is high.

Fourthly, the top result with CWTPD is: $Severity \leq 4.0 \wedge \#Appearances \geq 5$. Cases with this description occur relatively often at the Richmond Provincial Court, in total for 36,361.25 days.

Finally, when we compare the results found with MCC with those found with PCD, we see that MCC finds smaller subgroups. The last row in Table 7 details a subgroup found with PCD for which $\overline{c}_{S\cap(t=+)}$ value is not very different from $\overline{c}_{S\cap(t=-)}$, hence the subgroup will not score high according to MCC. For the top subgroup found with MCC, the observed number (105) in the TP cell is less than expected (148.37). This is an example of a subgroup that represents a higher average value in terms of the costs, but is not deviating in terms of the binary target. MCC is relatively well-suited for finding such subgroups.

6 Conclusions

We develop quality measures for Subgroup Discovery in a setting with a compound target. On the one hand, a binary target indicates membership of a group, and on the other hand, a real-valued target indicates a commodity. The most straightforward commodity for which this setting is applicable is a dataset involving money. One can think for instance of patients in a medical dataset that have a total amount of money claimed with their insurance company. To detect malicious claiming behavior (or even fraud) on a level that is interesting to the insurer from an economic point of view, i.e. not on the level of single patients, but on the aggregated level of medical practitioners such as pharmacies, the additional binary target comes into play, indicating whether a patient received care from a given provider. In this paper, we develop quality measures for this setting, whose values have an intuitive meaning in terms of the commodity: if a quality measure can be interpreted in terms of money recovered when the corresponding subgroup is dealt with, this enables the domain expert

to make informed executive decisions. To this end, we introduce six new quality measures, whose relative merits are discussed in Section 4.1.

Although a monetary-valued dataset is a prime example of the problem settings one could approach with these quality measures, the setup is more generic, allowing any kind of real-valued commodity. In fact, for any commodity we have in our dataset for which it is somehow desirable to limit its use, the associated qualities of subgroups found with these measures will always have a more or less straightforward interpretation in terms of the commodity. We illustrate this with experiments on a dataset detailing the number of court days necessary to reach a verdict in a court case. Found subgroups can be delivered to a domain experts with an associated number of court days that could potentially be salvaged, which is valuable information for making executive decisions.

From our primary application domain, health care, we conclude that subgroups are more interesting because more money is involved. Even very small subgroups were found using the cost-based quality measures. These subgroups would not be considered interesting according to commonly-used Subgroup Discovery quality measures, but such small groups are more actionable and easier to investigate than larger subgroups. In case such a subgroup consists of clear outliers, money can be claimed back. Other subgroups that are found with cost-based quality measures are often not directly related to fraud, however they do indicate excessive claiming behavior. Such new subgroups can function as a benchmark statistic, to compare claiming behavior of medical practitioners. When the subgroup is new to the insurance company, this is very valuable information. Medical practitioners that are identified to claim differently can be made aware of this. Usually this will result in a lower amount claimed by the practitioner in the future.

References

Atzmueller, M., & Lemmerich, F. (2009). Fast subgroup discovery for continuous target concepts. In J. Rauch, W. Raś, Z., P. Berka, T. Elomaa (Eds.), *Foundations of intelligent systems. Lecture notes in computer science* (Vol. 5722, pp. 35–44). Berlin: Springer.

Bay, S., & Pazzani, M. (2001). Detecting group differences: mining contrast sets. *Data Mining and Knowledge Discovery*, *5*(3), 213–246.

Chan, R., Yang, Q., Shen, Y.-D. (2003). Mining high utility itemsets. In *Third IEEE international conference on data mining, 2003* (pp. 19–26). IEEE.

Dong, G., & Li, J. (1999). Efficient mining of emerging patterns: discovering trends and differences. In *Proceedings of KDD '99* (pp. 43–52). New York.

Elkan, C. (2001). The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence* (Vol. 17, pp. 973–978). Citeseer.

Grosskreutz, H. (2010). Cascaded subgroups discovery with an application to regression. In *LeGo-08 - from local patterns to global models: ECML/PKDD-08 workshop* (p. 16).

Grosskreutz, H., Rüping, S., Wrobel, S. (2008). Tight optimistic estimates for fast subgroup discovery. In W. Daelemans, B. Goethals, K. Morik (Eds.), *Machine learning and knowledge discovery in databases. Lecture notes in computer science* (Vol. 5211, pp. 440–456). Berlin: Springer.

Hernández-Orallo, J., Flach, P.A., Ramirez, C.F. (2011). Technical note: towards roc curves in cost space. *CoRR*, ArXiv abs/1107.5930.

Jorge, A.M., Azevedo, P.J., Pereira, F. (2006). Distribution rules with numeric attributes of interest. In J. Fürnkranz, T. Scheffer, M. Spiliopoulou (Eds.), *Knowledge discovery in databases: PKDD 2006. Lecture notes in computer science* (Vol. 4213, pp. 247–258). Berlin: Springer.

Knobbe, A., & Ho, E. (2006). Pattern teams. In J. Fürnkranz, T. Scheffer, M. Spiliopoulou (Eds.), *Knowledge discovery in databases: PKDD 2006. Lecture notes in computer science* (Vol. 4213, pp. 577–584). Berlin: Springer.

Konijn, R.M., & Kowalczyk, W. (2012). Hunting for fraudsters in random forests. In E. Corchado, V. Snasel, A. Abraham, M. Wozniak, M. Grana, S.-B. Cho (Eds.), *Hybrid artificial intelligent systems. Lecture notes in computer science* (Vol. 7208, pp. 174–185). Berlin: Springer.

Konijn, R.M., Duivesteijn, W., Kowalczyk, W., Knobbe, A. (2013a). Discovering local subgroups, with an application to fraud detection. In J. Pei, V. Tseng, L. Cao, H. Motoda, G. Xu (Eds.), *Advances in knowledge discovery and data mining. Lecture notes in computer science* (Vol. 7818, pp. 1–12). Berlin: Springer.

Konijn, R.M., Duivesteijn, W., Meeng, M., Knobbe, A. (2013b). Cost-based quality measures in subgroup discovery. In *New frontiers in applied data mining - PAKDD 2013 international workshops - QIMIE 2013*.

Lavrač, N., Flach, P., Zupan, B. (1999). Rule evaluation measures: a unifying view. In S. Džeroski, P. Flach (Eds.), *Inductive logic programming. Lecture notes in computer science* (Vol. 1634, pp. 174–185). Berlin: Springer.

Liu, Y., Liao, W.K., Choudhary, A. (2005). A fast high utility itemsets mining algorithm. In *Proceedings of the 1st international workshop on utility-based data mining* (pp. 90–99).

Meeng, M., & Knobbe, A. (2011). Flexible enrichment with cortana (software demo). In *Proceedings Benelearn* (pp. 117–120).

Pieters, B.F.I., Knobbe, A., Džeroski, S. (2010). Subgroup discovery in ranked data, with an application to gene set enrichment. In *Proceedings preference learning workshop (PL 2010)*

Reid, A.A., Tayebi, M.A., Frank, R. (2013). Exploring the structural characteristics of social networks in a large criminal court database. In *Proceedings of the IEEE intelligence and security informatics conference (ISI 2013)* (pp. 209–214).

Wrobel, S. (1997). An algorithm for multi-relational discovery of subgroups. In *Proceedings of PKDD* (pp. 78–87).