Under consideration for publication in Knowledge and Information Systems

Exceptionally Monotone Models — the Rank Correlation Model Class for Exceptional Model Mining[1]

Lennart Downar[2] and Wouter Duivesteijn[3]

[2]Fakultät für Informatik, LS VIII, Technische Universität Dortmund, Germany, lennart.downar@udo.edu; [3]Data Science Lab & iMinds, Universiteit Gent, Belgium, wouter.duivesteijn@ugent.be

Abstract. Exceptional Model Mining strives to find coherent subgroups of the dataset where multiple target attributes interact in an unusual way. One instance of such an investigated form of interaction is Pearson's correlation coefficient between two targets. EMM then finds subgroups with an exceptionally linear relation between the targets. In this paper, we enrich the EMM toolbox by developing the more general rank correlation model class. We find subgroups with an exceptionally monotone relation between the targets. Apart from catering for this richer set of relations, the rank correlation model class does not necessarily require the assumption of target normality, which is implicitly invoked in the Pearson's correlation model class. Furthermore, it is less sensitive to outliers. We provide pseudocode for the employed algorithm and analyze its computational complexity, and experimentally illustrate what the rank correlation model class for EMM can find for you on six datasets from an eclectic variety of domains.

Keywords: Rank correlation; Exceptional Model Mining; Monotonicity; Subgroup Discovery; Data mining

1. Introduction

Identifying where the attributes of your dataset interact in an unusual way is an important component of understanding the underlying concepts that play a role in the dataset at hand. In physics, for example, it is nowadays common to reduce

Received Dec 13, 2015
Revised Jun 28, 2016
Accepted Jul 16, 2016

[1] This paper is an extended version of our paper (Downar and Duivesteijn, 2015) presented at ICDM 2015, which builds upon the work from the Bachelor's thesis (Downar, 2014).

the enormous amount of available data, in order to be able to process it. However, only existing domain knowledge is used to decide which parts of a dataset to keep and which to discard. Discovery of new insights is thus of great interest to researchers in these fields. Finding subsets of a dataset that might be of interest presents thus an important discovering and filtering task. Exceptional Model Mining (EMM) (Leman et al., 2008; Duivesteijn, 2013; Duivesteijn et al., 2016) is a framework dedicated to reporting such subareas of a dataset in a form that can be easily interpreted by a domain expert. The focus lies on providing *understanding*: we do not want to highlight an incoherent set of outliers, but rather define a coherent subgroup in terms of other attributes in the dataset on which exceptional interaction takes place.

Exceptional interaction can come in many different forms. One of the most straightforward forms is the Pearson correlation between two designated target attributes. This form of interaction has been studied in the *correlation model class* for EMM (Leman et al., 2008). With this model class, one can find subgroups of the dataset where the linear relation between two targets is substantially different from that same relation on the complement of the subgroup. In this paper, we introduce another model class studying the interaction of two designated targets, but then in terms of rank correlation (Spearman, 1904; Kendall, 1938). The *rank correlation model class* comes with three advantages over the existing correlation model class: the rank correlation model class does not need the assumption of target normality present in the correlation model class, it is less sensitive to outliers, and the gauged form of interaction is richer. After all, with the rank correlation model class, one can find subgroups of the dataset where the monotone relation between two targets is substantially different from that same relation on the complement of the subgroup, and monotone relations encompass linear relations.

Rank correlation has been employed on an eclectic variety of domains, including bioinformatics (Balasubramaniyan et al., 2005), information retrieval (Yilmaz et al., 2008), recommender systems (Breese et al., 1998), and determining molecular structure by lanthanide shift reagents (Li and Lee, 1980). Finding coherent subgroups of the dataset at hand displaying exceptional interaction between two targets, as measured through rank correlation, should be interesting to practitioners in these fields. For example in particle physics (cf. Section 5.6), it is quite common to have attributes, which are correlated in a monotone way. However the measurement of those variables is not directly possible and thus is done indirectly through reconstruction. Since reconstruction can be prone to errors, finding subsets in a dataset where the relationship deviates, e.g. the expected correlation is not found, would help detecting errors or noise in the reconstruction process. We define quality measures for the rank correlation model class based on Spearman's rank correlation coefficient r_s (Spearman, 1904), and on Kendall's τ_b (Kendall, 1938), and experimentally illustrate the model class and measures on six datasets.

1.1. Main Contribution

The main contribution of this paper is the development of the rank correlation model class for Exceptional Model Mining. In this model class, two attributes of the dataset are identified as the targets; these must be numeric or ordinal. The goal of the model class is to find subgroups representing a schism in monotone re-

lations between the targets: a subgroup is deemed interesting if the monotonicity of the relation between the targets deviates substantially from the monotonicity of the same relation on the complement of the subgroup in the dataset. A collateral contribution is the overview (provided in Section 2.3) of alternative correlation measures available in the literature, highlighting the potential for future research into their underlying statistical theory.

1.1.1. Innovations in the Extended Version

As mentioned in a footnote on the opening page of this paper, this paper is an extended version of our paper (Downar and Duivesteijn, 2015) presented at ICDM 2015. With respect to the original publication, this paper provides the following additional innovations:

- added experiments on two more datasets, increasing the number of datasets from four to six. The new datasets stem from substantially different domains than the other four. The results can be found in Sections 5.2 (South African Heart Disease Study) and 5.3 (Ozone Dataset);
- included more details behind the algorithm that is being used to search for subgroups. Consequently, we have divided the old experimental section into two sections: Section 4 now discusses the experimental setup, and Section 5 discusses the experimental results. The pseudocode for Algorithm 1 is not new; it has been published in (Duivesteijn, 2013; Duivesteijn et al., 2016). The schematic description of how the algorithm works, i.e., Section 4.1, is new. The analysis of the computational complexity of the Algorithm for a general model class has been published in (Duivesteijn, 2013; Duivesteijn et al., 2016). The analysis of its computational complexity for the rank correlation model class (i.e., all of Section 4.2 except for the first paragraph) is new;
- carried out one of the future work extensions that was proposed in the original paper. There we wrote that it would be good to investigate whether the rank correlation model class for EMM was compatible with the GP-Growth algorithm developed in (Lemmerich et al., 2012), and if so, define the corresponding valuation basis. We have concluded that this is not possible; reasons are outlined in Section 4;
- found and fixed a bug in the programming code, which invalidates the results in Tables IIc, IIIc, and IVc of the original paper. Section 4.3 gives a link to the repaired version of the code, and the correct results are given in this paper in Tables 1c, 4c, and 5c;
- broadened the scope of the related work. Section 2 now also discusses related work in slightly less directly related data mining areas. Particularly, work on Conceptual Clustering and Multi-Label Classification is included. We have cited one major work from each of these fields, and referred to other papers where a more extensive overview of the relation between work in these fields and work in Exceptional Model Mining is given;
- consulted a domain expert in the field of experimental physics: a fellow researcher whose research focus revolves around the CERN large hadron collider experiment, and hence a domain expert on the experimental results presented in Section 5.6. While interpretation of individual subgroups required more detailed information behind the data generating process than was available in the documentation (Adam-Bourdarios et al., 2014), the domain expert was

able to illuminate why seeking exceptionally monotone models is particularly interesting for particle physicists, thus contributing to the motivation of the overall paper (cf. Sections 1 and 5.6).

2. Related Work

Pattern mining (Hand et al., 2002; Morik et al., 2005) is the broad subfield of data mining where only a part of the data is described at a time, ignoring the coherence of the remainder. One class of pattern mining problems is *theory mining* (Mannila and Toivonen, 1997), whose goal is finding subsets S of the dataset Ω that are interesting somehow:

$$S \subseteq \Omega \quad \Rightarrow \quad \text{interesting}$$

Typically, not just any subset of the data is sought after: only those subsets that can be formulated using a predefined *description language* $\mathcal{L}$ are allowed. A canonical choice for the description language is conjunctions of conditions on attributes of the dataset. If, for example, the records in our dataset describe people, then we can find results of the following form:

$$\text{Age} \geq 30 \wedge \text{Smoker} = \text{yes} \quad \Rightarrow \quad \text{interesting}$$

Allowing only results that can be expressed in terms of attributes of the data, rather than allowing just any subset, ensures that the results are relatively easy to interpret for a domain expert: the results arrive at his doorstep in terms of quantities with which he should be familiar. A subset of the dataset that can be expressed in this way is called a *subgroup*.

In the best-known form of theory mining, *frequent itemset mining* (Agrawal et al., 1996), the interestingness of a pattern is gauged in an unsupervised manner. Here, the goal is to find patterns that occur unusually frequently in the dataset:

$$\text{Age} \geq 30 \wedge \text{Smoker} = \text{yes} \quad \Rightarrow \quad \text{(high frequency)}$$

The most extensively studied form of *supervised* theory mining is known as *Subgroup Discovery* (SD) (Herrera et al., 2011). Typically, one binary attribute t of the dataset is singled out as the *target*. The goal is to find subgroups for which the distribution of this target is unusual: if the target describes whether the person develops lung cancer or not, we find subgroups of the following form:

$$\text{Smoker} = \text{yes} \quad \Rightarrow \quad \text{lung cancer} = \text{yes}$$

Exceptional Model Mining (EMM) (Leman et al., 2008; Duivesteijn, 2013) can be seen as the multi-target generalization of SD. Rather than singling out one attribute as the target t, in EMM there are several target attributes $t_1, \ldots, t_m$. Interestingness is not merely gauged in terms of an unusual *marginal* distribution of t, but in terms of an unusual *joint* distribution of $t_1, \ldots, t_m$. Typically, a particular kind of unusual *interaction* between the targets is captured by the definition of a *model class*, and subgroups are deemed interesting when their model is exceptional, which is captured by the definition of a *quality measure*.

To illustrate this abstract form of exceptionality, we will flesh out the details of the one existing model class that is particularly relevant in this paper — correlation between two numerical targets (Leman et al., 2008) — in Section 2.1. Other investigated model classes are variance of a single target (Lemmerich

et al., 2012)[4], association between two nominal targets (Duivesteijn et al., 2016), simple linear regression on two targets (Leman et al., 2008), behavior of a hard classifier (Leman et al., 2008), total variation on a contingency table of any size (Moens and Boley, 2014), distance over a multivariate mean model (Moens and Boley, 2014), structure of a Bayesian network on any number of nominal targets (Duivesteijn et al., 2010), linear regression on any number of targets (Duivesteijn et al., 2012a), and SCaPE (Soft Classifier Performance Evaluation) (Duivesteijn and Thaele, 2014).

Notice that the interpretability is a fundamental characteristic of both Subgroup Discovery and Exceptional Model Mining. In these tasks, and hence in this paper, we are not merely interested in *pointing out* parts of the dataset that deviate from the norm; we are interested in *finding reasons why* parts of the dataset deviate from the norm. This sets SD and EMM apart from techniques such as clustering, outlier detection, and anomaly detection, where the focus typically lies on finding a distributional difference on the target space. In EMM, delivering a concise description is just as important as the exceptionality of the target interaction: a distributional target deviation that does not come with an associated description is not interesting from an EMM point of view.

2.1. The Correlation Model Class for EMM

Suppose that there are two target attributes: a person's height (t_1), and the average height of his/her grandparents (t_2). We may be interested in Pearson's standard correlation coefficient between t_1 and t_2; we then say we study EMM with the *correlation model class* (Leman et al., 2008). Given a subset $S \subseteq \Omega$, we can estimate the correlation between the targets within this subset by the sample correlation coefficient. We denote this estimate by r^S. Now we can define the following quality measure (adapted from (Leman et al., 2008)):

$$\varphi_{\text{abs}}(S) = \left| r^S - r^{\Omega \setminus S} \right|$$

EMM then strives to find subgroups for which this quality measure has a high value. Effectively, we search for subgroups coinciding with an exceptional correlation between a person's height and his/her grandparents' average height:

$$\text{Subgroup } S \quad \Rightarrow \quad \left| r^S - r^{\Omega \setminus S} \right| \text{ is high}$$

There is an undeniable elegance in the simplicity of the correlation model class. The subsequent three sections discuss its drawbacks.

2.1.1. Assumption of Normality?

Whether or not the use of Pearson's correlation coefficient implies the assumption that the targets in question are normally distributed, is a very subtle issue that

[4] Whether this model class falls under the *spirit* of EMM is debatable; having only a single target prohibits investigating target interaction. Careful reading of EMM literature (Leman et al., 2008; Duivesteijn et al., 2016) reveals that the framework (accidentally) allows model classes where $m = 1$. Hence, we cannot formally say that this model class doesn't fall under the *letter* of EMM. Since the authors of (Lemmerich et al., 2012) introduced this model class as an EMM instance, and we cannot formally reject it as such, we adopt it into the EMM canon.

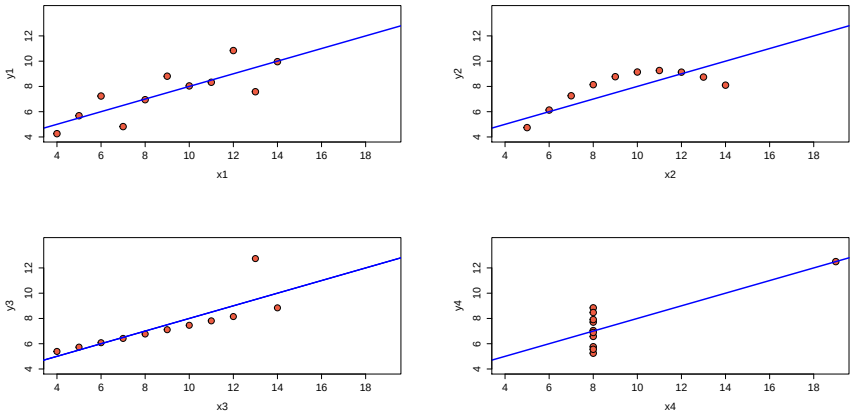


Fig. 1. Anscombe's quartet.

is open for debate. Kowalski's experimental evaluation (Kowalski, 1972) shows, however, that the distribution of r is sensitive to non-normality:

"normal correlation analyses should be limited to situations in which (X, Y) is (at least very nearly) normal" (Kowalski, 1972, Section 6).

Without the normality assumptions, many statistical tests on r become meaningless or at least hard to interpret. Considering that normality cannot be assumed for many real-life examples and datasets, it is questionable if Pearson's r is still a suitable measure. The normality assumption therefore limits the scope of application for this model class.

2.1.2. Sensitivity to Outliers

Pearson's correlation coefficient is well-known to be easily affected by outliers. This has been eminently illustrated by Anscombe's quartet (Anscombe, 1973), displayed in Figure 1, which consists of four different datasets with almost identical basic statistical properties (e.g., all four share the same Pearson coefficient). Francis Anscombe presented it to emphasize the importance of visualization when analyzing data. All four datasets have a Pearson correlation of 0.816. The effect of outliers can be seen quite clearly in sets 3 and 4, two datasets featuring two substantially different relations between the two displayed variables.

2.1.3. Linear Versus Monotone

The third point is not necessarily a drawback of the correlation model class per se, but more a point on which a newly developed model class could contribute. Pearson's correlation focuses on linear relations between the two targets. Hence, EMM with the correlation model class will find subgroups where this linear relation is exceptional. Rank correlation measures focus on the richer class of monotone relations between the two targets. Hence, EMM with a rank correlation model class will find subgroups where the monotone relation between the targets is exceptional. The class of monotone relations encompasses the class of linear relations. Hence, the types of target interaction for which EMM can find

exceptional subgroups, are less diverse for the existing correlation model class than they are for a rank correlation model class. This also implies, however, that the correlation model class is more specialized, while a rank correlation model class is more generalized. One can have domain-specific reasons to prefer the one over the other, and hence we would absolutely not claim that a rank correlation model class makes the correlation model class redundant. The correlation model class serves a clear purpose, but a rank correlation model class allows more to be explored.

2.2. Tasks Related to EMM

Local Pattern Mining tasks that are similar to SD are Contrast Set Mining (Bay and Pazzani, 2001) and Emerging Pattern Mining (Dong and Li, 1999). Both these tasks do not consider multiple target attributes simultaneously, and do not directly model unusual interactions. The relation between Contrast Set Mining, Emerging Pattern Mining, and Subgroup Discovery was studied extensively in (Kralj Novak et al., 2009). Explicitly seeking a deviating model over a target is performed in Distribution Rules (Jorge et al., 2006), where there is only one numeric target, and the goal is to find subgroups on which the target distribution over the entire target space is the least fitting to the same distribution on the whole dataset. This can be seen as an early instance of EMM with only one target. However, there is no multi-target interaction. Umek et al. (Umek and Zupan, 2011) do consider SD with multiple targets. They approach the attribute partition in the reverse way of EMM: candidate subgroups are generated by agglomerative clustering on the targets, and predictive modeling on the descriptors strives to find matching descriptions. This work does not allow freely expressing when target interaction is unusual. Redescription Mining (Galbrun and Miettinen, 2012) seeks multiple descriptions inducing the same subgroup. This models unusual interplay, but on the descriptor space rather than the target space.

Arguably, in striving to find descriptions of groups for which certain attribute values are distributed differently from the rest of the data, EMM finds kindred spirits in the fields of conceptual clustering (Fisher and Langley, 1986) and multi-label classification (Tsoumakas and Katakis, 2007). Due to differences in scope and capabilities of these methods, it is beyond scope of this paper to discuss these relations in full here; the relation between EMM and clustering methods is fleshed out further in Section 7.3 of (Duivesteijn et al., 2016), and methods on the crossroads of EMM and multi-label classification are discussed in Section 8 of (Duivesteijn et al., 2012b).

2.3. Alternative Correlation Measures

From Section 1.1 onwards, we will only consider correlation measures for which a straightforward adaptation of a well-known statistical test (cf. Section 3.3) exists. This enables the formulation of quality measures for EMM defined in terms of p-values corresponding to that statistical test. Thus, the quality measures that we will define in Section 3.3 have a solid basis in statistics, and come with the additional benefit that the interpretation of their values is straightforward. Alternative correlation measures exist for which, to the best of our knowledge, no statistical theory is available that would allow us to compare results on different

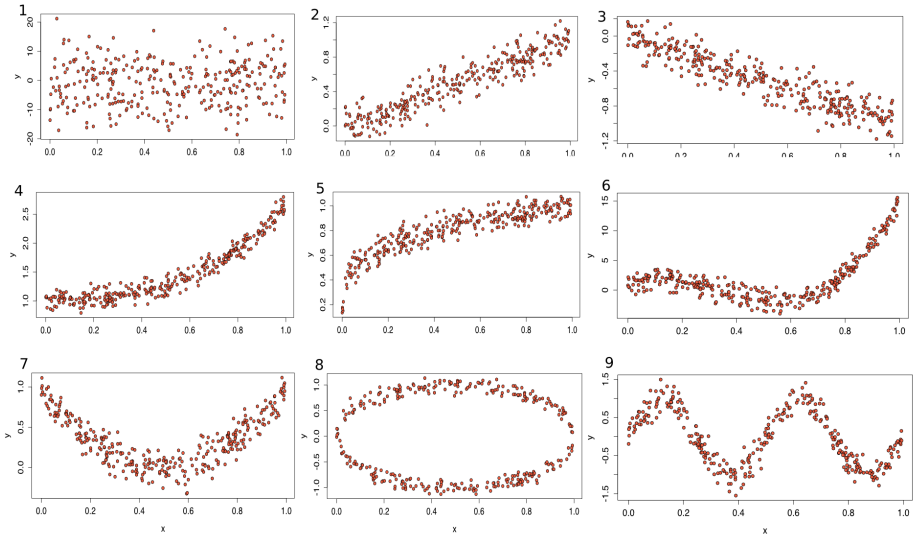


Fig. 2. Target relations, detectable by alternative correlation measures.

data samples. Developing the statistical theory necessary to base a mathematically well-supported EMM quality measure on these correlation measures is beyond the scope of this paper, but these alternative correlation measures are, obviously, relevant related work, and they are described as such in the remainder of this section.

A comparison of correlation measures has been given in (Clark, 2013). Apart from Pearson's r and Spearman's r_s, Clark examines three other measures, which promise to measure relations beyond linear and monotone behavior. Examples for datasets that exhibit such behavior can be seen in Figure 2, where Pearson would only be able to detect patterns 2 and 3, and to some extent 4 and 5.

Contrary to sample correlation coefficients such as Spearman's r_s, Kendall's τ, and Pearson's r, Hoeffding (Hoeffding, 1948) developed a test of independence that can be used to detect a much broader class of relations beyond monotone association. Hoeffding's statistic, denoted by D, is non-parametric and, similar to Spearman and Kendall, based on ranks. A similar statistic proposed by Blum et al. (Blum et al., 1961) can be used as a large-sample approximation for D (Hollander and Wolfe, 1999).

Distance correlation (*dCor*) has been introduced by Székely et al. (Székely et al., 2007) to widen the limited scope of the Pearson correlation coefficient towards non-linear relations. It is based on distance matrices for the target variables and can take values between 0 and 1. According to Clark (Clark, 2013), a ranked-based version of *dCor* could also be incorporated.

Reshef et al. (Reshef et al., 2011) have developed the Maximal Information Coefficient (*MIC*). It is based on the concepts of *Entropy* and *Mutual Information* from information theory. Clark points out that *MIC* could be seen as the continuous variable counterpart to mutual information. Similar to *dCor*, *MIC* takes on values between 0 and 1, with zero indicating independence.

2.3.1. Evaluation

After comparing these alternatives on several non-linear relations, Clark notes:
"Hoeffding's D only works in some limited scenarios." (Clark, 2013)
In the experiments, D did pick up some of the non-linear relations (e.g., a quadratic relation or a circle pattern), but the computed values were relatively small (mean ranging from 0 to 0.1), which was exacerbated when noise was added to the data (mean ranging from 0 to 0.02). Even though D does pick up some non-linear relations, due to the small values one cannot get a good sense of the measured dependence.

dCor and *MIC* performed better at finding relations beyond linear ones. However, when noise is present, both become less predictable and the strength of detected associations can vary strongly. Thus, *dCor* and *MIC* might provide alternatives to more classical approaches for picking up a wider variety of relations, but they are not perfect either. Some additional problems with *MIC* are described by Kinney and Atwal (Kinney and Atwal, 2014). Consequently, Clark concludes:
"[we] still need to be on the lookout for a measure that is both highly interpretable and possesses all the desirable qualities we want." (Clark, 2013)

2.3.2. Other Approaches

As pointed out by Clark (Clark, 2013), the development of satisfying general dependence measures that go beyond simple forms of relations is still far from finished. Other approaches therefore have been introduced in recent years. Gretton et al. (Gretton et al., 2005) developed the Hilbert-Schmidt Independence Criterion (*HSIC*), which is based on an empirical estimate of the Hilbert-Schmidt norm of the cross-covariance operator. Lopez-Paz et al. (Lopez-Paz et al., 2013) proposed the Randomized Dependence Coefficient (*RDC*), which is an estimate of the Hirschfeld-Gebelein-Rényi Maximum Correlation Coefficient (*HGR*) defined by Gebelein (Gebelein, 1941) in 1941. However, HGR is not computable and thus represents only an abstract concept.

3. The Rank Correlation Model Class for EMM

In the rank correlation model class for EMM, we assume a dataset Ω, which is a bag of N records of the form $r = \left(a^1, \ldots, a^k, x, y\right)$. We call $\left\{a^1, \ldots, a^k\right\}$ the *descriptive attributes* or *descriptors*. Their domain is unrestricted. The other two attributes, x and y, are the *target attributes* or *targets*. Their domain should at least be ordinal; for simplicity of notation we will assume that they are real-valued in the remainder of this paper, but the minimum requirement is that one should be able to rank the values of x and y. If we need to distinguish between particular records of the dataset, we will do so by subscripted indices: r_i is the i^{th} record, x_i and y_i are its values for the targets, and a_i^j is its value for the j^{th} descriptor. When we are considering a particular subgroup $S \subseteq \Omega$, we will denote the number of records belonging to the subgroup by n.

3.1. Spearman's Rank Correlation Coefficient

Spearman's rank correlation coefficient (usually denoted by ρ but also by r_s; we will use r_s to avoid confusion with the population correlation coefficient) has

been developed by Charles Spearman (Spearman, 1904). It uses the difference between rankings of a pair x_i and y_i as a statistic to measure rank correlation:

$$r_s = 1 - \frac{6 \sum_i d_i^2}{n(n^2-1)} \tag{1}$$

where d_i is the difference between the ranks of x_i and y_i If no ties are present, this is equivalent to computing the Pearson coefficient over the ranks of the data. With R_i and S_i corresponding to the ranks of x_i and y_i and $\bar{R}$ and $\bar{S}$ describing their respective means, we can thus write:

$$r_s = \frac{\sum \left(R_i - \bar{R}\right)\left(S_i - \bar{S}\right)}{\sqrt{\sum \left(R_i - \bar{R}\right)^2 \sum \left(S_i - \bar{S}\right)^2}} \tag{2}$$

In case of ties, Conover (Conover, 1971) suggest using Equation (2). If the number of ties is at most moderate, Equation (1) will still function as a good approximation and should be preferred due to its computational simplicity.

3.2. Kendall's Tau

Where Spearman's r_s uses the difference of ranks in individual pairs, Kendall's τ (Kendall, 1938) defines a statistic based on the agreement (concordances) of ranks to measure the correlation of a sample, making it less sensitive to outliers. A pair of observations $(x_i, y_i), (x_j, y_j)$ is said to be *concordant* if $(x_i < x_j) \wedge (y_i < y_j)$ or $(x_i > x_j) \wedge (y_i > y_j)$. The pair is said to be *tied* if $x_i = x_j$ or $y_i = y_j$, and it is said to be *discordant* otherwise.

The total number of pairs that can be constructed for a sample size of n is $M = \binom{n}{2} = n(n-1)/2$. For the following coefficients we define a number of values:

C = number of concordant pairs
D = number of discordant pairs
T_x = number of pairs tied only on the x-value
T_y = number of pairs tied only on the y-value
T_{xy} = number of pairs tied both on the x- and y-value

Hence, we can decompose M into: $M = C + D + T_x + T_y + T_{xy}$. Many correlation measures exist that involve the numerator C-D but differ in the normalizing denominator. We will take the most widely applied version of Kendall's measure, τ_b, as representative for this class of measures.

Kendall's τ_b accounts for ties by normalizing with a term expressing the geometric mean between the number of pairs untied on the x-value and untied on the y-value:

$$\tau_b = \frac{C - D}{\sqrt{(C + D + T_x)(C + D + T_y)}}$$

3.3. Encapsulating Spearman's r_s and Kendall's τ_b in Quality Measures for Exceptional Model Mining

Although it is common to view the presented rank correlation coefficients as alternatives to Pearson's coefficient, this notion has little mathematical justifi-

cation, as we can see by the definitions in the preceding two sections. We will therefore keep in mind that they should rather be regarded as measures for different kinds of associations.

Rank correlation naively inspires two simple quality measures by way of direct comparison of the correlation coefficients for the subgroup and its complement. The bigger the difference between a subgroup and its complement, the more interesting the subgroup:

$$\varphi_{\text{abs\_}r_s}(S) = \left| r_s^S - r_s^{\Omega\setminus S} \right| \qquad \varphi_{\text{abs\_}\tau_b}(S) = \left| \tau_b^S - \tau_b^{\Omega\setminus S} \right|$$

These quality measures do not make any assumptions on the distribution of the targets. Their drawback is that the size of the subgroups is not taken into account. Hence, they are prone to overfitting: it should be relatively easy to find small subgroups that display extreme rank correlation values, but these subgroups are not necessarily interesting. A straightforward solution to this problem is to determine whether the difference in rank correlation is statistically significant. Ideally, we would want to test:

$$H_0 : \rho_1 = \rho_2 \quad \text{against} \quad H_1 : \rho_1 \neq \rho_2$$

for two groups of data (e.g., a subgroup and its complement). A standard procedure to test for difference between independent Pearson correlations is to perform a Fisher z-transformation on both values to make them normally distributed:

$$z = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right) = \operatorname{arctanh}(r)$$

The transformed value z is normally distributed with variance $\text{var}_z(S) = \frac{1}{n-3}$.

We can then treat the difference between the transformed values as a random normal variable, with mean zero and variance $\text{var}_{\rho_1-\rho_2}(S_1, S_2) = \frac{1}{n_1-3} + \frac{1}{n_2-3}$. By comparing it with a standard normal distribution, a p-value for the difference can then be calculated. Even if the distribution of the z-score is not strictly normal, it tends to normality rapidly as the sample size increases for any value of the actual population correlation coefficient (Fisher, 1970). Hence, the authors of (Leman et al., 2008) defined one minus the p-value from this statistical test to be the quality measure φ_{scd} for the correlation model class.

Fieller et al. (Fieller et al., 1957) have transferred this approach to rank correlation, enabling comparisons of Kendall's τ_b and Spearman's r_s. His experiments suggested the following variances for the transformed values:

$$\text{var}_{r_s}(S) = \frac{1.06}{n-3} \quad \text{and} \quad \text{var}_{\tau_b}(S) = \frac{0.437}{n-4}$$

Accordingly, we define two quality measures for the rank correlation model class: the Fieller-Kendall quality measure φ_{fk} and the Fieller-Spearman quality measure φ_{fs}. Let z_{r_s} and z_{τ_b} denote the Fisher z-transformed values for r_s and τ_b, respectively. Then, both

$$z^*_{r_s} = \frac{z^S_{r_s} - z^{\Omega\setminus S}_{r_s}}{\sqrt{\text{var}_{r_s}(S) + \text{var}_{r_s}(\Omega\setminus S)}}$$

and

$$z^*_{\tau_b} = \frac{z^S_{\tau_b} - z^{\Omega\setminus S}_{\tau_b}}{\sqrt{\mathrm{var}_{\tau_b}(S) + \mathrm{var}_{\tau_b}(\Omega\setminus S)}}$$

approximately follow a standard normal distribution under H_0. Mirroring the development of φ_{scd} in (Leman et al., 2008), we take one minus the computed p-values for $z^*_{r_s}$ and $z^*_{\tau_b}$ as our quality measures φ_{fs} and φ_{fk}, respectively, so that their values range between zero and one, and higher values indicate subgroups that are more exceptional.

3.4. Limitations

In Section 2.1, we identified three limitations of the existing correlation model class. In this section, we revisit those limitations for the rank correlation model class.

The quality measures introduced in the previous section do not suffer as strongly from the sensitivity to outliers as highlighted in Section 2.1.2, and they capture the monotone relations that were discussed to be desirable in Section 2.1.3. Recall that the third limitation of the Pearson correlation coefficient, as identified in Section 2.1.1, is that it assumes a normal distribution over the targets. The rank correlation measures presented in Sections 3.1 and 3.2, as well as the naive quality measures $\varphi_{\mathrm{abs\_}r_s}$ and $\varphi_{\mathrm{abs\_}\tau_b}$, do not have this assumption. However, indirectly it comes into play again when applying the slight modifications of the Fisher z-transformation presented in (Fieller et al., 1957) (which is relevant for the more sophisticated quality measures φ_{fs} and φ_{fk}), because these again assume a normal distribution of the underlying population. However, Fieller argues that this might not be a necessary assumption: "The results [...] can clearly be extended to a much wider class of parental distributions". His experiments support that this assumption is reasonable, but since his test only included datasets having between 10 and 50 samples, he notes that for bigger samples this "is a field in which further investigation would be of considerable interest" (Fieller et al., 1957, page 3). Remarkably, to the best of our knowledge, in the half-century since this paper was published, no further investigation has occurred.

4. Experimental Setup

Lemmerich et al. have developed an exhaustive algorithm for Exceptional Model Mining: GP-Growth (Lemmerich et al., 2012). This algorithm captures all information that is relevant for the computation of the quality measure into a concept called *valuation basis*. This valuation basis is then stored in a GP-tree (Lemmerich et al., 2012), exactly as the frequencies are stored in an FP-tree (Han et al., 2000). The efficiency of the FP-Growth algorithm can be leveraged for the GP-Growth algorithm, but only if the employed model class and quality measure satisfy a specific constraint:

"[there must be] a parallel single-pass algorithm with sublinear memory requirements to compute the model from a given set of instances [...]" (Lemmerich et al., 2012, Theorem 1)

To the best of our knowledge, no such algorithm exists; we do not see how one could get around first making a pass over the data to replace the raw values

of x and y into ranks, and subsequently making another pass over the data to compute correlations between ranks. Hence, we turn to heuristic search.

4.1. The Employed Search Algorithm

Algorithm 1 Beam Search for Top-q Exceptional Model Mining (Duivesteijn, 2013; Duivesteijn et al., 2016)

Input: Dataset Ω, quality measure φ, refinement operator η, beam width w, beam depth d, result set size q, Constraints $\mathcal{C}$
Output: PriorityQueue resultSet

```
1 : candidateQueue ← new Queue;
2 : candidateQueue.enqueue({});                    ▷ Start with empty description
3 : resultSet ← new PriorityQueue(q);
4 : for (Integer level ← 1; level ≤ d; level++) do
5 :     beam ← new PriorityQueue(w);
6 :     while (candidateQueue ≠ ∅) do
7 :         seed ← candidateQueue.dequeue();
8 :         set ← η(seed);
9 :         for all (desc ∈ set) do
10 :            quality ← φ(desc);
11 :            if (desc.SATISFIESALL(C)) then
12 :                resultSet.insert_with_priority(desc,quality);
13 :                beam.insert_with_priority(desc,quality);
14 :    while (beam ≠ ∅) do
15 :        candidateQueue.enqueue(beam.get_front_element());
16 : return resultSet;
```

Early in Section 3, we defined the domain of the descriptive attributes to be unrestricted. We think it is important for the general applicability of EMM model classes to allow the user to run it on datasets with as wide a range of attributes as possible. Hence, apart from the two targets (which, in the rank correlation model class, are compelled to be ordinal or real-valued), all attributes can be binary, nominal, and even real-valued. Accommodating for this, however, restricts the scope of our search algorithm. For our experiments in this paper, we use the top-q Exceptional Model Mining beam search algorithm introduced in (Duivesteijn, 2013, Algorithm 1, page 19) and also described in (Duivesteijn et al., 2016). We reproduce the algorithm here in pseudocode, as Algorithm 1.

Beam Search is a heuristic search algorithm that considers candidate subgroups in a general-to-specific order. At the core of the algorithm lies the *refinement operator* η, which controls how a seed subgroup can be refined to generate a new set of more specialized candidate subgroups. In this paper, we use the canonical description language $\mathcal{L}$ of conjunctions of conditions on attributes of the dataset. This description language suggests a straightforward choice for the refinement operator η. Suppose that η is fed a seed subgroup whose description is a conjunction of n conditions. It will then return a set of subgroups whose description is a conjunction of $n+1$ conditions. The first n conditions of each returned description are identical to those of the seed subgroup. The last condition is different for each returned description, and the full set of these conditions spans

all conditions on all attributes of the dataset that make sense, i.e. it must ensure that the newly generated subgroups are proper subsets of the seed subgroup. For a more formal definition of this choice for η, see (Duivesteijn, 2013, Section 4.1). Initially, the refinement operator is seeded with the empty description, a conjunction over zero conditions, which corresponds to the subgroup covering the entire dataset.

Having defined η, the Beam Search algorithm is largely controlled by two user-set parameters; the *beam width* w and the *search depth* d. The first, w, determines how many subgroups are to be refined on each level of the search. On every level, we select the top-w subgroups (as evaluated by the quality measure φ), and these subgroups are used as the seeds for the next level. Hence, the parameter w controls where the algorithm finds itself on the axis between a purely greedy approach ($w = 1$) and an exhaustive approach ($w \to \infty$). A rule of thumb is that reasonable settings of w lie between 10 and 100; the lower end of that scale might lead to underexploration of the search space but makes the algorithm run quickly, which the higher end of that scale typically explores the search space more than thorough enough while being at risk of returning a redundant result set. The second parameter, d, is an upper bound on how many levels of the search are run. Hence, every resulting subgroup will be described as a conjunction of *at most* d conditions on attributes. Setting d to a reasonable level keeps the algorithm runtime in check, while also guaranteeing that the resulting subgroups remain interpretable.

Parameters of the algorithm that haven't been introduced yet are q, the user-specified number of subgroups the algorithm should return, and $\mathcal{C}$, which is a set of constraints a domain expert could come up with. Exceptional Model Mining delivers results in a language that is relatively easy for a domain expert to understand. Therefore, we find it important to provide a means that lets the domain expert tailor the algorithm output to their needs. From a computer science point of view, this set of constraints is typically not very demanding, and for all practical purposes, we will ignore it in the remainder of this paper.

4.2. Computational Complexity

The computational complexity of Algorithm 1 has been analyzed (Duivesteijn, 2013; Duivesteijn et al., 2016) to be $\mathcal{O}\left(dwkN\left(c + M(N,m) + \log(wq)\right)\right)$. In this expression, k and N are the number of descriptors and records in the dataset, and w and d are the user-set parameters of the beam search algorithm (where a typical generous setting would be in the order of magnitude of $w = 100$ and $d = 3$). The other two quantities in the expression, c and $M(N,m)$, depend on the chosen model class: c is the cost of comparing two models, and $M(N,m)$ is the cost of learning a model from N records on m targets.

For $M(N,m)$, we have exactly two targets in the rank correlation model class, so we are actually looking at $M(N,2)$. In a naive implementation, one would have to recompute the ranks of x and y for every subgroup under consideration. This requires sorting both vectors, which costs $\mathcal{O}\left(2N \log N\right)$. Afterwards, computing the rank correlations corresponding to both Spearman's r_s (cf. Section 3.1) and Kendall's τ_b (cf. Section 3.2) can be done in linear time, with a single pass over the dataset. Hence,

$$M(N,m) = M(N,2) = \mathcal{O}\left(2N \log N + N\right) = \mathcal{O}\left(N \log N\right)$$

For c, we need to extract the quality measure values out of the statistics available so far. Taking the statistics from the computations involved in computing the rank correlations in the $M(N, m)$ step, we can perform the necessary computations (cf. Section 3.3) in constant time (taking into account that quantities like the subgroup size are available from preceding algorithm steps). Hence,

$$c = \mathcal{O}(1)$$

Plugging these components in the EMM framework algorithm, we find that the computational complexity of beam search for Top-q Exceptional Model Mining with the rank correlation model class is:

$$\begin{aligned}\mathcal{O}\left(dwkN\left(c + M\left(N, m\right) + \log\left(wq\right)\right)\right) &= \mathcal{O}\left(dwkN\left(1 + N\log N + \log\left(wq\right)\right)\right) \\ &= \mathcal{O}\left(dwkN\left(N\log N + \log\left(wq\right)\right)\right) \quad (3) \\ &= \mathcal{O}\left(dwkN^2\log N\right) \quad (4)\end{aligned}$$

When moving from Equation (3) to Equation (4), we use the fact that generous settings for the paramaters w and q would be $w = q = 100$, which would make $\log(wq) < 14$. On the other hand, $N\log N > 14$ from $N = 6$ onwards, which would make for a tiny dataset indeed. Hence, we use the fact that $N\log N \gg \log(wq)$ for datasets which are not unreasonably small.

4.3. Implementation

We have implemented our work within the RapidMiner analytics platform (Mierswa et al., 2006). The code of the RapidMiner extension, encompassing the rank correlation model class, the original correlation model class, and the top-q Exceptional Model Mining beam search algorithm, is available online at `https://bitbucket.org/lennardo/rancor-emm` .

5. Experimental Results

To put the model class to the test, we perform experiments to find subgroups with the new quality measures φ_{fs} and φ_{fk}, and compare the results with subgroups found with the corresponding quality measure φ_{scd} in the original correlation model class (Leman et al., 2008). We have performed experiments on six datasets, two of which stem from the UCI machine learning repository (Lichman, 2013). In the following sections, we present results of experiments on the Windsor Housing dataset (Anglin and Gençay, 1996), the South African Heart Disease Study dataset (Rousseauw et al., 1983), the Ozone dataset (Hastie et al., 2010), the Contraceptive Method Choice (CMC) dataset (Lim et al., 2000; Lichman, 2013), the Iris dataset (Fisher, 1936; Lichman, 2013), and the real-life Higgs Boson Machine Learning Challenge dataset (Adam-Bourdarios et al., 2014).

Notice that the subgroups reported in Tables 1c, 4c, and 5c are different from the ones reported in (Downar and Duivesteijn, 2015, Tables IIc, IIIc, and IVc). A bug in the code corrupted those results; the subgroups reported in this paper are the correct ones.

Table 1. Windsor Housing: top-3 subgroups found with each of the correlation variants. The variable names have the following meaning (every records is one house). *fb*: number of full bathrooms. *drv*: does it have a driveway? *sty*: number of stories excluding basement. *bdms*: number of bedrooms. *rec*: does it have a recreational room? *ca*: does it have central air conditioning? *ghw*: does it use gas for hot water heating?

Subgroup	φ_{scd}	r	n
fb $\leq 2 \wedge$ drv $= 1 \wedge$ sty ≤ 2	0.99993	0.4740	383
bdms $\geq 3 \wedge$ rec $= 1 \wedge$ drv $= 1$	0.99992	0.1186	77
fb $\geq 2 \wedge$ rec $= 1 \wedge$ drv $= 1$	0.99989	-0.0894	35

(a) Pearson's r.

Subgroup	φ_{fs}	r_s	n
fb $\geq 2 \wedge$ rec $= 1 \wedge$ drv $= 1$	0.9999823	-0.1385	35
fb $\leq 1 \wedge$ drv $= 1 \wedge$ ca $= 0$	0.9999821	0.4319	247
fb $\geq 2 \wedge$ rec $= 1 \wedge$ bdms ≥ 3	0.9999781	-0.0932	36

(b) Spearman's r_s.

Subgroup	φ_{fk}	τ_b	n
bdms $\geq 3 \wedge$ rec $= 1 \wedge$ drv $= 1$	0.9999826687622296	0.072	77
bdms $\geq 3 \wedge$ rec $= 1 \wedge$ ca $= 0$	0.999925710197723	-0.0071	38
bdms $\geq 3 \wedge$ drv $= 0 \wedge$ ghw $= 0$	0.9998984860436247	0.0618	52

(c) Kendall's τ_b.

5.1. Windsor Housing

The Windsor Housing dataset contains 546 samples of houses that were sold in Windsor, Canada in 1987. Each sample consists of 12 attributes such as the lot size, the price at which the house was sold, number of bathrooms, and whether the house was located in a preferable area. The results for the Spearman (Table 1b) and Pearson (Table 1a) measures confirm the experiments performed on the Windsor Housing dataset in (Leman et al., 2008), as both return the subgroup:

$$S_0 : fb \geq 2 \wedge rec = 1 \wedge drv = 1$$

The subgroup S_0 encompasses 35 houses that have a driveway, a recreation room and at least two bathrooms. Leman et al. (Leman et al., 2008) reason that S_0 might describe "houses in the higher segments of the market where the price of a house is mostly determined by its location and facilities. The desirable location may provide a natural limit on the lot size, such that this is not a factor in the pricing." The subgroup S_0 occurs as third-ranked subgroup in the Pearson experiment and top-ranked in the Spearman experiment, but not in the top of the Kendall experiment. This behavior will remain constant over the following experiments: with Spearman rank correlation we find subgroups similar to the ones found with Pearson correlation, but Kendall rank correlation finds different results. In the Windsor Housing data, we see that the first measures focus on houses featuring a driveway, whereas Kendall focuses on houses featuring many bedrooms, which are large in a different manner.

Table 2. South Africa Heart Disease Study: top-3 subgroups found with each of the correlation variants.

Subgroup	φ_{scd}	r	n
age $\leq$ 25 $\wedge$ alcohol $\leq$ 2.42	0.9999999966	-0.0039	45
age $\leq$ 19 $\wedge$ typea $\leq$ 59	0.9999999965	-0.041	41
age $\leq$ 19 $\wedge$ ldl $\leq$ 3.98	0.999999987	-0.0138	41

(a) Pearson's r.

Subgroup	φ_{fs}	r_s	n
age $\leq$ 25 $\wedge$ alcohol $\leq$ 2.42	0.9999958	0.2137	45
age $\leq$ 25 $\wedge$ typea $\leq$ 56.0	0.99988	0.3286	43
age $\leq$ 25 $\wedge$ ldl $\leq$ 3.28	0.99976	0.3441	43

(b) Spearman's r_s.

Subgroup	φ_{fk}	τ_b	n
age $\leq$ 25 $\wedge$ alcohol $\leq$ 2.42	0.999984	0.1659	45
age $\leq$ 19 $\wedge$ famhist = Absent	0.999863	0.2119	43
age $\leq$ 19 $\wedge$ sbp $\leq$ 136	0.9995233	0.2418	43

(c) Kendall's τ_b.

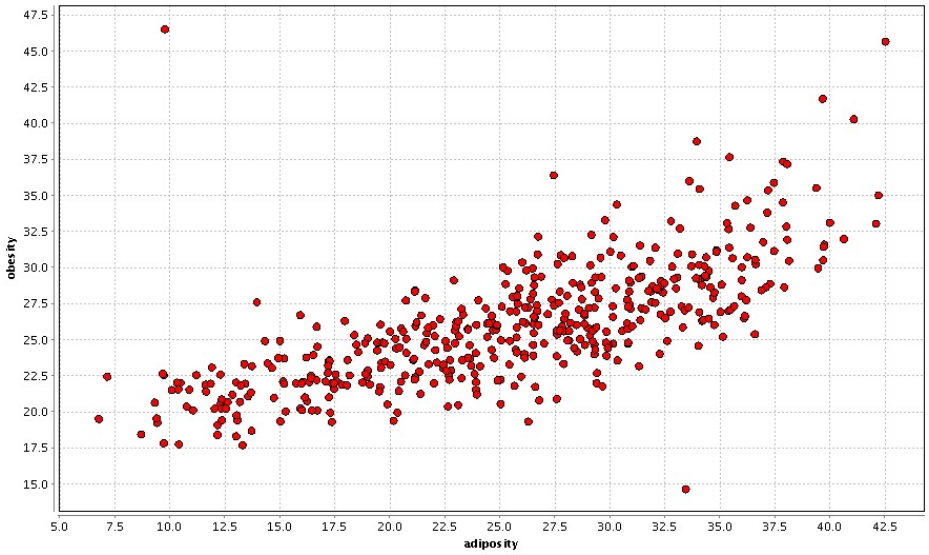
5.2. South African Heart Disease Study

This dataset consists of 462 retrospective samples of males from a heart-disease high-risk region of the Western Cape, South Africa. It contains attributes such as *alcohol* consumption, *age* or the systolic blood pressure (*sbp*). Some of these men have been diagnosed with coronary heart disease, indicated by the attribute *chd*. This dataset is an excerpt from a larger dataset, described in (Rousseauw et al., 1983).

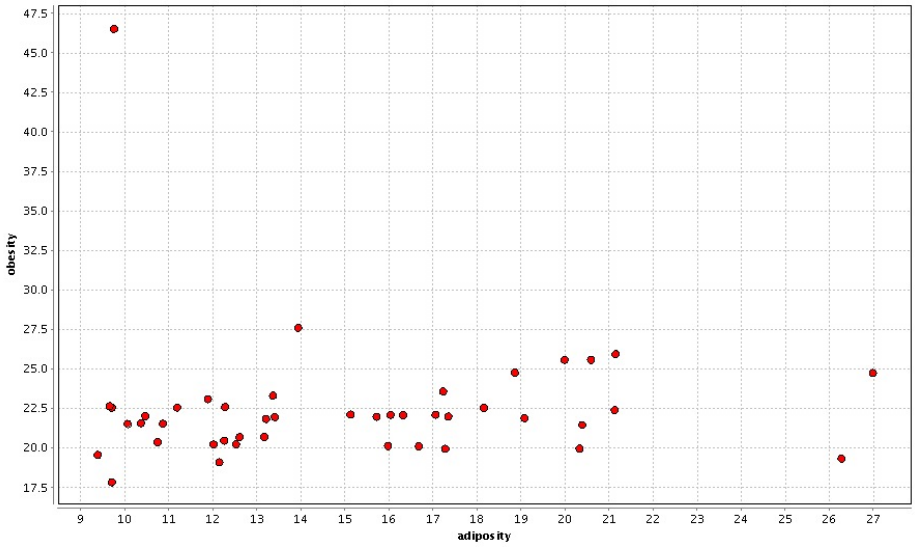
If we plot adiposity (a measure of the body fat percentage) against obesity (in terms of BMI, a weight-to-height ratio), we can observe a monotonically increasing relationship between the two values. This is not surprising as an increase in body fat naturally increases the weight and the body-mass-index. The most exceptional subgroups found with the three quality measures are contained in Table 2. However in the best subgroup found by Spearman we can see that the monotonicity must not always be the case. The target distribution within this subgroup, and within the entire dataset, are plotted in Figure 3. Even when disregarding the clear outlier in the top left corner of both figures, a marked difference in trend is visible. A clear monotonic increase takes place on the whole dataset, whereas the subgroup appears to display variance around a flatline. The found group describes young men with low alcohol consumption, which would generally be considered a healthy group.

5.3. Ozone Dataset

This dataset contains 111 measurements of daily ozone concentration (*ppb*), wind speed (mph), daily maximum temperature (Fahrenheit) and solar radiation (langleys) from May to September 1973 in New York. As a high ozone level may be dangerous for humans, it would be interesting to know how the ozone level is affected by different weather factors. Plotting ozone against the other three at-



(a) Entire South African Heart Disease Study dataset.



(b) Subgroup age $\leq 25 \wedge$ alcohol ≤ 2.42.

Fig. 3. Target distribution on subgroups found on the South African Heart Disease Study dataset.

Table 3. Ozone dataset: top-3 subgroups found with each of the correlation variants.

Subgroup	φ_{scd}	r	n
temperature $\geq 65 \wedge$ temperature ≤ 74	0.9944353	-0.12827	26
temperature $\leq 73 \wedge$ temperature ≥ 62	0.99037704	-0.1651	29
temperature $\leq 77 \wedge$ temperature $\geq 62 \wedge$ radiation ≥ 131	0.9853849	-0.20299	28

(a) Pearson's r.

Subgroup	φ_{fs}	r_s	n
temperature ≤ 69	0.9928254	-0.0759	25
temperature $\leq 77 \wedge$ temperature $\geq 59 \wedge$ radiation $\leq$193.0	0.9447063	-0.3342	27
temperature $\leq 73 \wedge$ temperature ≥ 62	0.9404784	-0.2966	29

(b) Spearman's r_s.

Subgroup	φ_{fk}	τ_b	n
temperature ≤ 69	0.993904	-0.052	25
temperature $\leq 77 \wedge$ temperature $\geq 59.0 \wedge$ radiation $\leq$193	0.958155	-0.2156	27
temperature $\leq 73 \wedge$ temperature ≥ 62	0.926970	-0.2205	29

(c) Kendall's τ_b.

tributes reveals that the ozone concentration monotonically increases with higher radiation and temperature, while decreasing monotonically with higher wind speeds (cf. Figure 4a).

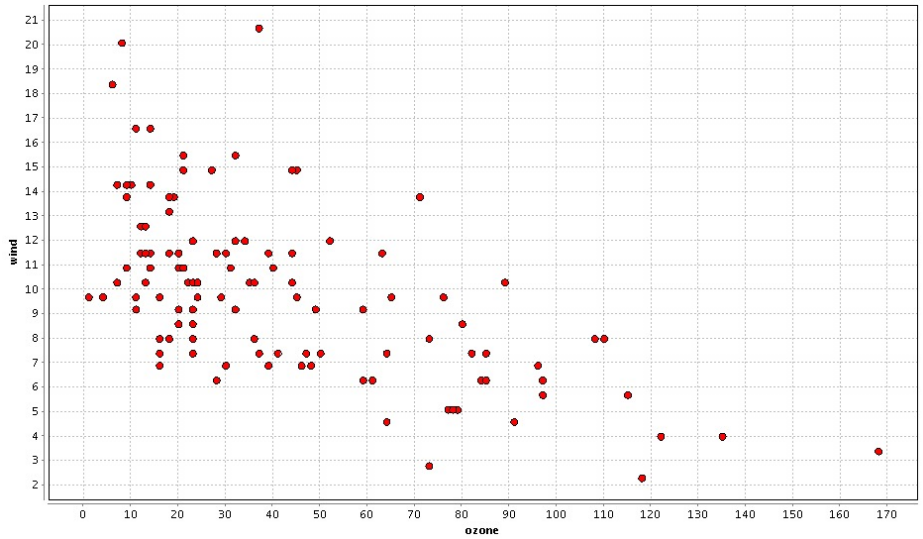
However, these relations do not always hold. Taking ozone concentration and wind speeds as targets, the most exceptional subgroups found with the three quality measures can be found in Table 3. The best subgroup Spearman finds is defined by the measurements with daily maximum temperature of 69°F or less (cf. Figure 4b). This suggests that on a milder temperatured day the concentration level is not too high in general and thus won't be affected by wind.

5.4. Contraceptive Method Choice

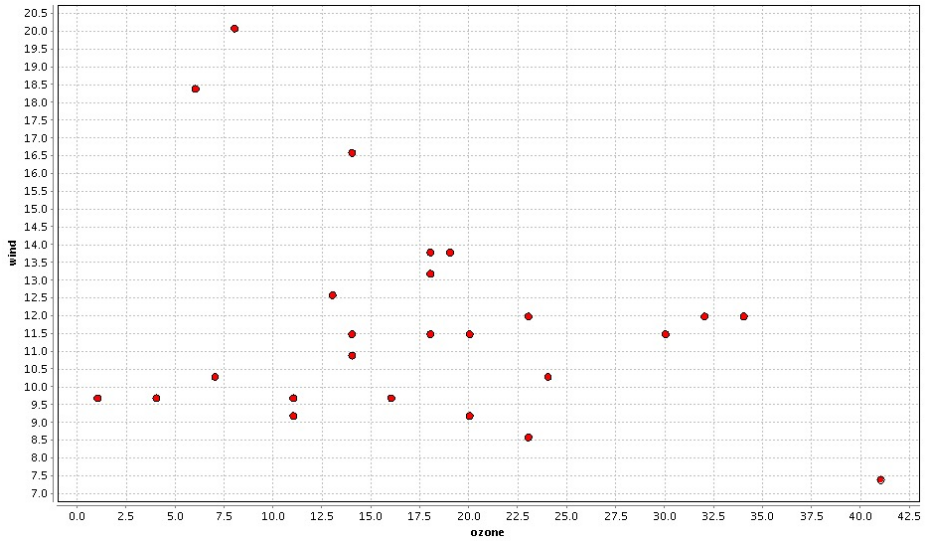
This dataset is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey. The dataset contains 1 473 samples of married women who were either not pregnant or did not know if they were at the time of interview.

One hypothesis could be that women with a higher education are more likely to employ long term contraception methods than women with a lower education and therefore also plan their pregnancy, resulting in motherhood at an older age. To investigate this assumption we selected *Wife's age* and *Number of children ever born* as target attributes.

The top-three results from both Pearson (cf. Table 4a) and Spearman (cf. Table 4b) are similar (the first-ranked subgroups are the exact same); they describe women with high education that employ long-term contraception methods, thus supporting our hypothesis of correlation between education and employed contraception method. Kendall (cf. Table 4c) finds descriptions that focus on the standard of living, targeting smaller subgroups compared to Pearson and Spearman.



(a) Entire Ozone dataset.



(b) Subgroup temperature ≤ 69.

Fig. 4. Target distribution on subgroups found on the Ozone dataset.

5.5. Iris

The Iris flower dataset (Fisher, 1936) contains 150 samples from three different species of Iris flowers: Setosa, Versicolor, and Virginica. Each sample has been examined with respect to four quantities: sepal length, sepal width, petal length and petal width. Sepal and petal are characteristic elements of a flowering plant. Setosa falls under the Iris series Tripetalae, whereas Versicolor and Virginica fall

Table 4. CMC: top-3 subgroups found with each of the correlation variants.

Subgroup	φ_{scd}	r	n
Wifes_edu = 4 ∧ Cont_method ≥ 2 ∧ Media_exp = 0	0.99998127	0.6725	398
Wifes_edu = 4 ∧ Cont_method = 2	0.99997633	0.7158	207
Wifes_edu = 4 ∧ Cont_method ≥ 2	0.99997175	0.6693	402

(a) Pearson's r.

Subgroup	φ_{fs}	r_s	n
Wifes_edu = 4 ∧ Cont_method ≥ 2 ∧ Media_exp = 0	0.999999986	0.7236	398
Wifes_edu = 4 ∧ Cont_method ≥ 2 ∧ Husbands_occu ≤ 2	0.999999983	0.7407	307
Wifes_edu = 4 ∧ Cont_method ≥ 2 ∧ Husbands_occu ≥ 1	0.999999966	0.7185	402

(b) Spearman's r_s.

Subgroup	φ_{fk}	τ_b	n
Wifes_edu ≥ 2 ∧ Std_living ≤ 2 ∧ Cont_method ≥ 2	0.9999999298	0.641	142
Wifes_edu ≥ 1 ∧ Std_living ≥ 3 ∧ Cont_method ≤ 1	0.9999994985399	0.346	432
Cont_method ≥ 3 ∧ Std_living ≤ 2 ∧ Wifes_edu ≤ 3	0.9999994569	0.66	104

(c) Kendall's τ_b.

Table 5. Iris: top-3 subgroups found with each of the correlation variants.

Subgroup	φ_{scd}	r	n
petalwidth ≥ 0.5 ∧ sepalwidth ≥ 2.2	0.999999988	0.8183	101
sepalwidth ≤ 4.1 ∧ petalwidth ≥ 0.3	0.999999557	0.8305	115
sepalwidth ≥ 2.5 ∧ petalwidth ≤ 0.3	0.999995618	0.2382	40

(a) Pearson's r.

Subgroup	φ_{fs}	r_s	n
petalwidth ≥ 2.1 ∧ sepalwidth ≤ 2.8	1	1	4
sepalwidth ≤ 4.1 ∧ petalwidth ≥ 0.3	0.999999655	0.8444	115
petalwidth ≤ 0.3	0.9999931	0.2736	41

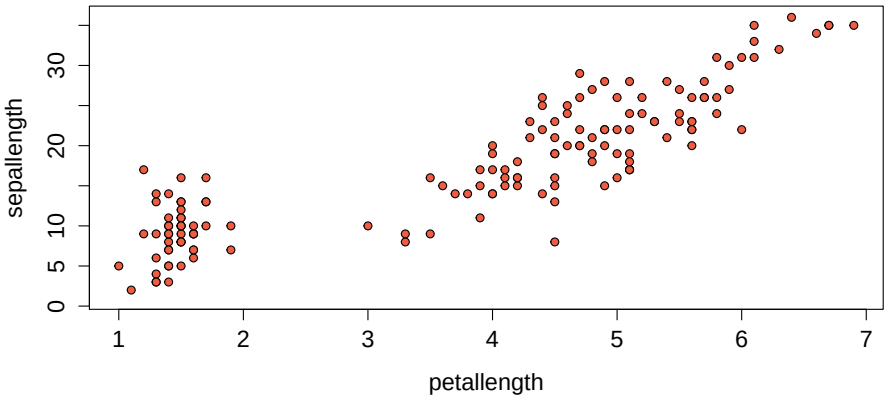
(b) Spearman's r_s.

Subgroup	φ_{fk}	τ_b	n
petallwidth ≤ 0.5 ∧ sepalwidth ≥ 3.2 ∧ petalwidth ≥ 0.1	0.999999985	0.118	36
sepalwidth ≥ 3.7 ∧ petalwidth ≤ 0.4 ∧ petalwidth ≥ 0.1	0.9999999785	-0.3234	13
petalwidth ≥ 0.3	0.9999998199	0.68	116

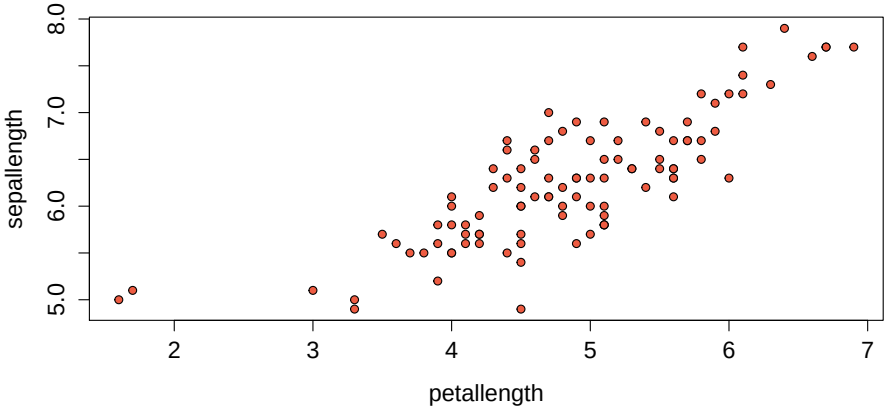
(c) Kendall's τ_b.

under the Iris series Laevigatae. Using simple cuts on single attributes of the dataset is enough to distinguish between the two Iris series, but usually it is not enough to distinguish between the two species within the same series. Instead a more complex interaction of attributes is necessary to separate Versicolor from Virginica. To that end, in these experiments, we take the petal and sepal length as our targets. A scatterplot of the overall target distribution is displayed as Figure 5a.

Experiments with the Iris dataset show that subgroups are found which separate the data with respect to their label. Pearson (cf. Table 5a) and Spearman (cf. Table 5b) both find subgroups excluding samples whose flower species is Setosa, while Kendall (cf. Table 5c) mirrors this behavior by returning subgroups consisting only of examples whose flower species is Setosa.



(a) Entire Iris dataset.



(b) Subgroup petalwidth $\geq 0.5 \wedge$ sepalwidth ≥ 2.2.

Fig. 5. Target distribution on subgroups found on the Iris dataset.

We observe that Pearson and Spearman's measures report more or less similar subgroups and relations, while Kendall's measure returns also subgroups whose targets feature weaker relations compared to their complements. For instance, Figure 5b contains the scatterplot of the targets, only for the records belonging to the best subgroup found with Pearson's r (hence quality measure φ_{scd}). The group of records in the lower left corner of Figure 5a, which appears to have the two targets correlated at most very weakly, has been removed almost completely in this subgroup, resulting in an apparently strongly correlated subgroup.

Table 6. Cern: top-3 subgroups found with two correlation variants.

Subgroup	φ_{scd}	r	n
PRI_lep_eta $\geq$ 2.0 $\wedge$ PRI_jet_leading_phi $\geq$ 2.497	$1 - 0.07302 \cdot 10^{-8}$	0.2163	817
PRI_lep_eta $\leq$ -1.99 $\wedge$ PRI_jet_leading_pt $\geq$ 134.551	$1 - 0.37190 \cdot 10^{-8}$	-0.2143	784
PRI_lep_eta $\leq$ -1.99 $\wedge$ PRI_jet_all_pt $\geq$ 215.471	$1 - 1.16989 \cdot 10^{-8}$	-0.2065	795

(a) Pearson's r.

Subgroup	φ_{fs}	r_s	n
PRI_lep_eta $\leq$ -1.99 $\wedge$ PRI_jet_all_pt $\geq$ 215.471	$1 - 0.8109 \cdot 10^{-8}$	-0.2027	795
PRI_lep_eta $\leq$ -1.99 $\wedge$ PRI_jet_leading_pt $\geq$ 134.551	$1 - 0.8400 \cdot 10^{-8}$	-0.2036	784
PRI_lep_eta $\geq$ 1.999 $\wedge$ PRI_jet_leading_phi $\geq$ 2.499	$1 - 9.7541 \cdot 10^{-8}$	0.1952	712

(b) Spearman's r_s.

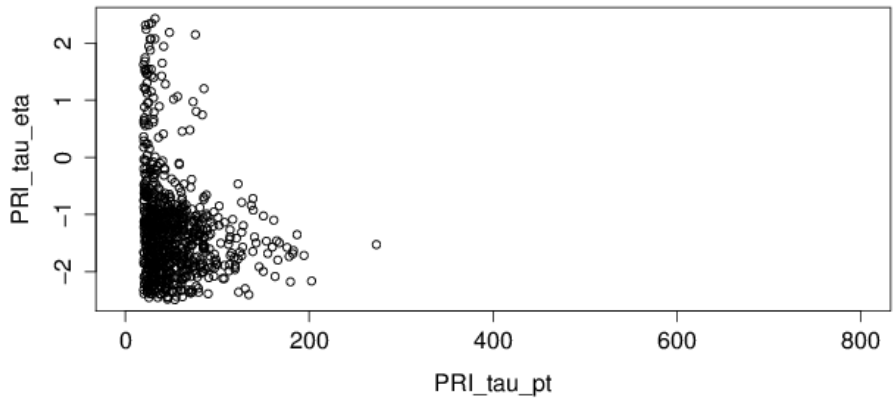
5.6. Higgs Boson Machine Learning Challenge

The Higgs boson is an elementary particle, which has recently been confirmed by experiments and is considered to be the particle (quantum) that provides other particles with mass. The ATLAS experiment at CERN provides simulated data used by physicists as a challenge to optimize the analysis of the Higgs boson. The dataset encompasses 250 000 simulated proton collisions (so-called events), which are characterized by a set of measured quantities, such as the energy momentum of the particle and the spatial coordinates of the resulting quarks. All quantities and their respective meanings can be found in the documentation (Adam-Bourdarios et al., 2014). The goal of the challenge is to improve classification of events. However, classification is not our primary goal; we will more generally explore whether we can find interesting subgroups in the data.

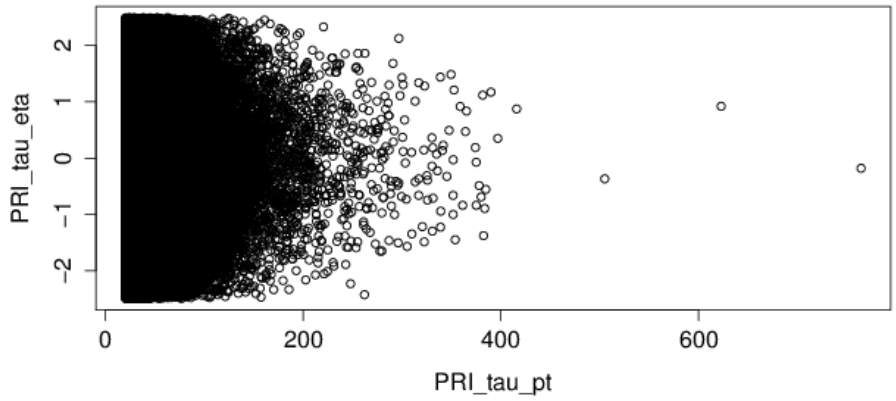
For the experiments, the attributes "Weight", "Label" and "Event ID" were excluded from the datasets, as they only served classification and identification purposes of the dataset. We also omitted all derived values (values starting with DER) as they are simply derived from the also present primitive values and should therefore not contribute significant knowledge about the relations of the measured quantities. Additionally, we imposed a restriction on the size of the subgroup, allowing only subgroups with a maximum coverage of 2 000 samples. Otherwise the found subgroups were too big and a sensible interpretation of their respective scatterplots was not possible.

On the Cern dataset, results again indicate that Spearman and Pearson will find similar subgroups. As an example, we present the found subgroups for gauging the relation between the attributes "PRI_tau_pt" and "PRI_tau_eta". The choice here is arbitrary as the drawn conclusion fits any of the experiments we performed with different targets. In Table 6a and Table 6b we present the top-3 subgroups found using Pearson and Spearman correlation, respectively. The second- and third-best subgroup found with Pearson correlation also appear in the top-three found with Spearman correlation. The odd ones out (top-ranked by Pearson and third-ranked by Spearman) have an almost identical definition.

In Figure 6a, displaying the subgroup that is ranked second by both Pearson and Spearman, we can see a concentration in the lower left corner as opposed to the structure of the complement in Figure 6b. This concentration suggests an interesting subgroup. However, for an evaluation, deeper knowledge about the data and how it was generated is necessary, which is not accessible to us. Nevertheless EMM is particularly interesting in this field, since, in real experiments



(a) Subgroup PRI_lep_eta $\leq$ -1.99 $\wedge$ PRI_jet_leading_pt $\geq$ 134.551.



(b) Complement of the subgroup presented in Figure 6a.

Fig. 6. Target distribution for subgroups found on the Cern dataset

the attributes are usually reconstructed because they cannot be measured directly. Thus a typical error source is a wrong reconstruction procedure. Finding subgroups that do not exhibit a correlation might indicate errors in the overall reconstruction procedure or noise in the measurements, which are important to filter and detect.

Additionally, these experiments do illustrate that the rank correlation model class for EMM is scalable beyond UCI-sized datasets.

6. Conclusions

We introduce the *rank correlation model class* for Exceptional Model Mining, a local pattern mining framework dedicated to finding subgroups for which multiple designated target attributes interact in an unusual manner. A model class in which this exceptional interaction was gauged in terms of Pearson's correlation between two targets had been developed previously. Our new rank correlation model class extends the EMM toolbox, by studying Spearman's rank correlation coefficient r_s and Kendall's τ_b between the two targets. This removes the assumption of target normality which is implicit in the existing correlation model class. Additional benefits of the rank correlation model class are the lower sensitivity to outliers, and the richer class of monotone target relations that can be explored.

Experiments on the Windsor Housing dataset, the South African Heart Disease Study dataset, the Ozone dataset, two UCI datasets (Contraceptive Method Choice and Iris), and the Higgs Boson ML Challenge dataset show that the subgroups found with the proposed Fieller-Spearman rank correlation quality measure φ_{fs} overlap with those found with the previously existing Pearson correlation quality measure φ_{scd}. The subgroups found with the proposed Fieller-Kendall rank correlation quality measure φ_{fk} overlap with the Pearson measure as well while occasionally also observing a different focus. This behavior makes sense: rank correlation gauges the strength of the *monotonic* relation between two targets, Pearson correlation gauges the strength of the *linear* relation between two targets, and the class of monotonic relations encompasses the class of linear relations. This provides corroborating evidence of the soundness of our experimental results: the set of subgroups found with the rank correlation model class encompasses the set of subgroups found with the previously introduced correlation model class.

Possible alternatives to the presented models that could be investigated in the future are the application of more experimental measures like *dCor*, *MIC* or other correlation quantifiers mentioned in Section 2.3. However, for a good quality measure it would also be necessary to investigate ways to compare these statistics on different subsets of datasets. A good statistical foundation is available for several alternative measures developed in the context of outlier detection. For instance, the method developed for outlier detection in (Keller et al., 2012) is probably not limited to the outlier detection domain, and the scalable selection of correlated groups of dimensions in (Nguyen et al., 2013) has been shown to be applicable on a range of data mining tasks. That range currently does not include Exceptional Model Mining, but this is a promising direction of research. Another promising extension would be to look at measures gauging correlation between a larger number of targets (rather than the two targets investigated in this paper), such as the multivariate maximal correlation analysis from (Nguyen et al., 2014).

Finally, in future work, we intend to experimentally validate Fieller's claim (cf. Section 3.4), regarding the suitability of his modification of the Fisher z-transformation for a class of parental distributions that is much wider than just normal distributions, on datasets of more than 50 records. In the year 1957, when the corresponding paper (Fieller et al., 1957) was originally published, such data may not have been readily available. Nowadays we could potentially evaluate the claim on a plethora of UCI datasets, and verify the veracity of a conjecture that has had the status of merely a conjecture for almost sixty years.

Acknowledgements. We would like to thank Dr. Johannes Albrecht (Emmy Noether group leader at the TU Dortmund, department of experimental physics, with research focus on the CERN LHCb Experiment) for fruitful discussion and helpful comments. This research is supported in part by the Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center SFB 876 "Providing Information by Resource-Constrained Analysis", project A1. This work was supported by the European Union through the ERC Consolidator Grant FORSIED (project reference 615517).

References

Adam-Bourdarios C, Cowan G, Cécile Germain IG, Kégl B, Rousseau D (2014) Learning to discover: The Higgs Boson Machine Learning Challenge. `http://higgsml.lal.in2p3.fr/documentation/`, accessed August 7th

Agrawal R, Mannila H, Srikant R, Toivonen H, Verkamo AI (1996) Fast Discovery of Association Rules. Advances in Knowledge Discovery and Data Mining, pp 307–328

Anglin PM, Gençay R (1996) Semiparametric Estimation of a Hedonic Price Function. Journal of Applied Econometrics 11(6):633–648

Anscombe FJ (1973) Graphs in Statistical Analysis. The American Statistician 27(1): 17–21

Balasubramaniyan R, Hüllermeier E, Weskamp N, Kämper J (2005) Clustering of gene expression data using a local shape-based similarity measure. Bioinformatics 21(7):1069–1077

Bay SD, Pazzani MJ (2001) Detecting Group Differences: Mining Contrast Sets. Data Mining and Knowledge Discovery 5(3):213–246

Blum JR, Kiefer J, Rosenblatt M (1961) Distribution Free Tests of Independence based on the Sample Distribution Function. Annals of Mathematical Statistics 32(2):485–498

Breese JS, Heckerman D, Kadie CM (1998) Empirical Analysis of Predictive Algorithms for Collaborative Filtering. Proc. UAI, pp 43–52

Clark M (2013) A Comparison Of Correlation Measures. Technical Report, University of Notre Dame

Conover WJ (1971) Practical Nonparametric Statistics. Wiley

Dong G, Li J (1999) Efficient Mining of Emerging Patterns: Discovering Trends and Differences. Proc. KDD, pp 43–52

Downar L (2014) A Rank Correlation Model Class for Exceptional Model Mining. Bachelor's thesis, TU Dortmund

Downar L, Duivesteijn W (2015) Exceptionally Monotone Models — the Rank Correlation Model Class for Exceptional Model Mining. Proc. ICDM, to appear

Duivesteijn W (2013) Exceptional Model Mining. PhD thesis, Leiden University

Duivesteijn W, Feelders A, Knobbe A (2012a) Different Slopes for Different Folks — Mining for Exceptional Regression Models with Cook's Distance. Proc. KDD, pp 868–876

Duivesteijn W, Feelders AJ, Knobbe A (2016) Exceptional Model Mining — Supervised Descriptive Local Pattern Mining with Complex Target Concepts. Data Mining and Knowledge Discovery 30:47–98

Duivesteijn W, Knobbe A, Feelders A, Van Leeuwen M (2010) Subgroup Discovery meets Bayesian Networks – An Exceptional Model Mining Approach. Proc. ICDM, pp 158–167

Duivesteijn W, Loza Mencía E, Fürnkranz J, Knobbe A (2012b) Multi-Label LeGo — Enhancing Multi-label Classifiers with Local Patterns. Technical Report TUD-KE-2012-02, TU Darmstadt

Duivesteijn W, Thaele J (2014) Understanding Where Your Classifier Does (Not) Work — the SCaPE Model Class for EMM. Proc. ICDM, pp 809–814

Fieller EC, Hartley HO, Pearson ES (1957) Tests for Rank Correlation Coefficients. I. Biometrika 44(4):470–481

Fisher DH, Langley PW (1986) Conceptual Clustering and its Relation to Numerical Taxonomy. In: Gale WA (Ed.) Artificial Intelligence and Statistics, Reading, MA: Addison-Wesley, pp. 77–116

Fisher RA (1936) The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics 7(2):179–188

Fisher RAS (1970) Statistical Methods for Research Workers. Oliver and Boyd, 14^{th} ed.

Galbrun E, Miettinen P (2012) From Black and White to Full Color: Extending Redescription Mining Outside the Boolean World. Statistical Analysis and Data Mining 5(4):284–303

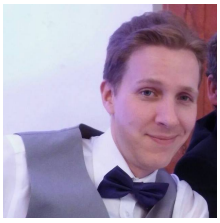
Gebelein H (1941) Das statistische Problem der Korrelation als Variations- und Eigenwert-

problem und sein Zusammenhang mit der Ausgleichsrechnung. Zeitschrift für Angewandte Mathematik und Mechanik 21:364–379
Gretton A, Bousquet O, Smola A, Schölkopf B (2005) Measuring Statistical Dependence with Hilbert-Schmidt Norms. Proc. ALT, pp. 63–77
Han J, Pei J, Yin Y (2000) Mining Frequent Patterns without Candidate Generation. Proc. SIGMOD, pp. 1–12
Hand D, Adams N, Bolton R (eds) (2002) Pattern Detection and Discovery. Springer, New York
Hastie T, Tibshirani R, Friedman, J (2010) The Elements of Statistical Learning. Springer, Stanford
Herrera F, Carmona CJ, González P, Del Jesus MJ (2011) An Overview on Subgroup Discovery: Foundations and Applications. Knowledge and Information Systems 29(3):495–525
Hoeffding W (1948) A Non-Parametric Test of Independence. Annals of Mathematical Statistics 19(4):546–557
Hollander M, Wolfe D (1999) Nonparametric Statistical Methods. Wiley Series in Probability and Statistics, Wiley, 2nd ed.
Jorge AM, Azevedo PJ, Pereira F (2006) Distribution Rules with Numeric Attributes of Interest. Proc. PKDD, pp 247–258
Keller F, Müller E, Böhm K (2012) HiCS: High Contrast Subspaces for Density-Based Outlier Ranking. Proc. ICDE, pp 1037–1048
Kendall MG (1938) A New Measure of Rank Correlation. Biometrika 30(1):81–93
Kinney JB, Atwal GS (2014) Equitability, Mutual Information, and the Maximal Information Coefficient. Proceedings of the National Academy of Sciences of the United States of America 111(9):3354–3359
Kowalski CJ (1972) On the Effects of Non-Normality on the Distribution of the Sample Product-Moment Correlation Coefficient. Journal of the Royal Statistical Society, Series C (Applied Statistics) 21(1):1–12
Kralj Novak P, Lavrač N, Webb GI (2009) Supervised Descriptive Rule Discovery: A Unifying Survey of Contrast Set, Emerging Pattern and Subgroup Mining. Journal of Machine Learning Research 10:377–403
Leman D, Feelders A, Knobbe AJ (2008) Exceptional Model Mining. Proc. ECML/PKDD (2), pp 1–16
Lemmerich F, Becker M, Atzmüller M (2012) Generic Pattern Trees for Exhaustive Exceptional Model Mining. Proc. ECML-PKDD (2), pp 277–292
Li WK, Lee SY (1980) Application of Rank Correlation to Lanthanide Induced Shift Data. Organic Magnetic Resonance 13(2):97–99
Lichman M (2013) UCI Machine Learning Repository. [`http://archive.ics.uci.edu/ml`], Irvine, CA: University of California, School of Information and Computer Science
Lim TS, Loh WY, Shih YS (2000) A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-three Old and New Classification Algorithms. Machine Learning 40:203–228
Lopez-Paz D, Hennig P, Schölkopf B (2013) The Randomized Dependence Coefficient. Advances in Neural Information Processing Systems, pp 1–9
Mannila H, Toivonen H (1997) Levelwise Search and Borders of Theories in Knowledge Discovery. Data Mining and Knowledge Discovery 1(3):241–258
Mierswa I, Wurst M, Klinkenberg R, Scholz M, Euler T (2006) YALE: Rapid Prototyping for Complex Data Mining Tasks. Proc. KDD, pp 935–940
Moens S, Boley M (2014) Instant Exceptional Model Mining Using Weighted Controlled Pattern Sampling. Proc. IDA, pp 203–214
Morik K, Boulicaut JF, Siebes A (eds) (2005) Local Pattern Detection. Springer, New York
Nguyen HV, Müller E, Böhm K (2013) 4S: Scalable Subspace Search Scheme Overcoming Traditional Apriori Processing. Proc. BigData, pp 359–367
Nguyen HV, Müller E, Vreeken J, Efros P, Böhm K (2014) Multivariate Maximal Correlation Analysis. Proc. ICML, pp 775–783
Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, Lander ES, Mitzenmacher M, Sabeti PC (2011) Detecting Novel Associations in Large Data Sets. Science 334:1518–1524
Rousseauw J, du Plessis, J, Benade A, Jordaan P, Kotze J, Jooste P, Ferreira J (1983) Coronary risk factor screening in three rural communities. South African Medical Journal 64:430–436
Spearman C (1904) The Proof and Measurement of Association Between two Things. American Journal of Psychology 15(1):72–101

Székely GJ, Rizzo ML, Bakirov NK (2007) Measuring and Testing Dependence by Correlation of Distances. Annals of Statistics 35(6):2769–2794
Tsoumakas G, Katakis I (2007) Multi-label Classification: an Overview. International Journal of Data Warehousing & Mining 3(3):1–13
Umek L, Zupan B (2011) Subgroup Discovery in Data Sets with Multi-Dimensional Responses. Intelligent Data Analysis 15(4):533–549
Yilmaz E, Aslam JA, Robertson S (2008) A New Rank Correlation Coefficient for Information Retrieval. Proc. SIGIR, pp 587–594

Author Biographies

Lennart Downar received a B.Sc. degree in Applied Computer Science from TU Dortmund University, Germany, in 2014. From 2012 to 2013, he was an exchange student at the University of Leiden, the Netherlands. Since 2015 he is working as a student employee at the Chair of Artificial Intelligence at the Faculty of Computer Science, TU Dortmund University. He is currently a Master student at the Department of Computer Science, TU Dortmund University. His research interests include data mining, machine learning and robotics.



Wouter Duivesteijn received B.Sc. degrees (Mathematics and Computer Science) from Utrecht Universiteit, the Netherlands, in 2005. M.Sc. degrees followed from the same university, in Mathematical Sciences (2007) and Applied Computing Science (2008). From 2009 until 2013 he worked as Assistent In Opleiding (∼ Ph.D. candidate) at Leiden University, the Netherlands, which awarded to Wouter his Ph.D. degree in Computer Science, in 2013. He has since worked as Wissenschaftlicher Mitarbeiter (∼ postdoctoral researcher) at the TU Dortmund, Germany (2013-2015), as honorary Research Associate (∼ postdoctoral researcher) at the University of Bristol, UK (2015-2015), and as postdoctoraal bursaal (∼ postdoctoral researcher) at Ghent University, Belgium (2015-2016), which is his current affiliation by time of paper acceptance. By the time you read this (after September 1, 2016), Wouter will have started working as Assistant Professor Data Mining at the Technische Universiteit Eindhoven.

Correspondence and offprint requests to: Wouter Duivesteijn, Data Science Lab & iMinds, Universiteit Gent, Belgium, `wouter.duivesteijn@ugent.be`