2015 IEEE International Conference on Data Mining

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

Lennart Downar
Fakultät für Informatik, LS VIII
Technische Universität Dortmund
lennart.downar@udo.edu

Wouter Duivesteijn
Fakultät für Informatik, LS VIII
Technische Universität Dortmund
wouterduivesteijn@gmail.com

***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.**

I. 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. Exceptional Model Mining (EMM) [2], [3], [4] 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 [2]. 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 [5], [6]. 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 [7], information retrieval [8], recommender systems [9], and determining molecular structure by lanthanide shift reagents [10]. 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. We define quality measures for the rank correlation model class based on Spearman's rank correlation coefficient r_s [5], and on Kendall's τ_b [6]. Experiments on the Windsor Housing dataset, UCI datasets and the real-life Higgs Boson ML Challenge dataset indicate that with Spearman we find results similar to the ones found with Pearson, while Kendall has a markedly different focus.

II. Related Work

Pattern mining [11], [12] 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* [13], 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* [14], the interestingness of a pattern is gauged in an

[1]This paper builds upon the work from the Bachelor's thesis [1].

DOI 10.1109/ICDM.2015.81

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) [15]. 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) [2], [3] 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 [2] — in Section II-A. Other investigated model classes are variance of a single target [16][2], association between two nominal targets [4], simple linear regression on two targets [2], behavior of a hard classifier [2], total variation on a contingency table of any size [17], distance over a multivariate mean model [17], structure of a Bayesian network on any number of nominal targets [18], linear regression on any number of targets [19], and SCaPE (Soft Classifier Performance Evaluation) [20].

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.

A. 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* [2]. 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 [2]):

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

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:

$$\text{Lives near nuclear plant} \quad \Rightarrow \quad \left| r^S - r^{\Omega \setminus S} \right| \text{ is high}$$

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

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 is open for debate. Kowalski's experimental evaluation [21] 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" [21, 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) 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 [22], 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 Person 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.

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 improve. 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. Quite a number of situations are imaginable in which such an exceptionally linear relation is interesting to detect. However, 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. Quite a number of situations are imaginable in which, rather than linearity, an exceptionally monotone relation is interesting to detect. The correlation model class serves a clear purpose, but there is more to explore.

[2]Whether this model class falls under the *spirit* of EMM is debatable; having only a single target prohibits investigating target interaction. However, careful reading of EMM literature [2], [4] 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 [16] introduced this model class as an EMM instance, and we cannot formally reject it as such, we adopt it into the EMM canon.

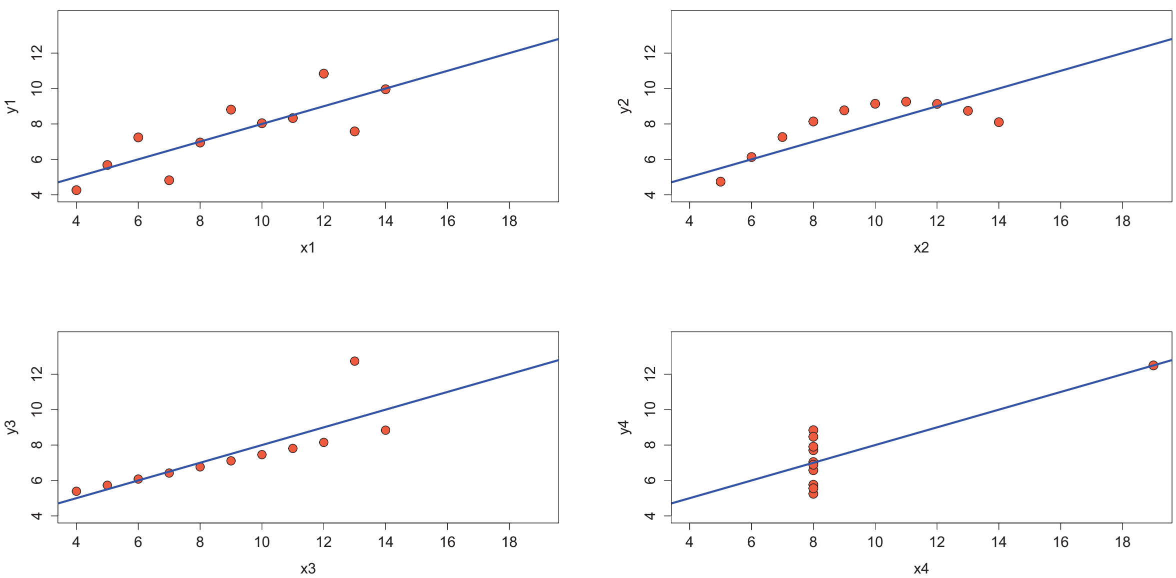


Fig. 1. Anscombe's quartet.

B. Tasks Related to EMM

Local Pattern Mining tasks that are similar to SD are Contrast Set Mining [23] and Emerging Pattern Mining [24]. 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 [25]. Explicitly seeking a deviating model over a target is performed in Distribution Rules [26], 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. [27] 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 [28] 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 and multi-label classification. 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 [4], and methods on the crossroads of EMM and multi-label classification are discussed in Section 8 of [29].

C. Alternative Correlation Measures

A comparison of several correlation measures has been given by Clark [30]. Apart from Pearson's r and Spearman's r_s, he 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 [31] 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. [32] can be used as a large-sample approximation for D [33].

Distance correlation (*dCor*) has been introduced by Székely et al. [34] 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 [30], a ranked-based version of *dCor* could also be incorporated.

Reshef et al. [35] 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.

1) Evaluation: After comparing these alternatives on several non-linear relations, Clark [30] notes that "Hoeffding's D only works in some limited scenarios". In the experiments, D did pick up some of the non-linear relations (e.g., a quadratic

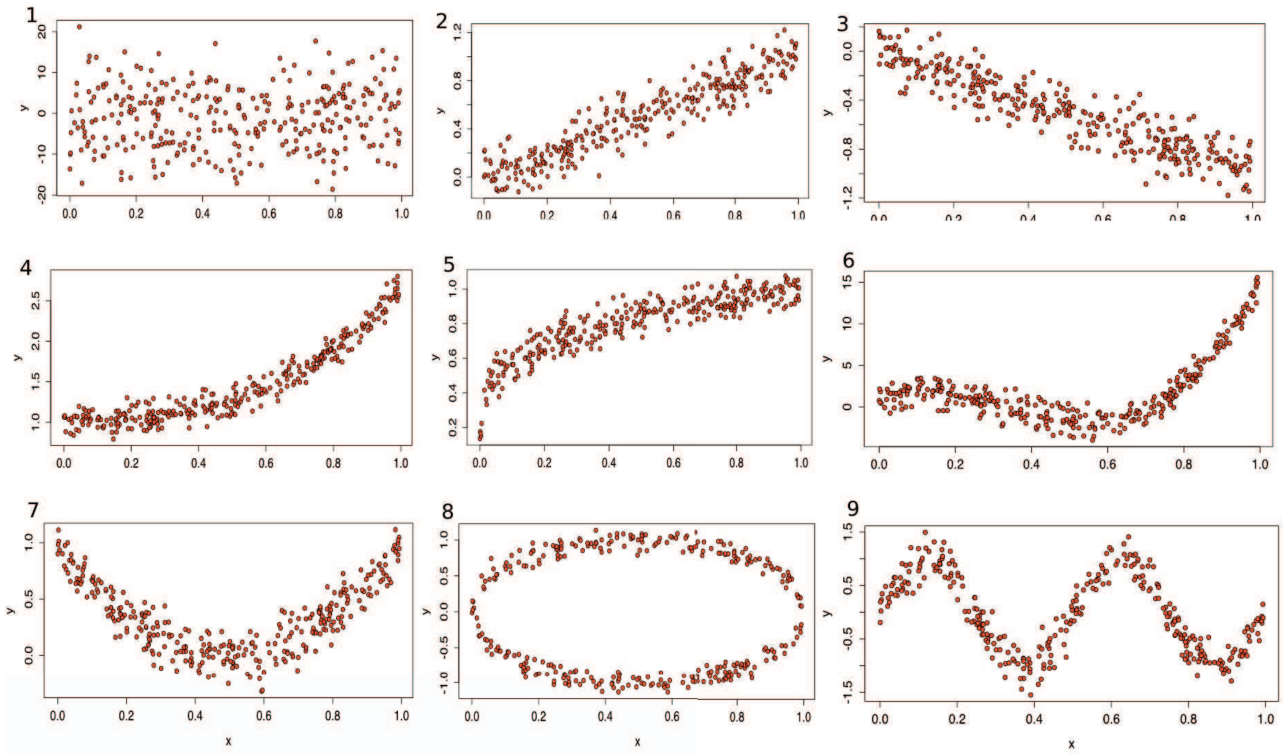


Fig. 2. Target relations, detectable by alternative correlation measures.

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 [36]. Consequently, Clark concludes [30] that we "still need to be on the lookout for a measure that is both highly interpretable and possesses all the desirable qualities we want".

2) Other Approaches: As pointed out by Clark [30], 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. [37] 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. [38] proposed the Randomized Dependence Coefficient (*RDC*), which is an estimate of the Hirschfeld-Gebelein-Rényi Maximum Correlation Coefficient (*HGR*) defined by Gebelein [39] in 1941. However, HGR is not computable and thus represents only an abstract concept.

In the remainder of this paper, we will only consider correlation measures for which a straightforward adaptation of a well-known statistical test (cf. Section IV-C) 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 IV-C have a solid basis in statistics, and come with the additional benefit that the interpretation of their values is straightforward. By contrast, for the alternative correlation measures introduced in this section, to the best of our knowledge no statistical way exists to compare results on different data samples. Developing the statistical theory necessary to base a mathematically well-supported quality measure on these measures is beyond the scope of this paper.

III. 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 relations 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 II-C) of alternative correlation measures available in the literature, highlighting the potential for future research into their underlying statistical theory.

IV. 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.

A. 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 of a dataset) has been developed by Charles Spearman [5]. 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:

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

In case of ties, Conover [40] 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.

B. Kendall's Tau

Where Spearman's r_s uses the difference of ranks in individual pairs, Kendall's τ [6] 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:

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

C. 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 justification, 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 [41]. Hence, the authors of [2] 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. [42] 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:

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

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

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

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 [2], 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.

D. Limitations

In Section II-A, 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 II-A2, and they capture the monotone relations that were discussed to be desirable in Section II-A3. Recall that the third limitation of the Pearson correlation coefficient, as identified in Section II-A1, is that it assumes a normal distribution over the targets. The rank correlation measures presented in Sections IV-A and IV-B, as well as the naive quality measures $\varphi_{\text{abs\_}r_s}$ and $\varphi_{\text{abs\_}\tau_b}$, do not have this assumption. However, indirectly it comes into play again when applying the slight modifications of the Fisher z-transformation presented in [42] (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" [42, page 3]. Remarkably, to the best of our knowledge, in the half-century since this paper was published, no further investigation has occurred.

V. Experiments

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 [2]. Notice that nothing could be learned from comparative experiments with results from other EMM model classes or other pattern mining methods: another model class or another pattern mining method is designed to find subgroups that display a kind of exceptional behavior different from the one gauged with the (rank) correlation model class. Hence, there is no way to compare the resulting subgroup sets that also does justice to both methods to be compared; no high-level conclusions could be drawn from such an experiment. Therefore, we compare the rank correlation model class only with the original correlation class. We compensate for this by analyzing some found subgroups in detail.

TABLE I. Long subgroup definitions.

Subgroup	Description
S_1	Wifes_edu = 4 ∧ Cont_method ≥ 2 ∧ Media_exp = 0
S_2	Wifes_edu = 4 ∧ Cont_method ≥ 2 ∧ Husbands_occu ≤ 2
S_3	Wifes_edu = 4 ∧ Cont_method ≥ 2 ∧ Husbands_occu ≥ 1
S_4	Wifes_edu ≤ 3 ∧ Std_living ≥ 3 ∧ Husbands_edu ≤ 3
S_5	Husbands_edu ≥ 1 ∧ Std_living ≥ 3 ∧ Husbands_edu ≤ 3
S_6	petalwidth ≤ 0.5 ∧ petalwidth ≥ 0.2 ∧ sepalwidth ≥ 2.9
S_7	sepalwidth ≥ 3.4 ∧ petalwidth ≤ 2.4 ∧ petalwidth ≥ 0.2
S_8	petalwidth ≤ 1.7 ∧ sepalwidth ≥ 3.7 ∧ petalwidth ≥ 0.1
S_9	PRI_lep_eta ≥ 2.0 ∧ PRI_jet_leading_phi ≥ 2.497
S_{10}	PRI_lep_eta ≤ -1.99 ∧ PRI_jet_leading_pt ≥ 134.551
S_{11}	PRI_lep_eta ≤ -1.99 ∧ PRI_jet_all_pt ≥ 215.471
S_{12}	PRI_lep_eta ≥ 1.999 ∧ PRI_jet_leading_phi ≥ 2.499

Early in Section IV, 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 datasets 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 [3, Algorithm 1, page 19] and also described in [4]. This is a heuristic search algorithm, whose complexity has been analyzed [3], [4] 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 user-set parameters of the beam search algorithm (where a typical generous setting would be in the order of magnitude of $w = 100$ and $d = 5$). 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 (in the rank correlation model class, $m = 2$).

We have implemented our work within the RapidMiner analytics platform [43]. 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 .

We have performed experiments on four datasets, two of which stem from the UCI machine learning repository [44]. In the following sections, we present results of experiments on the Windsor Housing dataset [45], the Contraceptive Method Choice (CMC) dataset [46], [44], the Iris dataset [47], [44], and the real-life Higgs Boson Machine Learning Challenge dataset [48]. If the found subgroups have definitions that would unnecessarily stretch the tables, we have given them a subgroup label S_i; their definitions can be found in Table I.

TABLE II. WINDSOR HOUSING: TOP-3 SUBGROUPS FOUND WITH EACH OF THE CORRELATION VARIANTS.

Subgroup	φ_{scd}	r	n
fb $\leq$ 2 $\wedge$ drv = 1 $\wedge$ sty $\leq$ 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 $\geq$ 3	0.9999781	-0.0932	36

(b) Spearman's r_s.

Subgroup	φ_{fk}	τ_b	n
fb = 1 $\wedge$ drv = 1	1	0.370	341
bdms $\leq$ 3 $\wedge$ drv = 1 $\wedge$ reg = 0	1	0.3329	277
bdms $\leq$ 3 $\wedge$ drv = 1 $\wedge$ ca = 0	1	0.31993	261

(c) Kendall's τ_b.

TABLE III. CMC: TOP-3 SUBGROUPS FOUND WITH EACH OF THE CORRELATION VARIANTS.

Subgroup	φ_{scd}	r	n
S_1	0.99998127	0.6725	398
Wifes_edu = 4 $\wedge$ Cont_method = 2	0.99997633	0.7158	207
Wifes_edu = 4 $\wedge$ Cont_method $\geq$ 2	0.99997175	0.6693	402

(a) Pearson's r.

Subgroup	φ_{fs}	r_s	n
S_1	0.999999986	0.7236	398
S_2	0.999999983	0.7407	307
S_3	0.999999966	0.7185	402

(b) Spearman's r_s.

Subgroup	φ_{fk}	τ_b	n
S_4	0.9999999999999842	0.3371	309
Wifes_edu $\leq$ 2 $\wedge$ Std_living $\geq$ 3	0.9999999999999442	0.3330	292
S_5	0.9999999999999382	0.3485	335

(c) Kendall's τ_b.

A. *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 IIb) and Pearson (Table IIa) measures confirm the experiments performed on the Windsor Housing dataset in [2], as both return the subgroup:

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

The subgroup S_0 encompasses 35 houses that have a driveway, a recreation room and at least two bathrooms. Leman et al. [2] 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 all three measures focus on houses featuring a driveway. However, whereas Pearson and Spearman tend to prefer houses in higher market segments (both Tables IIa and IIb feature a subgroup with the condition that the houses have at least three bedrooms), Kendall zooms in on lower market segments: the first subgroup in Table IIc encompasses houses with only one bathroom, and the other two encompass houses with at most three bedrooms. The condition *reg=0* expresses that the house is in a non-preferred neighborhood of Windsor, while *ca=0* expresses that the house does not have air conditioning.

B. *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.

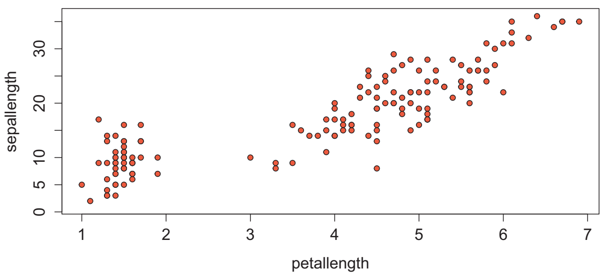
One hypothesis could be that women with a higher education are more likely to employ long term contraception methods than women with a lower education and therefore also plan their pregnancy, resulting in motherhood at an older age. To investigate this assumption we selected *Wife's age* and *Number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

The top-three results from both Pearson (Table IIIa) and Spearman (Table IIIb) 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 (Table IIIc) finds descriptions for groups where education in both partners is lower. The correlation between a woman's age and the number of children born in those groups is less then in their complements, confirming our hypothesis again, but from another angle.

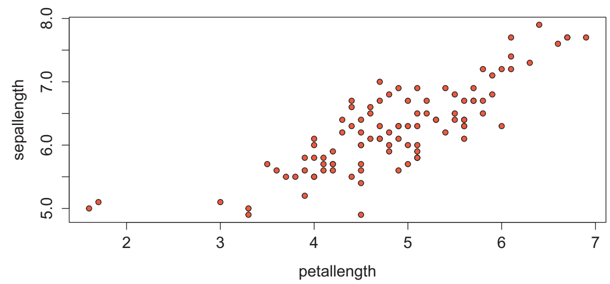
C. *Iris*

The Iris flower dataset [47] 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 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 3a.

Experiments with the Iris dataset show that subgroups are found which separate the data with respect to their label. Pearson (Table IVa) and Spearman (Table IVb) both find subgroups excluding samples whose flower species is Setosa, while Kendall (Table IVc) 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. 3. Target distribution on subgroups found on the Iris dataset.

TABLE IV. IRIS: TOP-3 SUBGROUPS FOUND WITH EACH OF THE CORRELATION VARIANTS.

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

(a) Pearson's r.

Subgroup	φ_{fs}	r_s	n
petalwidth $\geq 2.1 \wedge$ sepalwidth ≤ 2.8	1	1	4
sepalwidth $\leq 4.1 \wedge$ 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
S_6	0.9999999999999996	0.1515	42
S_7	0.999999999999993	0.2648	34
S_8	0.9999999999999911	-0.3234	13

(c) Kendall's τ_b.

We observe that Pearson and Spearman's measures report more or less similar subgroups and relations, while Kendall's measure returns subgroups whose targets feature weaker relations compared to their complements. For instance, Figure 3b 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 3a, 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.

D. 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 [48]. 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.

TABLE V. CERN: TOP-3 SUBGROUPS FOUND WITH TWO CORRELATION VARIANTS.

Subgroup	φ_{scd}	r	n
S_9	0.9999999992698	0.2163	817
S_{10}	0.9999999962810	-0.2143	784
S_{11}	0.9999999883011	-0.2065	795

(a) Pearson's r.

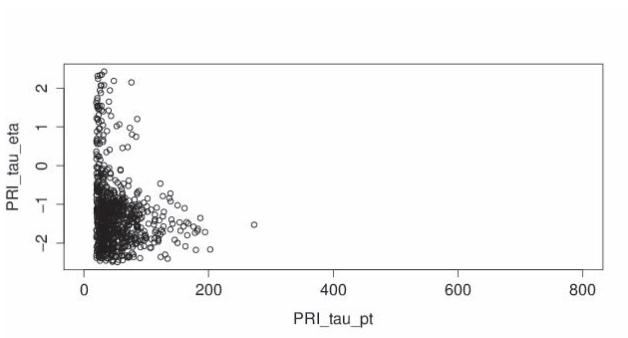
Subgroup	φ_{fs}	r_s	n
S_{11}	0.999999991891	-0.2027	795
S_{10}	0.999999991600	-0.2036	784
S_{12}	0.999999902459	0.1952	712

(b) Spearman's r_s.

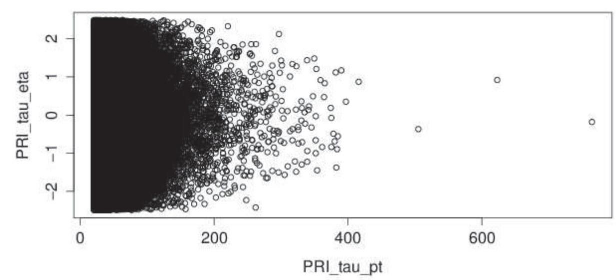
For the experiments, the attributes "Weight", "Label" and "Event ID" were excluded from the datasets, as they only served classification and identification purposes of the dataset. We also omitted all derived values (values starting with DER) as they are simply derived from the also present primitive 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 Va and Table Vb we present the top-3 subgroups found using Pearson and Spearman correlation, respectively. Of the top-3 subgroups found with Pearson correlation, two (S_{10} and S_{11}) also appear in the top-three found with Spearman correlation. The odd ones out, S_9 and S_{12}, have an almost identical definition (cf. Table I).

In Figure 4a, displaying subgroup S_{10}, we can see a concentration in the lower left corner as opposed to the structure



(a) Subgroup PRI_lep_eta ≤ -1.99 ∧ PRI_jet_leading_pt ≥ 134.551.



(b) Complement of the subgroup presented in Figure 4a.

Fig. 4. Target distribution for subgroups found on the Cern dataset

of the complement in Figure 4b. For thorough interpretation a domain expert should be consulted. We do not have such a domain expert at our disposal, but these experiments do illustrate that the rank correlation model class for EMM is scalable beyond UCI-sized datasets.

VI. Conclusions

We introduce the *rank correlation model class* for Exceptional Model Mining, a pattern mining framework dedicated to finding subgroups for which multiple designated target attributes interact in an unusual way. 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, two UCI datasets, 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}. On the other hand, the subgroups found with the proposed Fieller-Kendall rank correlation quality measure φ_{fk} have a different focus. This behavior makes sense: rank correlation gauges the strength of the *monotone* relation between two targets, Pearson correlation gauges the strength of the *linear* relation between two targets, and the class of monotone 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 II-C. 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 fundament is available for several alternative measures developed in the context of outlier detection. For instance, the method developed for outlier detection in [49] is probably not limited to the outlier detection domain, and the scalable selection of correlated groups of dimensions in [50] 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 [51]. Finally, as future work we strive to develop a valuation basis for the GP-Growth algorithm [16]. This development would allow for exhaustive search with the rank correlation model class, as opposed to the heuristic beam search performed in this paper.

VII. Acknowledgments

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.

References

[1] L. Downar, *A Rank Correlation Model Class for Exceptional Model Mining*, Bachelor's thesis, TU Dortmund, 2014.

[2] D. Leman, A. Feelders, A. J. Knobbe, *Exceptional Model Mining*, Proc. ECML/PKDD (2), pp. 1–16, 2008.

[3] W. Duivesteijn, *Exceptional Model Mining*, PhD thesis, Leiden University, 2013.

[4] W. Duivesteijn, A. J. Feelders, A. Knobbe, *Exceptional Model Mining — Supervised Descriptive Local Pattern Mining with Complex Target Concepts*, Data Mining and Knowledge Discovery, to appear, 2015.

[5] C. Spearman, *The Proof and Measurement of Association Between two Things*, American Journal of Psychology 15 (1), pp. 72–101, 1904.

[6] M. G. Kendall, *A New Measure of Rank Correlation*, Biometrika 30 (1), pp. 81–93, 1938.

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

[8] E. Yilmaz, J. A. Aslam, S. Robertson, *A New Rank Correlation Coefficient for Information Retrieval*, Proc. SIGIR, pp. 587–594, 2008.

[9] J. S. Breese, D. Heckerman, C. M. Kadie, *Empirical Analysis of Predictive Algorithms for Collaborative Filtering*, Proc. UAI, pp. 43–52, 1998.

[10] W.-K. Li, S.-Y. Lee, *Application of Rank Correlation to Lanthanide Induced Shift Data*, Organic Magnetic Resonance 13 (2), pp. 97–99, 1980.

[11] D. Hand, N. Adams, R. Bolton (eds), *Pattern Detection and Discovery*, Springer, New York, 2002.

[12] K. Morik, J. F. Boulicaut, A. Siebes (eds), *Local Pattern Detection*, Springer, New York, 2005.

[13] H. Mannila, H. Toivonen, *Levelwise Search and Borders of Theories in Knowledge Discovery*, Data Mining and Knowledge Discovery 1 (3), pp. 241–258, 1997.

[14] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, A. I. Verkamo, *Fast Discovery of Association Rules*, Advances in Knowledge Discovery and Data Mining, pp. 307–328, 1996.

[15] F. Herrera, C. J. Carmona, P. González, M. J. del Jesus, *An Overview on Subgroup Discovery: Foundations and Applications*, Knowledge and Information Systems 29 (3), pp. 495–525, 2011.

[16] F. Lemmerich, M. Becker, M. Atzmüller, *Generic Pattern Trees for Exhaustive Exceptional Model Mining*, Proc. ECML-PKDD (2), pp. 277–292, 2012.

[17] S. Moens, M. Boley, Instant Exceptional Model Mining Using Weighted Controlled Pattern Sampling, Proc. IDA, pp. 203–214, 2014.

[18] W. Duivesteijn, A. Knobbe, A. Feelders, M. van Leeuwen, *Subgroup Discovery meets Bayesian Networks – An Exceptional Model Mining Approach*, Proc. ICDM, pp. 158–167, 2010.

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

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

[21] C. J. Kowalski, *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), pp. 1–12, 1972.

[22] F. J. Anscombe, *Graphs in Statistical Analysis*, The American Statistician 27 (1), pp. 17–21, 1973.

[23] S. D. Bay, M. J. Pazzani, *Detecting Group Differences: Mining Contrast Sets*, Data Mining and Knowledge Discovery 5 (3), pp. 213–246, 2001.

[24] G. Dong, J. Li, *Efficient Mining of Emerging Patterns: Discovering Trends and Differences*, Proc. KDD, pp. 43–52, 1999.

[25] P. Kralj Novak, N. Lavrač, G. I. Webb, *Supervised Descriptive Rule Discovery: A Unifying Survey of Contrast Set, Emerging Pattern and Subgroup Mining*, Journal of Machine Learning Research 10, pp. 377–403, 2009.

[26] A. M. Jorge, P. J. Azevedo, F. Pereira, *Distribution Rules with Numeric Attributes of Interest*, Proc. PKDD, pp. 247–258, 2006.

[27] L. Umek, B. Zupan, *Subgroup Discovery in Data Sets with Multi-Dimensional Responses*, Intelligent Data Analysis 15 (4), pp. 533–549, 2011.

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

[29] W. Duivesteijn, E. Loza Mencía, J. Fürnkranz, A. Knobbe, *Multi-label LeGo — Enhancing Multi-label Classifiers with Local Patterns*, Technical Report, Technische Universität Darmstadt TUD-KE-2012-02, 2012.

[30] M. Clark, *A Comparison Of Correlation Measures*, Technical Report, University of Notre Dame, 2013.

[31] W. Hoeffding, *A Non-Parametric Test of Independence*, Annals of Mathematical Statistics 19 (4), pp. 546–557, 1948.

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

[33] M. Hollander, D. Wolfe, *Nonparametric Statistical Methods*, Wiley Series in Probability and Statistics, Wiley, 2nd ed., 1999.

[34] G. J. Székely, M. L. Rizzo, N. K. Bakirov, *Measuring and Testing Dependence by Correlation of Distances*, Annals of Statistics 35 (6), pp. 2769–2794, 2007.

[35] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, P. C. Sabeti, *Detecting Novel Associations in Large Data Sets*, Science 334, pp. 1518–1524, 2011.

[36] J. B. Kinney and G. S. Atwal, *Equitability, Mutual Information, and the Maximal Information Coefficient*, Proceedings of the National Academy of Sciences of the United States of America 111 (9), pp. 3354–3359, 2014.

[37] A. Gretton, O. Bousquet, A. Smola, B. Schölkopf, *Measuring Statistical Dependence with Hilbert-Schmidt Norms*, Proc. ALT, pp. 63–77, 2005.

[38] D. Lopez-Paz, P. Hennig, B. Schölkopf, *The Randomized Dependence Coefficient*, Advances in Neural Information Processing Systems, pp. 1–9, 2013.

[39] H. Gebelein, *Das statistische Problem der Korrelation als Variations- und Eigenwertproblem und sein Zusammenhang mit der Ausgleichsrechnung*, Zeitschrift für Angewandte Mathematik und Mechanik 21, pp. 364–379, 1941.

[40] W. J. Conover, *Practical Nonparametric Statistics*, Wiley, 1971.

[41] R. A. S. Fisher, *Statistical Methods for Research Workers*, Oliver and Boyd, 14^{th} ed., 1970.

[42] E. C. Fieller, H. O. Hartley, E. S. Pearson, *Tests for Rank Correlation Coefficients. I*, Biometrika 44 (4), pp. 470–481, 1957.

[43] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, T. Euler, *YALE: Rapid Prototyping for Complex Data Mining Tasks*, Proc. KDD, pp. 935–940, 2006.

[44] M. Lichman, *UCI Machine Learning Repository* [http://archive.ics.uci.edu/ml], Irvine, CA: University of California, School of Information and Computer Science, 2013.

[45] P. M. Anglin, R. Gençay, *Semiparametric Estimation of a Hedonic Price Function*, Journal of Applied Econometrics 11 (6), pp. 633–648, 1996.

[46] T.-S. Lim, W.-Y. Loh, Y.-S. Shih, *A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-three Old and New Classification Algorithms*, Machine Learning 40, pp. 203–228, 2000.

[47] R. A. Fisher, *The Use of Multiple Measurements in Taxonomic Problems*, Annals of Eugenics 7 (2), pp. 179–188, 1936.

[48] C. Adam-Bourdarios, G. Cowan, I. G. Cécile Germain, B. Kégl, D. Rousseau, *Learning to discover: The Higgs Boson Machine Learning Challenge*, http://higgsml.lal.in2p3.fr/documentation/, accessed August 7th, 2014.

[49] F. Keller, E. Müller, K. Böhm, *HiCS: High Contrast Subspaces for Density-Based Outlier Ranking*, Proc. ICDE, pp. 1037–1048, 2012.

[50] H. V. Nguyen, E. Müller, K. Böhm, *4S: Scalable Subspace Search Scheme Overcoming Traditional Apriori Processing*, Proc. BigData, pp. 359–367, 2013.

[51] H. V. Nguyen, E. Müller, J. Vreeken, P. Efros, K. Böhm, *Multivariate Maximal Correlation Analysis*, Proc. ICML, pp. 775–783, 2014.