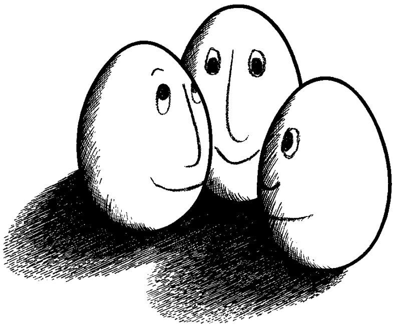
technische universität
dortmund

Bachelor-Arbeit

A Rank Correlation Model Class for Exceptional Model Mining

Lennart Downar



Technische Universität Dortmund
Fakultät Informatik
Lehrstuhl für Künstliche Intelligenz (LS-8)
http://www-ai.cs.uni-dortmund.de/

Dortmund, September 22, 2014

Betreuer:

Prof. Dr. Katharina Morik
Dr. Wouter Duivesteijn

Contents

List of Figures

List of Tables

1 Introduction

Since the rise of (personal) computers, the amount of data available to the public and researchers has been increasing dramatically, potentially giving us access to new insights and information. This growth has been fueled further by the recent advances of smartphones and embedded devices. Together with the amount, the complexity of data to be analyzed has increased. However these masses of data can also be a burden, as it might be impossible even for enthusiastic experts to sift through all the available data to find new insights. Therefore data mining techniques are needed to help analyze large amounts of data. One particular field of work is Local Pattern Mining, which finds patterns in data based on learning algorithms or statistical measures. This can be extremely helpful for domain experts as it can provide hints to possibly valuable/previously hidden information in the data, that might be interesting for research (e.g. research on gene expressions)

There are already many approaches to find and generate such interesting pre-selections. One of those is Subgroup Discovery (*SD*) which has recently been generalized into Exceptional Model Mining (*EMM*).

1.1 Aim of this thesis

In this thesis we are trying to extend the Correlation Model Class for *EMM* proposed by Leman et al. [1] to use non-parametric rank correlation measures and investigate suitable quality measures. Instead of the standard Pearson correlation coefficient, which can only measure linear relationships, we will apply rank correlation coefficients. The theoretical advantage is, that they measure the extent to which, as one variable increases, the other variable tends to increase without requiring that increase to be represented by a linear relationship. In Figure 1.1 we can see that some function could be fitted perfectly on the data. For such cases we would wish for a correlation measure value of 1 (or -1 depending if it is monotonically rising or falling), indicating a perfect monotonic function. A rank-based correlation measure such as Spearman's correlation fulfills this wish while the standard Pearson correlation only returns a value of 0.93 indicating a non-perfect linear relationship.

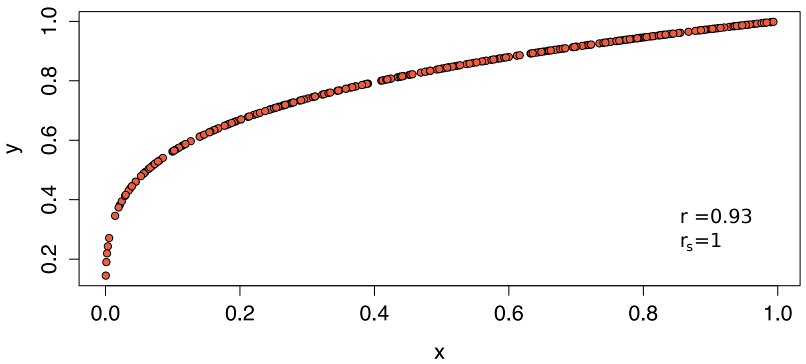


Figure 1.1: Comparing Pearson correlation with Spearman's rank correlation

We perform experiments to see if the theoretical advantages of a non-parametric measure have an impact in practice, i.e., whether we will find the same, similar or entirely different subgroups than with the standard correlation model class.

1.2 Structure

In Chapter 2 we describe the concept of Exceptional Model Mining as a generalization of Subgroup Discovery. Chapter 3 focuses on related work concerning *EMM* and correlation measures. In Chapter 4, we present 5 widely used and recognized correlation measures and choose 3 of them to develop a new Rank Correlation Model Class. Chapter 5 gives a quick overview on our implementation in RapidMiner. The chosen datasets as well as the results of our experiments can be found in Chapter 6. Finally, Chapter 7 presents a conclusion as well as a future outlook.

2 Exceptional Model Mining

2.1 Introduction

Data Mining is the scientific field of computer-aided extraction of useful, interesting or previously unknown information/patterns from (large) databases. The algorithms usually found in Data Mining are data-driven, meaning, not the algorithm alone decides what a good “solution” is, but that the data itself will indicate the direction.

Two different approaches to learning from data are supervised and unsupervised learning. In unsupervised learning, there is no feedback or labeling of “correct” instances, i.e., we do not tell the algorithm anything about the data. An unsupervised learning scheme has to discover the knowledge or hidden structures in the data by itself. In supervised learning on the other hand, there is some sort of feedback/measure or labeling for the data available. The algorithm then tries to infer relationships between variables of the data; or functions that could be used to map new examples.

When applying those approaches we should also keep in mind the differences between two concepts: local patterns and (global) models. A local pattern describes (small) subsets of a given dataset (e.g., people who buy milk, also often buy coffee), while a model is fitted on the whole data to describe some specific relationship (e.g., a regression line fitted on the age of a person and their income).

2.1.1 Local Pattern Mining

A standard and important task for Data Mining is to identify elements that differ from the norm, which techniques try to achieve by focusing on outlier detection. Another approach is Local Pattern Mining (*LPM*), where the goal is to find subgroups in the data. Subgroups are subsets of a dataset, that can be described by some conditions imposed on the attributes of the dataset.

The availability of descriptions for a subgroup makes them much more usable and interesting as Duivesteijn points out in [2]: “if we tell a pharmaceutical company that five given persons react badly to a certain type of medication, it is more difficult for them to act on the information than it would be if we could tell them that the group of smokers react badly to the medication.” What exactly constitutes a divergence from the norm can be defined in different ways, as we will see.

Unsupervised Local Pattern Mining

A traditional example for Local Pattern Mining is Frequent Itemset Mining [3], which was originally developed to analyze customer behavior regarding the purchased products. In this setting a frequent set of products describes how often specific items are bought together. One could for example find the subgroup of customers of a supermarket, that simultaneously buy coffee and milk. In *LPM* we thus have no designated target attribute; it is an *unsupervised* method.

2.1.2 Subgroup Discovery

The simplest form of *supervised* Local Pattern Mining is *Subgroup Discovery* [4] (*SD*), which is concerned with finding interesting or exceptional subgroups (i.e. relations between properties or variables) in a population with respect to a single target attribute. The target attribute is the value of interest, meaning we typically try to find subgroups where the distribution of the target attribute is significantly different from its distribution in the whole dataset.

In general, the interestingness of a subgroup is determined by a quality measure (e.g., distribution of the target attribute), which is defined to measure the difference between a potentially interesting subgroup and the complement (or the whole dataset).

“Since unusual distributions are more easily achieved in small subsets of the dataset, the typical quality measure also contains a component indicating the size of the subgroup. Thus, whether a description is deemed interesting depends on both its exceptionality and the size of the corresponding subgroup” [2].

Example: in [4], a dataset with four attributes is considered:

- Age = {Less than 25, 25 to 60, More than 60};
- Gender = {M, F};
- Country = {Spain, USA, France, Germany};
- Money = {Poor, Normal, Rich} ← target variable.

With respect to the target attribute “money”, traditional SD could find the following rules:

- R_1 : (Age = Less than 25 AND Country = Germany)→ Money=Rich;
- R_2 : (Age = More than 60 AND Gender = F)→Money=Normal.

Here, R_1 stands for a subgroup of German people less than 25 years old, for which the probability of being rich is unusually high compared to the rest of the population, and R_2 represents that women of more than 60 years old are more likely to have normal wealth than the rest of the population.

2.2 Exceptional Model Mining

2.2.1 Overview

In *SD* we find subgroups where the target attribute shows an unusual distribution. However, as already mentioned, we could also think of applying and comparing models on the data. This leads to a generalization of Subgroup Discovery: Excpetional Model Mining. With *EMM* we can find subgroups of the data where multiple target attributes show an unusual distribution. This is achieved by applying a model (e.g., regression) on the target attributes and comparing subgroups based on the fitted model - hence Exceptional Model Mining.

An Exceptional Model Mining Class consists of a fixed model (e.g., regression) and its model parameters, which vary depending on how the model is fitted on the data (e.g., slope value for a fitted regression model). We then strive to identify descriptions of subgroups, for which the model parameters deviate considerably from those of the model built from the entire or complement dataset. "Formally this is accomplished by using an exceptionality measure that maps a subgroup (pattern) to a real number corresponding to its quality (interestingness) based on its model parameters"[5].

2.2.2 Example

A simple example for *EMM* has been given by the authors of [1]; we consider a simple linear regression model:

$$P_i = a + bS_i + e_i$$

Where P_i is the sales price of a house, S_i is the lot size, and e_i the random error term. One could now fit this model to a subgroup of the dataset, for example a group of houses situated in a desirable location (what makes a location desirable is of course debatable). We could then perform a statistical test to gauge whether the slopes of the two fitted models are significantly different.

2.2.3 Comparison

To illustrate the differences of *LPM*, *SD* and *EMM*, we compare them in terms of the found subgroups/rules. We can then again see that by allowing more than one target attribute, *EMM* is simply a generalization of *SD*.

Unsupervised Local Pattern Mining:

- no target attribute;
- example: find a subgroup of customers of a supermarket, that simultaneously buy coffee and milk.

Subgroup Discovery:

- one target attribute;
- example: find a subgroup of smokers, whose lung cancer incidence is above average.

Exceptional Model Mining:

- multiple target attributes;
- example: find a subgroup of inner city houses, for which the correlation between price of a house and its lot size is substantially weaker than for the average house.

2.2.4 Definitions

We will introduce the definitions necessary to formalize model classes for *EMM*.

In general we will have a dataset with several attributes from which we pick our target attributes. The remaining attributes are used for subgroup description.

Definition 1
A dataset Ω *is a bag of* N *records* $r \in \Omega$ *of the form:* $r = (a_1, ..., a_k, l_1, ..., l_m)$

We call the attributes $a_1, ..., a_k$ *the descriptive attributes and attributes* $l_1, ..., l_m$ *the target attributes.*

To define a subgroup we first need a way of describing it. In general this is achieved by using a description language $\mathcal{D}$ from which we can build descriptions D.

Definition 2
A description is a function $D : (a_1^i, ..., a_k^i) \to \{0, 1\}$

A description D *covers a record* r *if and only if* $D(a_1^i, ..., a_k^i) = 1$

With the description we indirectly already have a subgroup, as it simply consists of the records covered by the description.

Definition 3
A subgroup corresponding to a description D *is the bag of records* $G_D \subseteq \Omega$ *that* D *covers:*

$G_D = \{r^i \in \Omega \mid D(a_1^i, ..., a_k^i) = 1\}$

To evaluate a candidate description in a given dataset, we use a quality measure to gauge how interesting the inferred subgroup is.

Definition 4
A quality measure is a function $\varphi : \mathcal{D} \to \mathbb{R}$ *that assigns a unique numeric value to a description* D

2.2.5 Defining the task/problem

As Duivesteijn [2] remarks: "The goal is to find interesting subgroups of a dataset, for whatever instantiation of "interesting" the user of *EMM* cares for, which is intrinsically subjective. Therefore, any formal definition of the *EMM* task will only concern a subset of what we attempt to achieve with *EMM*." Nevertheless for a formal discussion it is necessary to provide a general definition. An attempt of doing so is described in [2]:

Top-q Exceptional Model Mining

Given a dataset Ω, a description language $\mathcal{D}$, a quality measure φ, a positive integer q and a set of constraints $\mathcal{C}$. The Top-q Exceptional Model Mining task delivers the list $\{D_{1,...}, D_q\}$ of descriptions in the language $\mathcal{D}$ such that:

- $\forall\, 1 \leq i \leq q : D_i$ satisfies all constraints in $\mathcal{C}$
- $\forall\, i, j : i < j \Rightarrow \varphi(D_i) \geq \varphi(D_j)$
- $\forall\, D \in \mathcal{D} \backslash \{D_1, ..., D_q\} : D$ satisfies all constraints in $\mathcal{C} \Rightarrow \varphi(D) \leq \varphi(D_q)$

Constraints that can be imposed are e.g., a minimum support level, a minimum threshold for the quality measure or the complexity of the subgroup description.

Comparing subgroups

As already briefly mentioned in Section 2.2.1, a subgroup $G_\mathcal{D}$ is only of interest if it differs in some way from its complement $G_\mathcal{D}^C$ or the whole dataset Ω. The question remains, to which one should it actually be compared? Duivesteijn [2] notes that there is no general answer to this question. However the choice is not arbitrary, as the real-life problem can give us some directions for this decision. Suppose we are interested in deviations from the norm, then we should compare to Ω. If we are more interested in finding schisms in the dataset, a comparison to $G_\mathcal{D}^C$ makes more sense, as this implies a partitioning of Ω.

Sometimes the model class itself can dictate an answer: if learning models from the data has a nontrivial computational expense, with regards to efficiency we might not have a free choice, because "when comparing n descriptions to Ω, learning $n+1$ models suffices, but when comparing them to $G_\mathcal{D}^C$, learning $2n$ models is required". Furthermore "a statistically inspired quality measure may require choosing either Ω or $G_\mathcal{D}^C$, to prevent violation of mathematical assumptions" [2].

2.2.6 Basic Approach

For the top-q Exceptional Model Mining task we need:

- a refinement operator that generates possible subgroups;
- an algorithm that traverses the space of possible subgroups (heuristic or exhaustive);
- a model class, defined by a fixed model (e.g., regression) and its model parameters, which vary depending on how the model is fitted on the data.

Multiple target concepts can be explored by exchanging the model class.

Refining Subgroups

A description is created by conjunctions of basic conditions provided by the chosen description language $\mathcal{D}$. In our setting these are equalities (=) and inequalities ($\neq$,$\leq$,$\geq$). Refining a description is achieved by simply adding a new condition to the description. This can be done by a refinement operator $\eta : \mathcal{D} \to 2^{\mathcal{D}}$. Usually η will be a specialization operator, meaning “that every description D_i that is an element of the set $\eta(Dj)$, is more specialized that the description Dj itself” [2].

We start with the empty description; a new set is built by looping over the descriptive attributes a_i and the specialization is generated depending on the attribute type.

We can formalize this as follows [2]:

- if a_i is binary: add $D \cap (a_i = 0)$ and $D \cap (a_i = 1)$ to $\eta(D)$;
- if a_i is nominal, with values $\nu_1, ..., \nu_g$: add $\{D \cap (a_j = \nu_j), D \cap (a_j \neq \nu_j)\}_{j=1}^{g}$ to $\eta(D)$;
- if a_i is numeric: order the values of a_i that are covered by the description D; this gives us a list of ordered values $a_{(1)}, ..., a_{(n)}$ ($n = \mid G_{\mathcal{D}} \mid$). From this list we select the split points $s_1, ..., s_{b-1}$ letting
$$\forall_{j=1}^{b-1} : s_j = a_{(\lfloor j\frac{n}{b} \rfloor)}$$
Then, add $\{\mathcal{D} \cap (a_{(i)} \leq s_j), \mathcal{D} \cap (a_{(i)} \geq s_j)\}_{j=1}^{b-1}$ to $\eta(\mathcal{D})$.

However, it should be noted, that these are not the only ways of refining subgroup descriptions. Generating optimal subgroup descriptions is still subject of ongoing research. Hence more sophisticated approaches are also explored, e.g., by Mampaey et al. [6] who introduce linear-time algorithms for finding optimal sets for *Set-valued conditions* (e.g., *Country* $\in$ {*Spain, France*}) and *Interval conditions* (e.g., *Age* $\in [25, 60]$). However, as these “algorithms operate by only considering subgroup refinements that lie on a convex hull in ROC space” [6], they are only applicable for concepts that can be described in ROC space. Thus, this approach does not fit our scenario, where we have multiple (numeric) targets.

Algorithm

The procedure is as follows: we select the descriptive attributes to generate subgroup descriptions. On each level, the best-ranking w patterns, that fulfill the constraints $\mathcal{C}$, are refined by the specified refinement operator η to form the candidates for the next level. Once a specified level d of search depth is reached, we report the q best-ranked descriptions.

Algorithm 1 Top-q Exceptional Model Mining

```
Input: $\Omega$, $\varphi$, $\eta$, $w$, $d$, $q$, $\mathcal{C}$
Output: resultSet
candidateQueue $\leftarrow$ new Queue
resultSet $\leftarrow$ new PriorityQueue(q);
for (int level $\leftarrow$ 1, level $\leq$ d, level++) do
    beam $\leftarrow$ new PriorityQueue(w);
    while candidateQueue$\neq \emptyset$ do
        seed $\leftarrow$ candidateQueue.dequeue();
        set $\leftarrow$ $\eta$(seed);
        for all (desc $\in$ set) do
            quality $\leftarrow$ $\varphi$(desc);
            if (desc.satisfiesAll(C)) then
                resultSet.insertWithPriority(desc,quality);
                beam.insertWithPriority(desc,quality)
    while beam $\neq \emptyset$ do
        candidateQueue.enqueue(beam.getFrontElement());
return resultSet;
```

3 Related Work

3.1 Exceptional Model Mining

Subgroup Discovery is a Data Mining technique that has undergone thorough research. It has first been introduced by Klösgen [7] and Wrobel [8] and has since then been refined through new applications and strategies. A broad overview on the history of Subgroup Discovery, developed algorithms and quality measures has been published by Herrera et al. [4].

The Exceptional Model Mining framework [1] has been proposed as a generalization of Subgroup Discovery. Exceptional Model Mining is thus a very recent development. A thorough introduction, several model classes and quality measures have been given in [2].

3.1.1 Existing EMM Model Classes

Correlation Model Class

The correlation model class computes Pearson's Correlation Coefficient between two attributes for possible subgroups and their complements. A significance test on the differences in the coefficient r is used as a quality measure.

Classification Model Class

As a more complex instance, the classification model class allows classifiers, such as decision trees, support vector machines or ensembles of classifiers, as a basis for measuring the quality of a subgroup. The goal is to find descriptions for which a classifier learned from the targets has deviating performance.

This instance of EMM can deliver important indications for data miners when a developed algorithm works particularly well, or when its performance is not satisfactory. The knowledge can be incorporated by classification algorithm developers to improve their algorithms. In [2, Chapter 5] logistic regression has been used to exemplify the application of the classification model.

Bayesian Network Model Class

The Bayesian network model class allows multiple nominal targets between which a Bayesian network is then learned. A description is deemed interesting, when the conditional dependence relations between the targets are substantially different from the description from these relations on the whole dataset. The descriptions are therefore validated on the interdependencies between the targets, instead of the target values themselves.

Regression Model Class

The regression as model class seeks descriptions for which (a subset of) the regression parameter vector β significantly deviates from the parameter vector estimated on the whole dataset.

3.1.2 Search strategies: heuristic or exhaustive?

One can imagine that the space of possible subgroups is potentially large and thus it is important to consider how this space can be explored. In Section 2.2.6 we have already presented a way of doing so in a heuristic way, where only a subset of all possible subgroups is actually explored. The advantage is achieved by not just arbitrarily selecting some of them, but choosing several “good” candidates and refining them further. This way we can avoid getting stuck in a local optima.

An exhaustive alternative has been developed, using a model class based on Generic Pattern Trees by Lemmerich et al. [5]. For their GP-Growth algorithm (based on FP-Growth) a special data structure, called *valuation basis*, is needed for every model class. As developing such a valuation basis for rank correlation is beyond the scope of this Bachelor thesis, we will leave this for future work. Van Leeuwen [9] has proposed the EMDM (Exception Maximisation and Description Minimisation) algorithm, another (iterative) alternative to the beam-search based approach presented in [1]. Along with EMDM, two additional quality measures (based on KL divergence and KRIMP) are described. However, the average “description complexities vary from 5 up to 25 conditions” [9]. Resulting in “more complex subgroups than are typically found” [9], which makes them somewhat uninspectable.

3.2 Measurements of correlation

A comparison of several correlation measures has been given by Clark [10]. Apart from Pearson's r and Spearman's r_s, he examines three other measures, which promise to measure relationships beyond linear and monotonic behavior. Examples for datasets that exhibit such behavior can be seen in Figure 3.1, where Pearson would only be able to detect paterns 2 and 3, and to some extent 4 and 5.

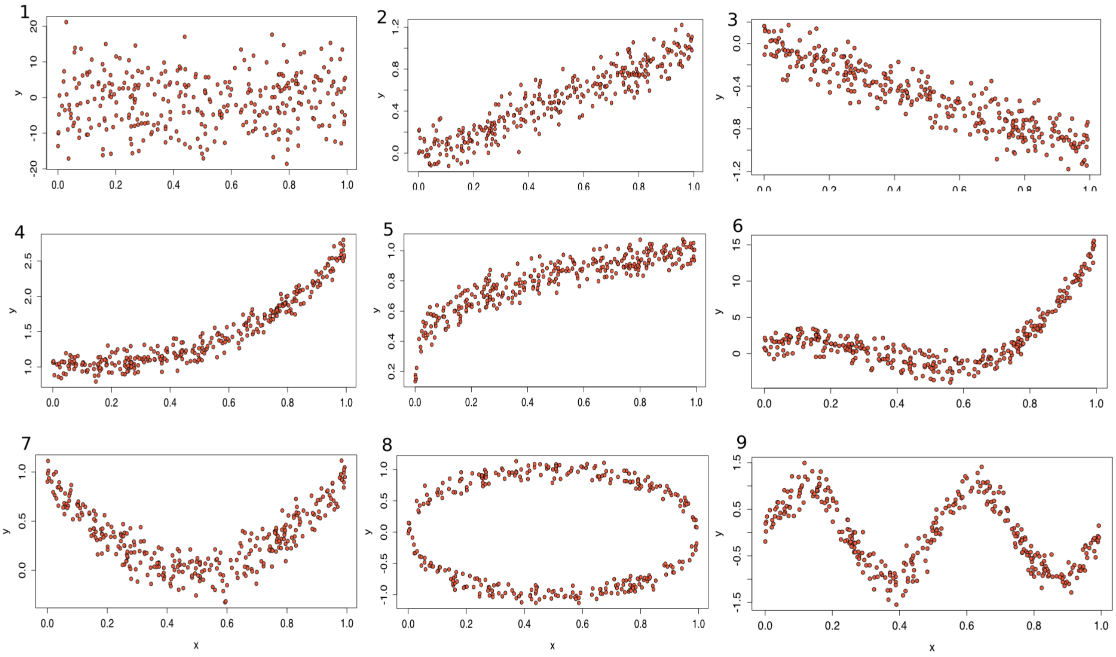


Figure 3.1: Collection of datasets with various dependency structures

Hoeffding's D

Contrary to sample correlation coefficients such as Spearman's r_s, Kendall's τ, and Pearsons's r, Hoeffding [11] developed a test of independence that can be used to detect a much broader class of dependence structures beyond monotonic association. Hoeffding's statistic, denoted as D, is non-parametric and, similar to Spearman and Kendall, based on ranks. For two random variables X and Y (with R_i and S_i denoting the ranks of X_i and Y_i respectively) D is defined as:

Definition 5

$$
\begin{aligned}
D &= \frac{Q - 2(n-2)R + (n-2)n-3)S}{n(n-1)(n-2)(n-3)(n-4)} \\
Q &= \sum_{i=1}^{n}(R_i - 1)(R_i - 2)(S_i - 1)(S_i - 2) \\
R &= \sum_{i=1}^{n}(R_i - 2)(S_i - 2)c_i \\
S &= \sum_{i=1}^{n} c_i(c_i - 1) \\
c_i &= \sum_{\alpha=1}^{n} \phi(X_\alpha, X_i)\phi(Y_\alpha, Y_i), \textit{ where } \phi(a,b) = \begin{cases} 1 & \textit{if } a < b \\ 0 & \textit{if } a \geq b \end{cases}
\end{aligned}
$$

A similar statistic proposed by Blum et al. [12] can be used as a large-sample approximation for D [13].

Distance Correlation

Distance correlation (*dCor*) has been introduced by Székely et al. [14] to address the deficiency of the Pearson correlation coefficient regarding non-linear relationships. It is based on distance matrices for the X and Y variables and can take values between 0 and 1. According to Clark, a ranked-based version *dCor* could also be incorporated.

Maximal Information Coefficient

Reshef et al. [15] have developd 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.

Evaluation

After comparing these alternatives on several non-linear relationships Clark notes, that "Hoeffding's D only works in some limited scenarios". In the experiments D did pick up some of the non-linear relationships (e.g., a quadratic relationship or a circle pattern) however, the computed values were relatively small (mean ranging from 0 to 0.1). This became even worse when noise was added to the data (mean ranging from 0 to 0.02). So even though it does pick up some non-linear relationships, due to the small values one cannot get a good sense of the measured dependence.

dCor and *MIC* performed better at finding various relationships 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 relationships, but they are not perfect either. Some additional problems with *MIC* are also described by Kinney and Atwal [16]. Consequently, Clark also concludes 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”.

Other approaches

As pointed out by Clark [10], the development of satisfying general dependence measures that go beyond simple forms of relationships is still far from finished. Other approaches therefore have been introduced in recent years. Gretton et al. [17] 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. [18] proposed the Randomized Dependence Coefficient (*RDC*), which is an estimate of the Hirschfeld-Gebelein-Rényi Maximum Correlation Coefficient (*HGR*). *HGR* was defined by Gebelein [19] in 1941, however it is not computable and thus represents only an abstract concept.

All of the aforementioned measures have in common that they are mostly recent developments and as such, further investigation into whether they truly fulfill their promises is needed. Additionally there is, to our knowledge, not yet a statistical way of comparing results of these measures on different data samples, thus making it difficult to formulate a mathematically well-supported quality measure based on them.

4 The Rank Correlation Model Class

4.1 Motivation for a new model class

Detecting a correlation between some variables can lead to useful insights. While the detection alone might not directly explain a potential underlying connection, domain experts can try to infer and analyze the reasons behind it. Even if there is no clear explanation, knowing that two things are in some way correlated can be useful. Especially in medicine, lots of treatments are based on the correlation between their usage and the healing of some illness.

As briefly highlighted in Section 3.1.1 a correlation model class for EMM exists [2, Chapter 4], which uses Pearson's standard correlation coefficient as a model to measure the exceptionality of a subgroup. However, this implies that the data follows a normal distribution, as experiments have shown that the distribution of r is sensitive to non-normality [20]. Without the normality assumptions, many statistical tests on r therefore become meaningless or at least hard to interpret. Considering that normality cannot be assumed for many real-life examples and datasets, its is questionable if Pearson's r is then still a suitable measure. The normality assumption therefore limits (at least theoretically) the scope of application for this model class. Furthermore, r is easily affected by outliers and in general only captures linear relationships between targets.

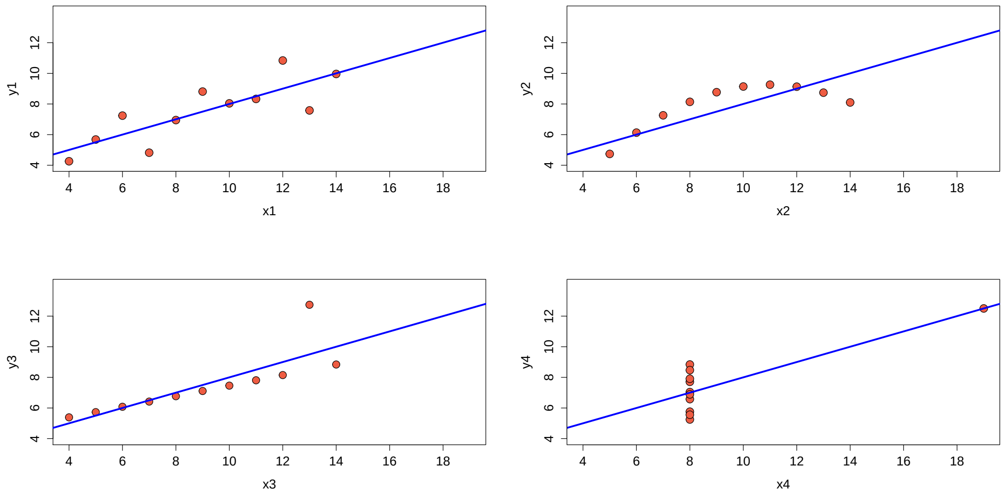


Figure 4.1: Anscombe's Quartet

These limitations can very well be illustrated with Anscombe's quartet [21], displayed in Figure 4.1, 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 very well in sets 3 and 4, two datasets with a clearly different relationship between the two displayed variables.

Figure 4.2 gives some examples of datasets and their Pearson coefficient, for several non-linear relationships not captured by r. However, one should keep in mind that rank correlation does not promise to recognize these relationships either, but can only expand upon Person's limitations by additionally capturing general monotonic relationships.

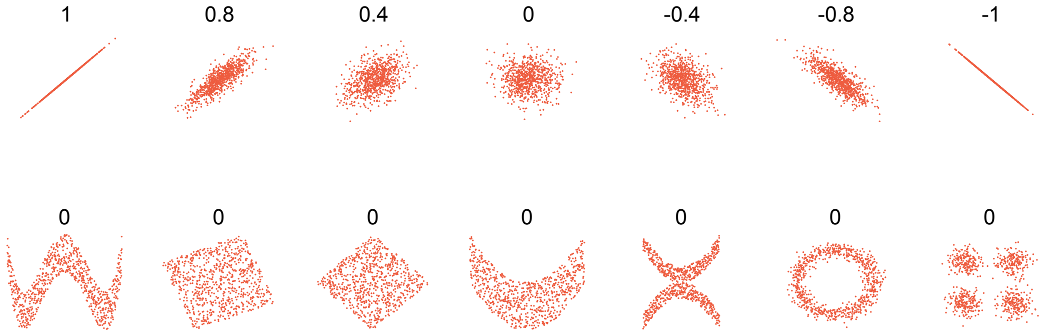


Figure 4.2: Various datasets and their respective Pearson coefficients

The question therefore is, if we can develop a model class that uses correlation as quality measure, but has fewer limitations and assumptions on the distribution of the target data.

Spearman's r_s [22] or Kendall's τ [23] both offer a measurement of the rank correlation, which can measure whether two targets have a monotonic relationship, thus making them also capable of capturing some non-linear relationships. Additionally they don't rely on assumptions for the distribution of targets. Another measurement of rank correlation is Goodman and Kruskal's γ [24]. We investigate if there are relevant differences between these coefficients, that would make one preferable over to other for a model class based on rank correlation.

4.2 Standard Correlation Measures

In general, the *population correlation coefficient* is denoted by ρ. However, it is never truly known, and thus many researchers have tried to develop estimators based on a sample from the population. We present some of the most widely used and established sample correlation coefficients in the following sections.

4.2.1 Pearson's r

Probably the most commonly used method to compute correlation is Pearson's product-moment correlation coefficient, often denoted as r.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}} \tag{4.1}$$

4.2.2 Spearman's r_s

Spearman's rank correlation coefficient (usually denoted as ρ but also as r_s, we will use r_s to avoid confusion with the *population correlation coefficient* of a dataset) has been developed by Charles Spearman [22] and uses the difference between rankings of a pair x_i and y_i as a statistic to measure the (rank) correlation.

$$r_s = 1 - \frac{6\sum_i d_i^2}{n(n^2 - 1)} \tag{4.2}$$

Where d_i is the difference between the ranks of x_i and y_i

If there are no ties 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:

$$r_s = \frac{\sum(R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum(R_i - \bar{R})^2 \sum(S_i - \bar{S})^2}} \tag{4.3}$$

In the case of ties, Conover [25] suggest using equation 4.3. Though if the number of ties is moderate, equation 4.2 will still function a good approximation and should be preferred due to its computational simplicity.

4.2.3 Kendall's τ

While Spearman uses the difference of rank in individual pairs, Kendall's τ, named after Maurice Kendall [23], defines a statistic based on the agreement (concordances) of ranks to measure the correlation of a sample, making it less sensitive to outliers.

Definition 6
A pair $(x_i, y_i), (x_j, y_j)$ is said to be concordant *if $(x_i < x_j) \land (y_i < y_j)$ or $(x_i > x_j) \land (y_i > y_j)$ hold, and* discordant *if this is not the case. Two observations are tied if $x_i = x_j$ and/or $y_i = y_j$.*

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:

Definition 7

$$\begin{aligned}
C &= \textit{number of concordant pairs}\\
D &= \textit{number of discordant pairs}\\
T_x &= \textit{number of pairs tied only on the x-value}\\
T_y &= \textit{number of pairs tied only on the y-value}\\
T_{xy} &= \textit{number of pairs tied both on the x- and y-value}
\end{aligned}$$

Therefore, we can decompose M as: $M = C + D + T_x + T_y + T_{xy}$.

We will see that several correlation measures exist, which make use of the numerator C-D but differ in the normalizing denominator.

Tau a

τ_a represents the surplus of concordant pairs over discordant pairs, as a percentage of all pairs. Because ties are not taken into account, τ_a is rarely used.

$$\tau_a = \frac{C-D}{M} \tag{4.4}$$

Tau b

τ_b is the most widely applied version of Kendall's measure. In contrast to τ_a, it accounts for ties by normalizing with a term representing the geometric mean between the number of pairs not tied on the x-value (i.e., $C+D+T_y$) and the number not tied on the y-value (i.e., $C+D+T_x$).

$$\tau_b = \frac{C-D}{\sqrt{(C+D+T_x)(C+D+T_y)}} \tag{4.5}$$

Tau c

If applied to asymmetric contingency tables, τ_b cannot reach -1 or 1 anymore, making its interpretation difficult. τ_c is an adjustment to this limitation, with k being the minimum of the number of rows and columns.

$$\tau_c = \frac{2k(C-D)}{(k-1)n^2} \tag{4.6}$$

4.2.4 Goodman and Kruskal's γ

Gamma represents the surplus of concordant pairs over discordant pairs, as a percentage of all pairs, ignoring ties.

$$\gamma = \frac{C - D}{C + D} \tag{4.7}$$

4.2.5 Somers' D

Somers' D is a modification of τ_b, developed by Somers [26]. If there are no ties, D is equal to Goodman and Kruskal's γ.

Asymmetric

For the asymmetric version only one variable is considered to be independent, while the other one is considered to be dependent. The asymmetric version of Somers' D differs from τ_b in that it only adjusts for tied pairs on the independent variable.

$$D = \frac{C - D}{C + D + T_x} (X\ dependent) \tag{4.8}$$

$$D = \frac{C - D}{C + D + T_y} (Y\ dependent) \tag{4.9}$$

Symmetric

The symmetric version of Somers' D differs from τ_b in that it uses the arithmetic mean between the number of pairs not tied on the x- and y-value to account for ties, instead of the geometric mean.

$$D = \frac{C - D}{C + D + (T_x + T_y)/2} \tag{4.10}$$

4.2.6 Comparison

Several measures use the difference in concordant and discordant pairs to gauge correlation, but differ in the way they normalize this statistic. We choose Kendall's τ_b to represent those measures, as it is the most prevalent and takes ties into account.

Although it is common to regard the presented rank correlation coefficients as alternatives to Pearson's coefficient, as we can see by the definitions above, this view has little mathematical basis. They all measure different types of relationships than the Pearson product-moment correlation coefficient and we will therefore keep in mind that these coefficients should rather be seen as measures of a different type of association.

4.3 Rank Correlation as a Quality Measure for EMM

4.3.1 Defining a Quality Measure

In order to define a quality measure we have to answer the question: when is a subgroup more or less interesting than another?

A simple idea could be the 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 it is. However, as indicated in Section 2.1.2, a quality measure should also consider the support of the subgroup (i.e., the number of records it covers) to prevent overfitting. As it is usually relatively easy to generate small subgroups with extreme correlation values ($-1, 1$ or 0) and thus probably also creating extreme difference values. When we get a big difference in correlation we would also like to know if it is significant at all. What we want to test is thus :

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

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.

Definition 8
Given a Pearson correlation coefficient r, Fisher's z-transformation is defined as:

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

The transformed value z is normally distributed with variance $\mathrm{var_z} = \frac{1}{n-3}$.

We can then treat difference between the transformed values as a random normal variable, with mean zero and variance $\mathrm{var}_{\rho_1\text{-}\rho_2} = \frac{1}{n_1-3} + \frac{1}{n_2-3}$. By comparing it with a standard normal distribution, a p-value for the difference can the 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 [27].

Fieller [28] has tried to transfer this approach for comparisons of Kendall's τ and Spearman's r_s. His experiments suggested the following variances for the transformed values:

$$\mathrm{var_{r_s}} = \tfrac{1.06}{n-3} \text{ and } \mathrm{var}_\tau = \tfrac{0.437}{n-4}$$

Limitations?

A direct limitation for applying this model class stems from the fact that we are using correlation as a quality measure, where target attributes can only be numeric. Datasets are thus restricted to have at least two numerical attributes.

Since our test of difference is only applicable to independent datasets, we have to compare a subgroup with its complement, as comparing with the whole dataset (which includes the subgroup data) would violate the independence assumption.

A limitation of the Pearson coefficient is that it assumes a bivariate normal distribution over the two variables to compare. While the other presented correlation measures do not have this assumption, indirectly it comes into play again when applying the slight modifications of the Fisher z-transformation presented in [28], 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 thus notes for bigger samples that this “is a field in which further investigation would be of considerable interest” [28, Page 3]. Sadly, to our best knowledge, almost 50 years later, no further investigation has happened as of today.

5 Implementation

We choose to implement the (rank) correlation model class for *EMM* in the popular data mining platform RapidMiner (`http://www.rapidminer.com/`). This *EMM* extension for RapidMiner is available under: `https://bitbucket.org/lennardo/rancor-emm`

Apart from the direct implementation of the top-q Exceptional Model Mining algorithm, important components of the implementation are:

QualityMeasure: an Interface for defining your own quality measure.

EMMCondition: a comparable single condition for our subgroup descriptions, extending on RapidMiner's *AttributeValueFilterSingleCondition* class.

RefinementOperator: an interface that can be used to define how a subgroup should be refined.

Subgroup: a class encapsulating all relevant information like the coverage, measure value and the list of conditions describing a subgroup.

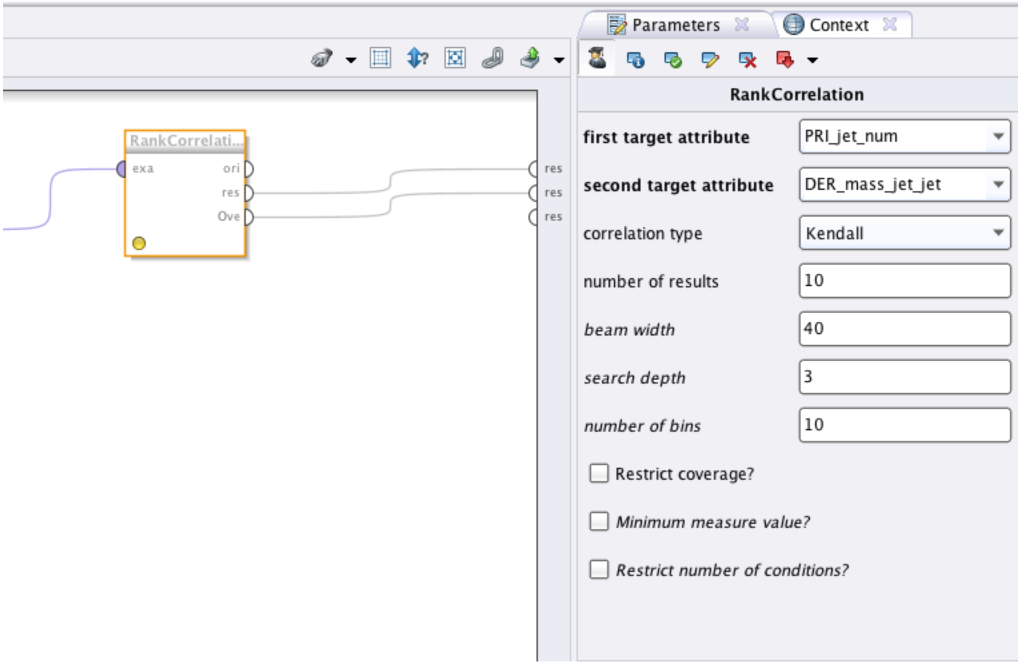
5.1 Computation of Kendall's tau

A direct computation of the numerator for Kendall's τ would require two nested iterations, resulting in a complexity of $\mathcal{O}(n^2)$. A faster algorithm for computing Kendall's τ has been developed in [29]. The approach is built on the concept of merge sort and has a resulting complexity of $\mathcal{O}(n \cdot log\ n)$. Other efficient ways (but none better than $\mathcal{O}(n \cdot log\ n)$) are detailed in [30]. For our implementation we choose Knight's algorithm.

5.2 RapidMiner Operator

The Operator can be configured via the following parameters:

- first and second target attribute;
- correlation type;
- number of results;
- min and max support level;
- min and max number of conditions defining the subgroups.



RankCorrelation

first target attribute	PRI_jet_num
second target attribute	DER_mass_jet_jet
correlation type	Kendall
number of results	10
beam width	40
search depth	3
number of bins	10

□ Restrict coverage?

□ *Minimum measure value?*

□ *Restrict number of conditions?*

Furthermore the following "expert" parameters are available for users who are familiar with the underlying algorithm

- beam width;
- search depth;
- number of bins (for refining numeric attributes);
- defining limits for the measure value.

6 Experiments

6.1 Design

For the experiments several datasets have been used to compare the newly developed model classes with the standard Pearson Correlation Model Class. We choose to compare the results of the existing Pearson correlation class, Spearman's r_s and Kendall's τ_b. Goodman/Kruskal's γ was not considered; as shown in Section 4.2 it is quite similar to τ_b in its interpretation and computation. We therefore have r_s and τ_b representing two different standard approaches of calculating rank correlation values.

6.2 Datasets

To test and evaluate the proposed model classes and their different quality measures, experiments based on datasets like the Iris dataset offered by the UCI repository [31] were performed. The goal was to investigate differences and to see if the theoretical benefits had any impact in practice, i.e., if the proposed model classes could find (relevant) subgroups on additional/fewer underlying concepts.

6.3 Experimental Results

6.3.1 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 prize it was sold for, number of bathrooms or whether the house was located in a preferable area.

The experimental results for the Spearman and Pearson measures confirm the experiments performed on the Windsor Housing dataset in [1], as both return the description:

$$D_1 : \mathit{fb} >= 2 \land \mathit{rec} = 1 \land \mathit{drv} = 1$$

Which describes a group of 35 houses that have a driveway, a recreation room and at least two bathrooms. Leman et al. reason that D_1 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.”

Subgroup	φ	r	n
fb <= 2 ∧ drv = 1 ∧ sty <= 2	0.99993	0.4740	383
bdms >= 3 ∧ rec = 1 ∧ drv = 1	0.99992	0.1186	77
fb >= 2 ∧ rec = 1 ∧ drv = 1	0.99989	-0.0894	35

Table 6.1: Housing: top-3 subgroups for Pearson

Subgroup	φ	ρ	n
fb >= 2 ∧ rec = 1 ∧ drv = 1	0.9999823	-0.1385	35
fb <= 1 ∧ drv = 1 ∧ ca = 0	0.9999821	0.4319	247
fb >= 2 ∧ rec = 1 ∧ bdms >= 3	0.9999781	-0.0932	36

Table 6.2: Housing: top-3 subgroups for Spearman

Subgroup	φ	τ_b	n
fb = 1 ∧ drv = 1	1	0.370	341
bdms <= 3 ∧ drv = 1 ∧ reg = 0	1	0.3329	277
bdms <= 3 ∧ drv = 1 ∧ ca = 0	1	0.31993	261

Table 6.3: Housing: top-3 subgroups for Kendall

6.3.2 Contraceptive Method Choice

This dataset is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey. The dataset contains 1473 samples of married women who were either not pregnant or did not know if they were at the time of interview.

One assumption 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 results from both Pearson and Spearman are similar and describe women with high education that employ long term contraception methods, thus supporting our assumption of correlation between education and employed contraception method.

Kendall finds description for groups, where education in both partners is lower. The correlation between a womens age and the number of children born in those groups is less then in their complements, confirming our assumption again, but from another angle.

Subgroup	φ	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

Table 6.4: Contraception: top-3 subgroups for Pearson

Subgroup	φ	r	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

Table 6.5: Contraception: top-3 subgroups for Spearman

Subgroup	φ	r	n
Wifes_edu <= 3 ∧ Std_living >= 3 ∧ Husbands_edu <= 3	0.9999999999999842	0.3371	309
Wifes_edu <= 2 ∧ Std_living >= 3	0.9999999999999442	0.3330	292
Husbands_edu >= 1.0 ∧ Std_living >= 3 ∧ Husbands_edu <= 3	0.9999999999999382	0.3485	335

Table 6.6: Contraception: top-3 subgroups for Kendall

6.3.3 Iris

The famous Iris flower dataset (introduced by Fisher [32]), 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 leaves of a flowering plant.

Experiments with the Iris dataset show that descriptions are found, which separate the data with respect to their class. Pearson and Spearman both find rules that exclude samples that are classified as setosa, while the Kendall class mirrors this behavior by returning subgroups consisting only of examples classified as setosa. The following tables and figures illustrate the experiment with target attributes *petallength* and *sepallength*.

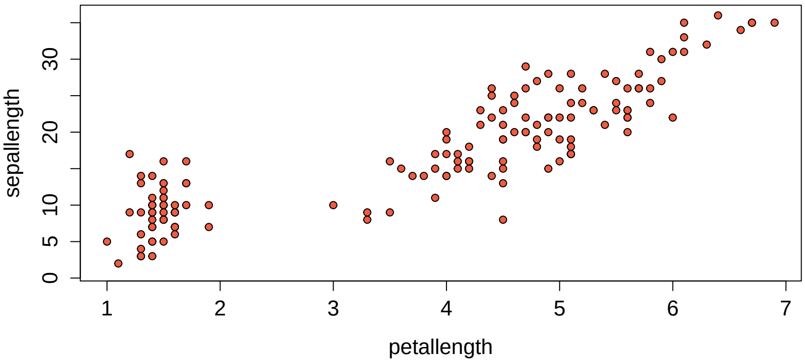


Figure 6.1: Iris dataset

Subgroup	φ	r	n
petalwidth >= 0.5 $\wedge$ sepalwidth >= 2.2	0.999999988	0.8183	101
sepalwidth <= 4.1 $\wedge$ petalwidth >= 0.3	0.999999557	0.8305	115
sepalwidth >= 2.5 $\wedge$ petalwidth <= 0.3	0.999995618	0.2382	40

Table 6.7: Iris: top-3 subgroups for Pearson

Subgroup	φ	ρ	n
sepalwidth <= 4.1 $\wedge$ petalwidth >= 0.3	0.999999655	0.8444	115
petalwidth <= 0.3	0.9999931	0.2736	41
petalwidth >= 2.1 $\wedge$ sepalwidth <= 2.8 coverage 4	1	1	4

Table 6.8: Iris: top-3 subgroups for Spearman

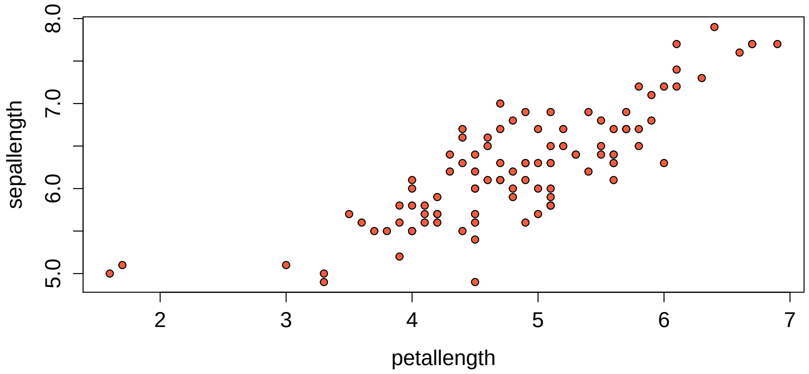


Figure 6.2: petalwidth >= 0.5 ∧ sepalwidth >= 2.2

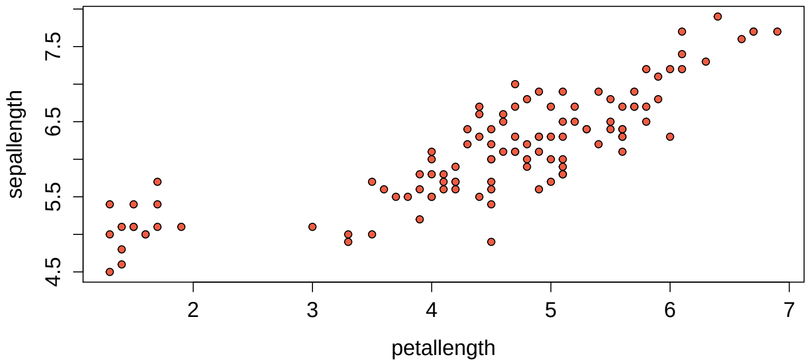


Figure 6.3: sepalwidth <= 4.1 ∧ petalwidth >= 0.3

Subgroup	φ	τ_b	n
petalwidth <= 0.5 ∧ petalwidth >= 0.2 ∧ sepalwidth >= 2.9	0.9999999999999996	0.1515	42
sepalwidth >= 3.4 ∧ petalwidth <= 2.4 ∧ petalwidth >= 0.2	0.999999999999993	0.2648	34
petalwidth <= 1.7 ∧ sepalwidth >= 3.7 ∧ petalwidth >= 0.1	0.9999999999999911	-0.3234	13

Table 6.9: Iris: top-3 subgroups for Kendall

In this experiment we could observe the same behavior as in the Contraceptive Method Choice dataset, as here again Pearson and Spearman's measures report more or less similar subgroups and relationships, while Kendall's measure returns subgroups that do not observe any relationship as opposed to their complements.

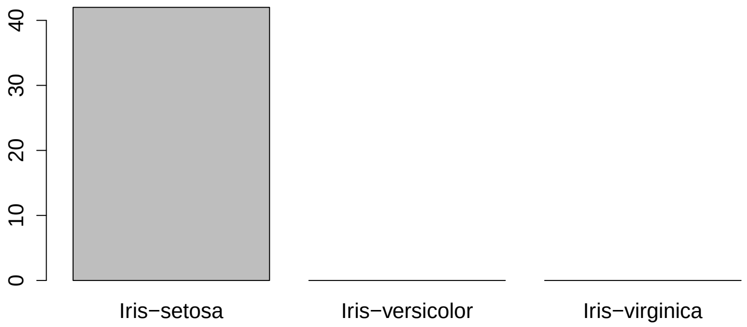


Figure 6.5: Iris: class distribution in the best subgroup (Kendall)

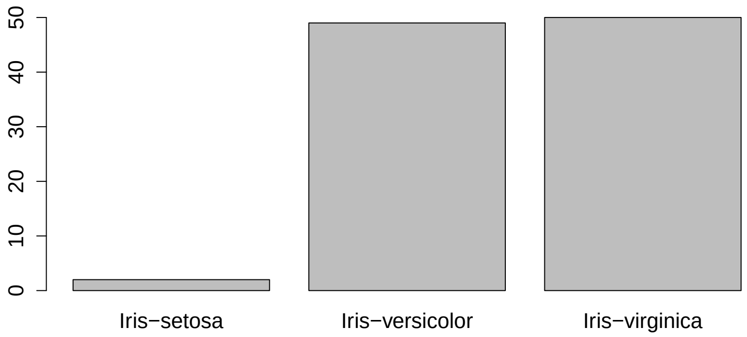


Figure 6.4: Iris: class distribution in the best subgroup (Pearson)

Looking at the distribution of the classes (Figure 6.4 and 6.5), our results agree with previous experiments on this dataset that found the setosa class to be linearly separable from virginica and versicolor. Fisher noted e.g., “It will be noticed [...] that there is some overlap of the distributions of *I. virginica* and *I. versicolor*, so that a certain diagnosis of these two species could not be based solely on these four measurements [...]”[32].

6.3.4 Higgs Boson Challenge

The Higgs boson is an elementary particle, that 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 provided simulated data used by physicists as a challenge to optimize the analysis of the Higgs boson. It contains a set of 250000 simulated proton collision (so-called events), which are characterized by a set of measured quantities, such as the energy momentum of the particle or spacial coordinates of the resulting quarks. All quantities and their respective meanings can be found in the documentation [33]. The goal of the challenge is to improve classification of events. However, classification is not our primary goal and we will more generally explore whether we can find subgroups in the data, that appear interesting (based on their scatterplots).

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 2000. Otherwise the found subgroups were too big and a sensible interpretation of their respective scatterplots was not possible.

Testing the Cern datasets, we again get results indicating that the Spearman and Pearson Correlation Classes will produce similar results. As an example we present the found subgroups for gauging the relationship between the attributes "PRI_tau_pt" and "PRI_tau_eta". The choice here is arbitrary as the drawn conclusion fit any of the experiments we did on different attributes. In Table 6.10 and Table 6.11 we present the top-3 subgroups found by the Pearson and the Spearman Correlation Model Class, respectively.

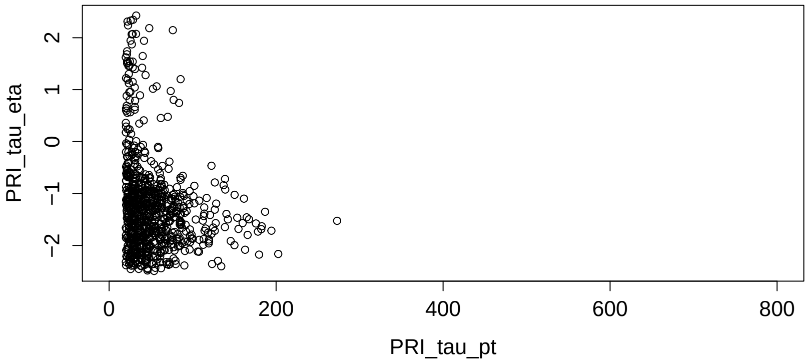


Figure 6.6: Subgroup corresponding to rule: PRI_lep_eta <= -1.99 ∧ PRI_jet_leading_pt >= 134.551

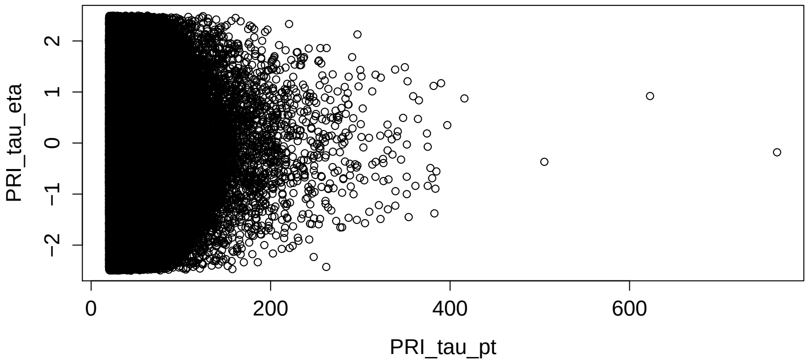


Figure 6.7: Cern: complement of the subgroup presented in figure 6.6

In Figure 6.6 we can see a concentration in the lower left corner as opposed to the structure of the complement in Figure 6.7

Subgroup	φ	τ_b	n
PRI_lep_eta >= 2.0 $\wedge$ PRI_jet_leading_phi >= 2.497	0.9999999992698	0.2163	817
PRI_lep_eta <= -1.99 $\wedge$ PRI_jet_leading_pt >= 134.551	0.9999999962810	-0.2143	784
PRI_lep_eta <= -1.99 $\wedge$ PRI_jet_all_pt >= 215.471	0.9999999883011	-0.2065	795

Table 6.10: Cern: top-3 subgroups for Pearson

Subgroup	φ	τ_b	n
PRI_lep_eta <= -1.99 $\wedge$ PRI_jet_all_pt >= 215.471	0.999999991891	-0.2027	795
PRI_lep_eta <= -1.99 $\wedge$ PRI_jet_leading_pt >= 134.551	0.999999991600	-0.2036	784
PRI_jet_leading_pt <= 117.866 $\wedge$ PRI_lep_eta >= 1.999 $\wedge$ PRI_jet_leading_phi >= 2.499	0.999999902459	0.1952	712

Table 6.11: Cern: top-3 subgroups for Spearman

7 Conclusions

In this thesis we explored the possibilities of replacing the Person Correlation Model Class with a Rank Correlation based Model Class to see if this would “improve” the performance. We therefore chose to examine two of the most popular rank correlation coefficients.

In the experiments the Spearman Correlation measure performed similar to the Pearson Correlation measure. We can therefore at least recommend it as an alternative as it seems to be as strong as Pearson’s correlation in detecting subgroups with linear relationship. Whether it can find subgroups with relationships beyond linear ones is still unknown. In our experiments we could not observe any behavior to indicate this, but at least the mathematical background suggests that it should also capture them.

Not an alternative but a different quality measure can be created with Kendall’s correlation. In the experiments the results for this class somewhat complemented the results of the two aforementioned measures; Spearman and Pearson return subgroups that have some sort of linear relationship as opposed to their complements. Kendall’s measure on the other hand returns subgroups having little or no correlation, while their complements do.

It might therefore be useful to apply Kendall’s correlation when the subgroups found with Pearson or Spearman do not show the desired result. In general Kendall seems to find subgroups that show no real relationship compared to the complement, therefore more or less filtering those groups, that do not observe any quantifiable relationship.

Another use could simply be to get a different view on the dataset in question, as suggested at the end of Section 4.2.6, since Kendall essentially measures different quantities than Pearson/Spearman.

7.1 Outlook

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 Chapter 3. However, for a good quality measure it would also be necessary to investigate ways to compare these statistics on different subsets of datasets. Regardless of other measures, the (indirect) assumption of an underlying normal distribution could be removed by applying the bootstrapping method used for permutation tests. With this approach the underlying distribution of the applied correlation coefficient is estimated through re-sampling from the dataset a fixed number of times. This estimation then can be used for a statistical test of difference in the applied correlation measure. However, as this approach is computationally expensive, it will only provide an alternative for smaller datasets and is unsatisfactory as a general approach for correlation-based Exceptional Model Mining.

As mentioned in Section 3.1.2, another direction could be to experiment with an exhaustive algorithm. For this approach, a valuation basis for the GP-Growth algorithm, proposed by Lemmrich et al. [5], could be developed. It would be interesting to see if an exhaustive approach for these models yields different results.

A slight modification of the implemented version could be derived from a paper written by Wilcox [34], who examines a new method of computing confidence intervals for the difference of two Pearson correlation coefficients proposed by Zou [35]. The general idea is to use Fisher's z-transformation to compute confidence intervals for the individual correlations. From those, a confidence interval for $\rho_1 - \rho_2$ can be computed. The described method might very well be adapted for Spearman's and Kendall's correlation, as both can be Fisher z-transformed. Further investigation here might result in more statistically backed-up results.

Bibliography

[1] D. Leman, A. Feelders, and A. Knobbe, "Exceptional model mining," *ECML PKDD'08*, vol. 5212, pp. 1–16, 2008.

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

[3] R. Agrawal, T. Imieliński, and A. Swami, "Mining Association Rules between Sets of Items in large Databases," *ACM SIGMOD Rec.*, vol. 22, no. May, pp. 207–216, 1993.

[4] F. Herrera, C. J. Carmona, P. González, and M. J. Jesus, "An overview on subgroup discovery: foundations and applications," *Knowl. Inf. Syst.*, vol. 29, pp. 495–525, Nov. 2010.

[5] F. Lemmerich, M. Becker, and M. Atzmueller, "Generic pattern trees for exhaustive exceptional model mining," in *Proceedings of the 2012 European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part II*, ECML PKDD'12, pp. 277–292, Springer-Verlag, 2012.

[6] M. Mampaey, S. Nijssen, A. Feelders, and A. Knobbe, "Efficient algorithms for finding richer subgroup descriptions in numeric and nominal data," in *IEEE Int. Conf. Data Min.*, pp. 499—-508, 2012.

[7] W. Klösgen, "Explora: a multipattern and multistrategy discovery assistant," *Adv. Knowl. Discov. data Min.*, pp. 249–271, 1996.

[8] S. Wrobel, "An algorithm for multi-relational discovery of subgroups," *Lect. Notes Comput. Sci.*, vol. 1263/1997, pp. 78–87, 1997.

[9] M. van Leeuwen, "Maximal exceptions with minimal descriptions," *Data Min. Knowl. Discov.*, vol. 21, pp. 259–276, July 2010.

[10] M. Clark, "A Comparison Of Correlation Measures," tech. rep., University of Notre Dame, 2013.

[11] W. Hoeffding, "A non-parametric test of independence," *Ann. Math. Stat.*, 1948.

[12] J. R. Blum, J. Kiefer, and M. Rosenblatt, "Distribution Free Tests of Independence based on the Sample Distribution Function," *Ann. Math. Stat.*, 1961.

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

[14] G. J. Székely, M. L. Rizzo, and N. K. Bakirov, "Measuring and testing dependence by correlation of distances," *Ann. Stat.*, vol. 35, pp. 2769–2794, Dec. 2007.

[15] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, "Detecting novel associations in large data sets.," *Science*, vol. 334, pp. 1518–24, Dec. 2011.

[16] J. B. Kinney and G. S. Atwal, "Equitability, mutual information, and the maximal information coefficient.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 111, pp. 3354–9, Mar. 2013.

[17] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf, "Measuring statistical dependence with Hilbert-Schmidt norms," *Lect. Notes Comput. Sci.*, vol. 3734, pp. 63–77, 2005.

[18] D. Lopez-Paz, P. Hennig, and B. Schölkopf, "The randomized dependence coefficient," *NIPS*, pp. 1–9, 2013.

[19] H. Gebelein, "Das statistische problem der korrelation als variations- und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung.," *Zeitschrift für Angewandte Mathematik und Mechanik*, vol. 21, pp. 364—379, 1941.

[20] C. J. Kowalski, "On the Effects of Non-Normality on the Distribution of the Sample Product-Moment Correlation Coefficient," *Appl. Stat.*, vol. 21, no. 1, pp. 1–12, 1972.

[21] F. J. Anscombe, "Graphs in statistical analysis," *Am. Stat.*, vol. 27, no. 1, pp. 17–21, 1973.

[22] C. Spearman, "The proof and measurement of association between two things," *Am. J. Psychol.*, vol. 15, no. 1, pp. 72–101, 1904.

[23] M. G. Kendall, "A new measure of rank correlation.," *Biometrika*, vol. 30, no. 1, pp. 81–93, 1938.

[24] L. A. Goodman and W. H. Kruskal, "Measures of association for cross classifications," *J. Am. Stat. Assoc.*, vol. 49, no. 268, pp. 732–764, 1954.

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

[26] R. H. Somers, "A new asymmetric measure of association for ordinal variables," *American sociological review*, pp. 799–811, 1962.

[27] R. A. S. Fisher, *Statistical methods for research workers.* Edinburgh : Oliver and Boyd, 14th ed., revised and enlarged ed., 1970. Previous ed. 1958.

[28] E. C. Fieller, H. O. Hartley, and E. S. Pearson, "Tests for rank correlation coefficients. I," *Biometrika*, vol. 44, no. 4, pp. 470–481, 1957.

[29] W. R. Knight, "A computer method for calculating Kendall's tau with ungrouped data," *J. Am. Stat. Assoc.*, vol. 61, no. 314, pp. 436–439, 1966.

[30] D. Christensen, "Fast algorithms for the calculation of Kendall's Tau," *Comput. Stat.*, pp. 51–62, 2005.

[31] K. Bache and M. Lichman, “UCI machine learning repository,” 2013.

[32] R. A. Fisher, “The Use of Multiple Measurements in Taxonomic Problems,” *Ann. Eugen.*, vol. 7, no. 2, pp. 179—-188, 1936.

[33] C. Adam-Bourdarios, G. Cowan, I. G. Cécile Germain, B. Kégl, and D. Rousseau, “Learning to discover: the higgs boson machine learning challenge,” 2014.

[34] R. R. Wilcox, “Comparing Pearson Correlations: Dealing with Heteroscedasticity and Non-Normality,” *Commun. Stat. - Simul. Comput.*, vol. 38, pp. 2220–2234, 2009.

[35] G. Y. Zou, “Toward using confidence intervals to compare correlations,” *Psychol. Methods*, vol. 12, pp. 399–413, Dec. 2007.

Erklärung

Hiermit erkläre ich, Lennart Downar, die vorliegende Bachelor-Arbeit mit dem Titel *A Rank Correlation Model Class for Exceptional Model Mining* selbständig verfasst und keine anderen als die hier angegebenen Hilfsmittel verwendet, sowie Zitate kenntlich gemacht zu haben.

Dortmund, September 22, 2014