Subgroup Discovery in Ranked Data, with an Application to Gene Set Enrichment

Barbara F.I. Pieters[1], Arno Knobbe[2], and Sašo Džeroski[3]

[1] Utrecht University, the Netherlands
[2] LIACS, Leiden University, the Netherlands
[3] Jožef Stefan Institute, Ljubljana, Slovenia
bpieters@cs.uu.nl

Abstract. We investigate a class of problems that deal with ranked data. Such data can be found in a variety of domains, ranging from inherently competitive fields such as sports and business, to more surprising applications such as relevance ranking and temporal data (where more recent events rank higher). In this paper, we deal with ranked data in a Subgroup Discovery setting, where we are looking to find subgroups that occur predominantly at the top of the ranking. The analysis of ranked data can be posed as either a regression or an ordinal discovery problem, and we introduce and review a number of quality measures for both settings. As an additional challenge, the possibilities of ties in the ranking (a so-called *partial* ranking) is accounted for. The techniques for ranked Subgroup Discovery are tested on an important application in cancer research, where a given ranking of 12k genes (based on their individual involvement in a tumour called neuroblastoma) is analysed for common functional themes among the top-ranking genes.

1 Introduction

This paper is concerned with the analysis of objects that are ranked. A ranking of objects provides information about the relationship between objects: high-ranking objects are somehow 'bigger', 'better' or 'more recent' (depending on the application domain) than all objects with a lower rank. We will be investigating the question of what makes some (groups of) objects appear at the top of the ranking, and others not. We will be approaching this question from a Subgroup Discovery (SD) [1, 5, 7, 8, 10] point of view, that is, we aim to discover subgroups that are over-represented at the top of the ranking, and are uncommon in lower regions.

One should note that 'top of the ranking' is a vague notion, and it is often undesirable to provide a clear cut-off between top-ranking objects and the remainder of the database. Therefore, approaching the analysis of ranked data as a binary classification problem (which is the usual SD setting) is not an option. A more logical setting would be to treat the rank of each object as a number, and thus opt for a *regression* setting. In this setting, good subgroups, includ-

ing mostly high ranking objects, will have a significantly lower average number[4] than can be expected from the database average. In fact, some of the solutions provided in Section 4 will be assuming such a regression setting. However, treating the rank as actual numbers implies a metric that may be entirely artificial, thus potentially producing artifacts in the results. For example, the difference between rank 1. and 2. need not be the same as that between 5. and 6. The obvious solution to this artificial metric is to treat the analysis as an *ordinal* problem [3], and we will investigate a number of ways to do this in the SD paradigm.

Subgroup Discovery for ordinal and regression problems is mostly a matter of providing *quality measures* that treat the target (typically the rank) in the appropriate manner, by evaluating subgroups that are over-represented at the top favourably. Section 4 introduces a range of quality measures, both for the ordinal and regression setting. Each measure has its own specific qualities that make it desirable or less desirable, depending on the specific needs of the application domain. Essentially, each quality measure answers a specific type of analysis question, and as such, one cannot really compare measures, and answer conclusively the question of which measure is the preferred choice. However, we do investigate different properties, and shed light on the specific advantages of each measure.

The measures we survey were taken from various backgrounds. We propose one new measure, called Median MAD Metric, and furthermore consider two measures from the numeric Subgroup Discovery field [5, 7, 14, 15], as well as measures that are inspired by well-known statistical tests, but as yet not employed to analyse ranked data. The collection of measures is selected such that the range of analysis questions imaginable for ranked data is covered by the available measures.

The analysis of ranked data is very widely applicable. Obvious examples come of course from domains that are competitive by nature, such as sports (e.g., FIFA World Ranking (soccer), FIDE Federations Ranking (chess)) and business (Fortune 500). The following example demonstrates how this might work:

Example 1. Table 1 shows the final ranking of countries participating in the 2010 Winter Olympics in Vancouver, based on the total number of medals obtained (irrespective of the type of medal). For reasons of presentation, we only consider the 24 countries that won at least one medal. The table includes information about the participation of countries, as well as general demographics of the country and its inhabitants, such as its population size, the most common language (family), and whether the country is a republic. Note that some countries share a particular rank, as they obtained equal numbers of medals.

One of the simplest subgroups that will be considered by a typical SD algorithm is the set of countries that have polar regions (*polar* = y). As the data

[4] We will use low numbers to denote high ranks, such that 1. denotes the best ranking object.

Table 1. Ranking of countries participating in the 2010 Olympic Winter Games.

Rank	Country	Medals	Athletes	Continent	Popul.	Lang. Family	Republic	Polar
1	United States	37	214	N. America	309	Germanic	y	y
2	Germany	30	152	Europe	82	Germanic	y	n
3	Canada	26	205	N. America	34	Germanic	n	y
4	Norway	23	100	Europe	4.8	Germanic	n	y
5	Austria	16	79	Europe	8.3	Germanic	y	n
6	Russian Fed.	15	179	Asia	142	Slavic	y	y
7	Korea	14	46	Asia	73	Altaic	y	n
9	China	11	90	Asia	1338	Sino-Tibetan	y	n
9	Sweden	11	107	Europe	9.3	Germanic	n	y
9	France	11	107	Europe	65	Italic	y	n
11	Switzerland	9	144	Europe	7.8	Germanic	y	n
12	Netherlands	8	34	Europe	16.5	Germanic	n	n
13.5	Czech Rep.	6	92	Europe	10.5	Slavic	y	n
13.5	Poland	6	50	Europe	38	Slavic	y	n
16	Italy	5	110	Europe	60	Italic	y	n
16	Japan	5	94	Asia	127	Japonic	n	n
16	Finland	5	95	Europe	5.3	Finno-Ugric	y	y
20	Australia	3	40	Australia	22	Germanic	y	n
20	Belarus	3	49	Europe	9.6	Slavic	y	n
20	Slovakia	3	73	Europe	5.4	Slavic	y	n
20	Croatia	3	18	Europe	4.5	Slavic	y	n
20	Slovenia	3	49	Europe	2	Slavic	y	n
23	Latvia	2	58	Europe	2.2	Slavic	y	n
25	Great Britain	1	52	Europe	61	Germanic	n	n
25	Estonia	1	30	Europe	1.3	Finno-Ugric	y	n
25	Kazakhstan	1	38	Asia	16	Turkic	y	n

shows, this is a reasonably interesting subgroup, as three of the top four countries are included (1. *United States*, 3. *Canada*, 4. *Norway*), and furthermore, only a single polar country appears in the lower half of the ranking (16. *Finland*). A somewhat more complicated subgroup that scores particularly well is the following:

$$\mathit{language\_family} = \mathit{Germanic} \wedge \mathit{athletes} \geq 60$$

These conditions select countries of which the majority of inhabitants speak a Germanic language (e.g., English, German, the Scandinavian languages, ...), and that have sent at least 60 competing athletes (this further excludes the Netherlands, Great Britain and Australia). The subgroup comprises 7 countries (1. *United States*, 2. *Germany*, 3. *Canada*, 4. *Norway*, 5. *Austria*, 9. *Sweden*, 11. *Switzerland*), the majority of which appear in the top 10 of the ranking.

Apart from applications in sports and business, the proposed techniques can be applied to temporal data of discrete nature. One could argue that a sequence of events is ranked according to time, with the most recent event having the

highest rank. Discovering current trends, fashions or developments would thus translate to finding subgroups that encompass mostly recent events. This temporal setting can be of benefit in applications such as finding descriptions for recent trends in movie releases, press releases or financial transactions, to name but a few.

Finally, the described methods can play an important role in making sense of objects somehow ranked by relevance, such as webpages returned by Google, and ranked by PageRank. One instance of such relevance ranking, that was our original motivation for considering extensions to the standard SD setting, can be found in the Bioinformatics domain. Here, the initial statistical analysis of high-throughput data such as from micro-arrays typically results in a list of genes ranked by relevance to the particular disease or phenotype under consideration. In a European project that all authors were involved in – the European Embryonal Tumour Pipeline (*EETP*) project [9] – a number of gene rankings were produced for a variety of tumour types involving young children. As the number of genes can be large, in this case over 12 thousand, and domain experts generally only are familiar with the most well-known ones, the task of making sense of such a large ranking of genes can be daunting. Subgroup Discovery on ranked data was used in order to shed light on the most relevant genes in the list, and find descriptions of the top-ranking genes in terms of a number of background sources of gene-related domain knowledge. These include relational information about protein-families, functional annotations, gene locations, and gene-to-gene interaction networks. In Section 6, we evaluate our methods on data relating to one particular tumour of the nervous system, known as *neuroblastoma* [2], for which data was collected during the EETP project.

2 Preliminaries

Subgroup discovery is performed on a dataset D, with individuals (records) $\boldsymbol{x} \in D$ of the form $\boldsymbol{x} = \{a_1, \ldots, a_m, t\}$, where m is a positive integer. The set of attributes $\{a_1, \ldots, a_m\}$ is denoted by $\boldsymbol{a}$, taken from domain $\mathcal{A}$. The target attribute t is assumed to be continuous and taken from domain $\mathbb{R}$, the ordinal target attribute is a special case of a continuous target attribute. The size of the dataset is denoted by $N = |D|$. The list of target values of all individuals in the dataset is denoted by $\boldsymbol{T} = \{T_1, \ldots, T_N\}$.

Rankings come in two types, *complete* and *partial*. In the complete case, all individuals have a unique value for their target attribute t. In a partial ranking however, at least two individuals will have the same value for t, such that their mutual order cannot be determined. The Olympic ranking of countries is an example of such a partial ranking. In the case of an ordinal approach to ranking, we will be assuming that individuals are labeled by integers ($\boldsymbol{T} = \{1, \ldots, N\}$), with 1 denoting the highest ranking individual. For ordinal rankings that are partial, we will be assuming the so-called *fractional ranking*: all equal individuals will be assigned the mean of the rank of these individuals, were ties arbitrarily broken. This is equivalent to $1 + n_+ + (n_= - 1)/2$, where n_+ is the number

of higher-ranked individuals, and $n_=$ the number of ties. For example China, Sweden and France share rank $1 + 7 + (3 - 1)/2 = 9$.

A subgroup of the dataset is a set of individuals $s \subseteq D$ that are covered by a certain *pattern* p (condition) in pattern language $\mathcal{P}$. A pattern is a function $p : \mathcal{A} \to \{0, 1\}$, where pattern p covers individual $\boldsymbol{x}^i \in D$ if and only if $p(\boldsymbol{a}^i) = 1$. The size of subgroup s is denoted by n. The list of target values of the individuals in the subgroup is denoted by $\boldsymbol{t} = \{t_1, \ldots, t_n\}$. Hence, a pattern corresponds to one subgroup, and a subgroup corresponds to a subset of individuals in the dataset. The complement of subgroup $\bar{s}$ is the subset of individuals $\bar{s} = D \setminus s$, i.e. all individuals not in subgroup s. The list of target values of all individuals in $\bar{s}$ is denoted by $\bar{\boldsymbol{t}} = \boldsymbol{T} \setminus \boldsymbol{t}$. The size of the complement of the subgroup is denoted by $\bar{n} = N - n$.

The quality of a subgroup is defined by a *quality* measure $\varphi(p) : \mathcal{P} \to \mathbb{R}$ that assigns a numeric value to a pattern p given a dataset D. The objective of a quality measure is to return patterns that are evaluated as best. What is considered as best, i.e. either a large value or a small value (even negative), depends on the characteristics of the quality measure.

It may be the case that two different qualities evaluate subgroups differently, but still produce lists of interesting subgroups in the same order. In this case, the two quality measures are called *order-equivalent*:

Definition 1 (order equivalence). *Two quality measure $\varphi_1(s)$ and $\varphi_2(s)$ are called order-equivalent $\varphi_1(s) \sim \varphi_2(s)$, iff $\varphi_1(s_1) > \varphi_1(s_2) \to \varphi_2(s_1) > \varphi_2(s_2) \wedge \varphi_1(s_1) = \varphi_1(s_2) \to \varphi_2(s_1) = \varphi_2(s_2)$ $\forall s_1, s_2 \in s$.*

3 Intuitions on Subgroups

In our view, quality measures are a formal expression of the kinds of subgroups a data analyst is interested in. In many cases, the analyst will have some informal ideas about the nature of the ideal subgroup, which should then be translated into a specific choice of quality measure. In this section we will review a number of such informal qualities of subgroups, which we will refer to as *intuitions*. The quality measures, and how they match the intuitions defined here, will then be discussed in the next two sections.

Somewhat informally, the following intuitions are typical for subgroups in ranked data:

I1 size: Larger subgroups (containing many individuals) will be preferred over smaller ones.
I2 rank: The majority of the subgroup individuals should be highly ranked.
I3 position: The 'middle' of the subgroup should differ from the middle of the ranking.
I4 deviation: The subgroup individuals should have similar ranks.

The first intuition is rather straightforward, and common to many mining paradigms. Larger subgroups represent more common and more reliable themes

in the data. One could argue that really large subgroups are less interesting because they are too general. This notion is however covered by the Intuitions 2 and 3, as the distribution of a really large subgroup starts to resemble that of the entire database.

Intuition 2 is the obvious intuition one has when dealing with rankings that high ranking individuals are important. It is a special case of Intuition 3, which specifies that subgroups should not be spread around the 'middle' (mean, median, ...) of the ranking. Intuition 3 allows for targets that are not a fractional ranking per se, for example when one has a numeric target where high values indicate high ranks.

Intuition 4 requires the individuals to have similar ranks, that is, the individuals should ideally form a block. If this is not the case, the distribution of the dataset and the subgroup can become alike, due to the more even spread of the subgroup individuals in the dataset.

4 Quality Measures

Currently, most quality measures in Subgroup Discovery are only applicable to nominal target attributes. Moreover, most quality measures require targets to be binary. To the best of our knowledge, only a few studies have been conducted on Subgroup Discovery involving continuous target variables. No studies have been conducted on Subgroup Discovery with ordinal targets. In [7], Klösgen presents a measure (*mean test*) to deal with continuous target attributes. This measure was later adopted by Grosskreutz [5]. Trajkovski [14, 15] has implemented a few quality measures for continuous target attributes, one of which is the *z-score*. Both the mean test and the z-score were also implemented by us and will be presented later. They can both deal with ordinal target attributes too, although they are not specifically designed for such targets.

In the next two sections, we provide a list of quality measures for ranked data. The first section is concerned with measures for continuous targets, whereas the second deals with measure for ordinal targets. With respect to their prior use for ranked data, the 8 measures are either, *a*) well-established in ranked Subgroup Discovery (Mean test, Standardized z-score), *b*) well-known statistical measures with a novel application to ranked SD (Average, t-statistic, Median χ^2 statistic, Area under the ROC curve, Wilcoxon-Mann-Whitney Ranks statistic), or *c*) entirely novel (Median MAD metric).

4.1 Quality Measures for Continuous Target Attributes

Average A relatively simple and effective quality measure is the average target value (mean, μ) of a subgroup. Depending on the subgroup search objectives, a maximum of all averages or minimum of all averages is best. For instance, if the number of medals obtained (see example 1) is the target, high averages are desired (*maximization*). On the other hand, if the rank of the country is the target, low averages are prefered (*minimization*). Given the list of subgroup target attribute values $\boldsymbol{t}$ with size n, the average is $\varphi_{avg}(s) = \frac{\sum_{i=1}^{n} t_i}{n}$.

Mean test A more complex measure is the mean test. It compares the distribution of the subgroup to the distribution of the dataset: $\varphi_{mt}(s) = \sqrt{n}(\mu - \mu_0)$. μ is the mean of $\boldsymbol{t}$ (the subgroup) and μ_0 is the mean of $\boldsymbol{T}$ (the entire dataset). The evaluation values of the mean test can assume values from the interval $(-\infty, \infty)$, where a negative value indicates that the subgroup mainly is positioned below the dataset mean, and vice versa.

(Standardized) z-score The z-score [3] for a group of individuals can be calculated using the standardized version of the z-score, and has been used for instance by Trajkovski et al. [14, 15]. The standardized z-score of a subgroup is defined by $\varphi_z(s) = \frac{\mu - \mu_0}{(\sigma_0/\sqrt{n})} = \frac{\sqrt{n}(\mu - \mu_0)}{\sigma_0}$. μ and μ_0 are defined as before, σ_0 is the standard deviation of $\boldsymbol{T}$. The standardized z-score measures how far the mean of the subgroup is away from the mean of D in terms of standard deviations. As $\varphi_z(s) = \frac{\varphi_{mt}(s)}{\sigma_0}$, and σ_0 is constant for a given dataset, the following holds:

Proposition 1 ($\varphi_z(s)$ and $\varphi_{mt}(s)$ order equivalence). $\varphi_{mt}(s) \sim \varphi_z(s)$

The evaluation values of the z-score range from $-\infty$ to ∞, with positive values indicating a subgroup mean above the dataset mean.

t-Statistic Compared to the z-score, the t-statistic [3] is more accurate for smaller sample sizes. It is thus more suited when subgroup sizes can or should be small. The t-statistic is related to the z-score in the sense that they both compare the distribution of the subgroup to the distribution of the dataset, but the t-statistic uses the subgroup deviation instead of the dataset deviation. This makes the t-statistic sensitive to differences in variances in subgroups. The range of the evaluation values of the t-statistic is $(-\infty, \infty)$. The sign of the evaluation value gives information of the position of the majority of the subgroup, similar to $\varphi_z(s)$. The t-statistic for a subgroup is calculated as follows: $\varphi_t(s) = \frac{\mu - \mu_0}{(\sigma/\sqrt{n})} = \frac{\sqrt{n}(\mu - \mu_0)}{\sigma}$, where n, μ and μ_0 are defined as before. σ is the standard deviation of $\boldsymbol{t}$.

Median χ^2 Statistic The median χ^2 statistic [3] uses the median of the dataset to calculate the difference in distributions. The median is either $\frac{T_{N/2} + T_{N/2+1}}{2}$ if N is even, or $T_{\frac{N+1}{2}}$ if N is odd. The distribution difference between the dataset and the subgroup is calculated by counting how many individuals in both the subgroup and the dataset lie above the dataset median and at or below the dataset median: $\varphi_{\chi^2}(s) = \frac{(n_l - N_l)^2}{N_l} + \frac{(n_s - N_s)^2}{N_s}$. n_l and N_l stand for the frequencies of individuals in the subgroup and the dataset respectively whose target values are larger than the dataset median. n_s and N_s consequently denote the frequencies of individuals in the subgroup and the dataset respectively whose target values are equal to or smaller than the dataset median. The evaluation values range from 0 to ∞, where 0 denotes that the subgroup individuals are equally divided around the dataset median and that the subgroup is large.

4.2 Quality Measures for Ordinal Subgroup Discovery

From a statistical point of view, ordinal data, such as ranked data, is data for which it is not known from what kind of distribution the data originates. More specifically, it is assumed that such data does not even follow a distribution. Therefore, nonparametric tests are used to make inferences on ordinal data [3]. The quality measures implemented for Subgroup Discovery on ordinal data are either based on nonparametric tests or inspired by them. These measures are less sensitive and more robust than their parametric counterparts, such as the t-statistic and z-score.

AUC of ROC The area under the Receiver Operating Characteristic (ROC) curve [6] is traditionally a metric to compare the performance of classifiers. The AUC of ROC is modified in such a way that it measures how interspersed the individuals of a subgroup are in the overall dataset. In other words, this measure is very useful to define the position of the subgroup individuals in the dataset and whether they are grouped together or more spread out. To do so, the AUC of ROC divides the individuals of the dataset into s and $\bar{s}$. The AUC of ROC is calculated as follows: $\varphi_{roc}(s) = \frac{\sum_{i=1}^{\bar{n}} \bar{t}_i - \frac{\bar{n}(\bar{n}+1)}{2}}{\bar{n}n}$

The $\varphi_{roc}(s)$ measure can only be used on complete rankings. The range of the $\varphi_{roc}(s)$ is $[0, 1]$. If $\varphi_{roc}(s) = 1$, the n individuals of the subgroup correspond to the first n individuals in the ranking. If the measure returns 0 however, the subgroup represent only the n lowest ranks. All other values indicate that the individuals are either not closely packed and/or the best (or worst) dataset individual is not present in the subgroup. Note that the size of the subgroup does not affect the value of $\varphi_{roc}(s)$.

Wilcoxon-Mann-Whitney Ranks statistic The Wilcoxon-Mann-Whitney Ranks (wmw) statistic [3] is derived from the nonparametric Wilcoxon-Mann-Whitney Ranks test. It is related to the z-score, since it calculates the difference of the means of the ranks through the z-statistic. The distribution of the subgroup is compared to the distribution of the complement of the subgroup. The wmw ranks statistic is defined as follows: $\varphi_{wmw}(s) = \frac{\sum_{i=1}^{n} t_i - \mu_0}{\sigma_0}$. Here, the mean and standard deviation of the dataset are defined on the ranks and differ from the usual definitions: $\mu_0 = \frac{n(N+1)}{2}$ and $\sigma_0 = \sqrt{\frac{n\bar{n}(N+1)}{12}}$. The range of $\varphi_{wmw}(s)$ is $(-\infty, \infty)$. The interpretation of the evaluation values is equal to the interpretation of the $\varphi_z(s)$ evaluation values.

Median MAD Metric Apart from the statistics described above, a new metric was developed, the median MAD metric ($mmad$). The mmad maximizes on the subgroup size and minimizes on the median and median absolute deviation (mad). The median and the mad are the robust, nonparametric counterparts of the mean and the standard deviation. The median and the mad are not sensitive to outliers, as long as there are only few. Thus, large subgroups with a relatively small portion of bad individuals are not penalized too heavily. This does happen when the mean or standard deviation are used.

The median mad metric *mmad* shows a bias towards large subgroups where the majority of the individuals have top ranks. The metric does not compare the subgroup distribution to the dataset distribution, but just calculates a ratio for the subgroup size and the position of the individuals (cluster) in the subgroup. The mmad is defined as follows: $\varphi_{mmad}(s) = \frac{n}{2 \cdot m + mad}$, where the mad is denoted by $mad(\boldsymbol{t}) = median(\boldsymbol{y})$. Here, $\boldsymbol{y} = \{|t_1 - m|, \ldots, |t_k - m|\}$, and m is the median of $\boldsymbol{t}$. The evaluation values range from 0 (worst) to ∞ (best).

5 On Quality Intuitions and Quality Measures

In this section, we investigate different characteristics of the quality measures presented in the previous section. We shed some light on how, and on which type of data each of the measures should be applied. Furthermore, in order to better understand the types of subgroups each measure favours, we loosely test to what extent each quality measure follows each intuition.

In table 2 below, we list some basic characteristics of the quality measures. The table shows what kind of targets the measures can deal with (ordinal, numeric, or both). Also, whether a measure can deal with partial or complete rankings is shown here (only φ_{roc} is limited to complete rankings). *Symmetry* indicates how values returned by the measures should be interpreted. A positive value (> 0) of the measure denotes that the subgroup distribution of the target is predominantly above that of the dataset distribution. A negative value indicates that the subgroup distribution lies below the dataset distribution. The fourth characteristic indicates what kind of information is needed, that is, whether the distributions on s, D and/or $\bar{s}$ are required for subgroup evaluation.

Table 2. Quality measures and their characteristics.

	φ_{avg}	φ_{mt}	φ_{z}	φ_{t}	φ_{χ^2}	φ_{roc}	φ_{wmw}	φ_{mmad}
Continuous/Ordinal targets	both	both	both	both	both	ordinal	ordinal	ordinal
Complete/Partial ranking	both	both	both	both	both	complete	both	both
Symmetric	no	yes	yes	yes	no	no	yes	no
Distribution information	s	$s\&D$	$s\&D$	$s\&D$	$s\&D$	$s\&\bar{s}$	$s\&\bar{s}$	s

Table 3 below shows information about the applicability of the measures. The first row, configurations, shows whether one can maximize on target attribute values, minimize or both maximize and minimize on the target attribute values (use the *absolute* values). The rest of the table tells when to maximize or minimize. The second row describes whether the measure can be applied to the continuous target value (rather than the rank). The last row explains whether measure values should be minimized or maximized when applied to the (partial) ranking.

Table 3. How to configure quality measures.

	φ_{avg}	φ_{mt}	φ_z	φ_t	φ_{χ^2}	φ_{roc}	φ_{wmw}	φ_{mmad}
Configurations (max/min/abs)	max&min	all	all	all	max	max	all	max
Original target	yes	yes	yes	yes	yes	no	no	no
Ranking (1 is best)	min	min	min	min	max	max	min	max

Table 4. Two artifical rankings, and four artificial subgroups defined on them. The value 1 indicates membership of s_i.

score	**rank 1**	**rank 2**	s_1	s_2	s_3	s_4
0.150	1.5	1	1	1	1	1
0.150	1.5	2	1	1	1	1
0.140	3	3	1	1	1	1
0.130	4.5	4	1	0	1	1
0.130	4.5	5	1	0	1	1
0.110	6	6	1	1	1	0
0.100	7	7	1	1	1	0
0.090	9	8	1	0	0	0
0.090	9	9	1	1	0	0
0.090	9	10	1	0	0	0
0.070	11.5	11	0	0	0	0
0.070	11.5	12	0	0	0	0
0.035	13.5	13	0	1	0	0
0.035	13.5	14	0	0	1	0
0.001	15	15	0	1	1	0
subgroup size			10	8	9	5

5.1 Intuitions

To illustrate the characteristics of the quality measures, we have performed a test on two artificial rankings, shown in Table 4. The objects are ranked in two ways: the first ranking (*rank 1*) is partial and based on the (artificial) values in the second column (*score*). *rank 2* is a complete ranking that is not related to the first two columns (as these have ties). The evaluation of the artificial subgroups $s_1, ..., s_4$ are shown in Table 5. The quality measures were used on the targets with which they can deal. Optimal values are in **bold**.

In Table 6, the quality intuitions are informally compared to the quality measures. As can be seen from Tables 4, 5 and 6, many quality measures ignore the *size* of a subgroup, especially for this small dataset. Most quality measures favour subgroup s_4 over s_1, although given the size intuition, subgroup s_1 clearly is more attractive than s_4. Only φ_{wmw} and φ_{mmad} favour different subgroups, where s_1 and s_4 tie when using φ_{wmw}, and s_3 and s_1 are very close when φ_{mmad} is used. φ_{mt}, φ_z and φ_t score reasonably well given the size intuition.

Table 5. Evaluation values on artificial dataset. Best evaluation values are in **bold**. Maximized values are shown in *italic*. The remaining values are minimized.

	φ_{avg}	φ_{mt}	φ_z	φ_t	φ_{χ^2}	φ_{roc}	φ_{wmw}	φ_{mmad}
				target=score				
s_1	*0.118*	*0.079*	*1.757*	*3.162*	*3.125*			
s_2	*0.097*	*0.011*	*0.251*	*0.21*	*3.696*			
s_3	*0.105*	*0.036*	*0.8*	*0.679*	*4.5*			
s_4	***0.14***	***0.105***	***2.335***	***10.51***	***8.571***			
				target=rank 1 (partial)				
s_1	5.5	-7.906	-1.781	-2.66	*5*		**-3.062**	*0.741*
s_2	7.063	-2.65	-0.597	-0.512	*3.4*		-0.868	*0.464*
s_3	6.278	-5.166	-1.164	-1.056	*2.7*		-1.827	***0.783***
s_4	**3**	**-11.18**	**-2.518**	**-7.454**	***7.5***		**-3.062**	*0.667*
				target=rank 2 (complete)				
s_1	5.5	-7.906	-1.768	-2.611	*3.571*	***1***	**-3.062**	*0.741*
s_2	7	-2.828	-0.632	-0.555	*3.411*	*0.64*	-0.926	*0.471*
s_3	6.333	-5.001	-1.118	-1	*3.696*	0.78	-1.768	***0.75***
s_4	**3**	**-11.18**	**-2.5**	**-7.072**	***8.125***	***1***	**-3.062**	*0.714*

On Intuition 4, *deviation*, all quality measures score reasonably well. Only φ_{χ^2} does not specifically take the deviation of subgroups into account. This is illustrated by subgroup s_2, where the evaluation value is relatively close to that of subgroups s_1 and s_3. φ_{roc} does not cover the third intuition (*position*) well. Subgroups s_4 and s_1 are judged equally good, although clearly the distribution of s_4 is much less like the dataset distribution than the distribution of s_1. The other quality measures, however, cover the third intuition quite well, not in the least since most compare the distribution of the subgroup to the dataset distribution. φ_{roc} however, covers the Intuition 2 (*rank*) very well, something that is illustrated again by s_1 and s_4. Except for the φ_{χ^2} measure, all measures capture the second intuition.

6 Experiments and Results

The presented quality measures for ordinal and numeric targets were implemented in the multi-relational data mining package Safarii [8], which already offers a generic Subgroup Discovery algorithm. The extensions for ranked data assume that the ranking of individuals in the database is determined by the target attribute, and that ordinal attributes start from the value 1, and use fractional ranks to represent ties, as described in Section 2. Safarii provides the possibility to mine multi-relational data, such that complex representations for individuals are no problem. In multi-relational data, it is assumed that the (ranked) individuals are represented by records in a so-called *target table*. All potential target attributes should therefore appear in this target table. The new implementation has been tested on a variety of ranked datasets. In this paper

Table 6. Informal qualification of quality measures given the intuitions. Values range from −− (very bad match) to ++ (very good match).

	φ_{avg}	φ_{mt}	φ_{z}	φ_{t}	φ_{χ^2}	φ_{roc}	φ_{wmw}	φ_{mmad}
I1: size	−−	+	+	−	−−	−−	++	++
I2: rank	++	+	+	+	−−	++	+	+
I3: position	+	++	++	++	++	−−	++	+
I4: deviation	+	+	+	++	−−	++	+	+

we present results on data related to the European Embryonal Tumour Pipeline project.

The generic SD algorithm in Safarii performs a heuristic search through the search space of candidate subgroups, guided by the selected quality measure. The heuristic search is essentially a beam-search with a configurable width and breadth, which is further bounded from below by a minimum support threshold. We have opted for heuristic search as in many cases, specifically of multi-relational nature, the search space prohibits the use of exhaustive methods. Safarii is able to deal with both numeric and nominal attributes in the subgroup description, although in the case of the biological application we present here, only nominal information was analysed (except for the target attribute of course). Although we present subgroups as conjunctions of (seemingly) propositional conditions, the descriptions are actually multi-relational, with implicit existential quantors accounting for the one-to-many relationships between the target table and the remaining background knowledge.

6.1 Neuroblastoma data

The EETP project is concerned with the analysis of genes potentially involved in a number of embryonal tumours. Our specific attention was focused on the childhood tumour of neuroblastoma. Since it is believed that neuroblastoma is not caused by environmental factors [2], research focuses on genetic factors, such as genes, gene expression, cell processes and so forth. Using a range of biological high-throughput analysis platforms, different aspects on these factors can be measured for activity in the tumour tissue. The typical outcome of such an analysis is a list of gene expressions measured over multiple patients, which in turn can be translated into a long list of genes, ranked by their differential expression. 'Differential' in this case refers to providing information about the difference between two classes of tumours, e.g., easily treatable and progressive. There are many ways to produce such a ranking of genes, most of them based on some measure of correlation between expression level and the class, such as the significance of a t-test or weighted relative accuracy [1]. For the neuroblastoma data, Safarii was used to produce such a primary ranking of genes from the expression data [9]. As the two target classes, we chose tumours from patients with no events after removal of the primary tumour, and tumours from patients who had a relapse, possibly with fatal outcome.

Given the ranking, the next obvious question is how to characterise the top-ranking genes, which apparently explain the difference between the two classes of tumours. In other words, one would like to gather additional information concerning each gene, and apply the presented Subgroup Discovery techniques in order to 'enrich' (the biological terminology) the ranking of genes [13–15]. The neuroblastoma-related gene ranking was joined with a substantial amount of background information that was obtained from a number of well-known online sources of genetic domain knowledge. The first two sources of gene-specific knowledge represent functional annotations (one or more functions selected from a large hierarchy) and come from the KEGG and GO databases (Gene Ontology [4]). Furthermore, information was added about which *protein families* [12] genes belong to. Protein families determine which proteins are related to each other (belong to the same family) given their chemical structure. Since proteins are translated from genes, protein families also give insight in gene families. As a third source of information, the genomic location of the gene was added, that is, the chromosome it appears on, and the exact location at different levels of detail (so-called *cytobands*). Finally, a gene interaction network was included, such that for each gene in the ranking, a set of genes can be specified with which it is known to interact. As such, we can look for secondary genes that interact with the majority of the top-ranking genes. Each of these four sources was represented as a separate table, such that the final database under investigation comprised 5 tables, of which the ranked table of genes was assigned the target table. Because of one-to-many relationships between some of these tables, our dataset constitutes a multi-relational problem [8, 14, 15]

Within the target table there are two attributes that can act as the target: an attribute (*score*) representing a measure of differential expression of the gene in question, and a rank attribute (*rank*) that is derived from the *score*. As a result, we have three options:

1. treat *score* as the numeric target
2. treat *rank* as an ordinal target
3. treat *rank* as a numeric target

For our experiments, we have tested all three options, but here only report on the last two[5]. For each of these, the appropriate quality measures were tested. As the data was partially ranked, our results do not include those for φ_{roc}. Although we have extensively experimented with parameter-settings that lead to large search spaces (for specific quality measures), here we present results for moderate settings only, for reasons of comparability and efficiency. We report subgroups of at most three conditions ($d = 3$), coming from all four domains mentioned. The width of the beam-search was set to one hundred ($w = 100$), and a minimum support of five genes per subgroup was required ($c = 5$).

Qualifying the Results Table 7 shows for each quality measure the first 5 subgroups that were found. Each line represents a multi-relational description

[5] The results on the *score* attribute are mostly comparable to those presented here, with only slight changes in the order of subgroups reported.

of a subgroup, with implicit references to the background table in question. As an example, the first description for $\varphi_{avg}(s)$ should be interpreted as follows: *all genes X for which there is an interacting gene Y called 'CDK3', and for which there is a interacting gene Z called 'CDC2'.*

An important observation is that, irrespective of the quality measure, there are patterns so strong that they are found by most quality measures. These patterns are, amongst others, the GO-terms *cell cycle*, *nucleus* and *mitosis*. Furthermore, as can be seen from Table 7, the largest subgroups have patterns in which a GO-term is included.

GO-terms and gene-to-gene interactions in relation to neuroblastoma tumours have been investigated in other studies. Some of the patterns we found, were also found in other studies. For instance, the GO-terms *DNA replication*, *cell cycle* and *DNA replication initiation* can be found in [11]. Other notable patterns that were found are interacting genes of the CDC, CDK and MCM families, such as CDC2, CDC7, CDC25A, CDK2, CDK3 MCM2, MCM3 and MCM10. These genes and families are known to play a role in cell division control, and their expression is specifically associated with carcinogenesis [2].

Several protein families also occurred frequently: Histone and MCM. A few of the most frequent chromosomes (or subsections of chromosomes) are chromosomes X, 6, 1, 11 and 17, along with cytobands Xq28, 6p22.1 and 17p11.2. Chromosomes 1, 11 and 17 or cytobands of these chromosomes have been identified as relevant for neuroblastoma in the past [11, 2]. Note that the important findings in the literature were produced with great effort, and using various biological analysis techniques, while our results are based on fairly standard expression data only. More extensive results from our experiments on neuroblastoma can be found in [10].

Comparison of Quality Measures The quality measures return different subgroups, with different conditions and different sizes, according to the measures' characteristics. Although the subgroups are not equal, we are interested in the performance of the quality measures in terms of how the quality of the subgroups develops. Since the scales of the measures are incomparable, we decided to normalize them. After normalization, the best performing subgroup is assigned the value 1, whereas the rest of the subgroups get assigned a value that lies between 1 and 0.

The normalized evaluation values for numeric and ordinal Subgroup Discovery are shown in Figure 1. The quality measures that produce average sized subgroups do not show a large decline. For instance, quality measures φ_{mt}, φ_z and φ_{wmw} return both large and small subgroups, with an approximate average size of 200 for the first 100 subgroups. The decline does not drop below 0.5. The results show that φ_{avg}, φ_t, φ_{χ^2} have a strong preference towards small subgroups. Also, not surprisingly, φ_{mmad} has a strong preference towards large subgroups. Note that φ_{χ^2} produces many subgroups on the minimum support level (5) with equal scores. The subgroups involve information from all domains, although the 5 subgroups shown in Table 7 suggest otherwise.

Table 7. First five subgroups returned by each quality measure. The ... for $\varphi_{\chi^2}(s)$ indicate that more than five subgroups share the same score and size. The five subgroups shown here are an arbitrary selection of this larger group

$\varphi_{avg}(s)$	**score**	**size**
gene2gene = CDK3 ∧ gene2gene = CDC2	-151.1	5
gene2gene = CDC25A ∧ gene2gene = CDK3	-243.5	5
gene2gene = CDC7 ∧ func = GO:0003677: DNA binding	-354.2	7
pfam = MCM ∧ func = KEGG:04110: Cell cycle	-409.3	6
gene2gene = CDK3 ∧ gene2gene = CDK2	-414.8	5

$\varphi_{mt}(s)$ **and** $\varphi_z(s)$	**score**	**size**
func = GO:0005634: nucleus ∧ func = GO:0007049: cell cycle	46692/9.34	230
func = GO:0005515: protein binding ∧ func = GO:0005634: nucleus	45416/9.08	1105
func = GO:0051301: cell division	44643/8.93	144
func = GO:0007049: cell cycle	43844/8.77	343
func = GO:0007067: mitosis	43542/8.71	111

$\varphi_t(s)$	**score**	**size**
gene2gene = CDC2 ∧ gene2gene = CDK3	137.3	5
func = GO:0007067: mitosis ∧ gene2gene = CDC20	136.1	6
func = GO:0005634: nucleus ∧ pfam = Kinesin	85.5	5
func = GO:0007067: mitosis ∧ gene2gene = E2F4	79.9	6
gene2gene = CDC25A ∧ gene2gene = CDK3	63.3	5

$\varphi_{\chi^2}(s)$	**score**	**size**
pfam = zf-UBR	17307	5
pfam = zf-C2HC	17307	5
pfam = zf-AN1	17307	5
pfam = WSC	17307	5
pfam = Vps4_C	17307	5
...		

$\varphi_{wmw}(s)$	**score**	**size**
func = GO:0005634: nucleus	9.54	3128
func = GO:0007049: cell cycle ∧ func = GO:0005634: nucleus	9.40	230
func = GO:0005515: protein binding ∧ func = GO:0005634: nucleus	9.39	1105
func = GO:0051301: cell division	8.97	144
func = GO:0007049: cell cycle	8.86	343

$\varphi_{mmad}(s)$	**score**	**size**
func = GO:0005634: nucleus	0.1577	3128
func = GO:0016020: membrane	0.1544	3466
func = GO:0005515: protein binding	0.1540	3152
func = GO:0016021: integral to membrane	0.1109	2543
func = GO:0016021: integral to membrane ∧ func = GO:0016020: membrane	0.098	2234

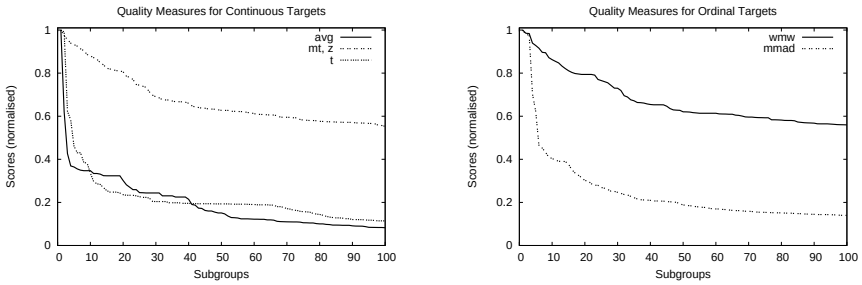

Fig. 1. The normalized evaluation values for numeric and ordinal Subgroup Discovery, for the first 100 subgroups.

7 Discussion and Conclusions

In this paper, we have presented a list of eight quality measures for ranked applications, with a variety of properties. Whether the different characteristics of each measure are beneficial or not often depends on the application domain and the specific demands of the analyst. By listing the basic properties in Tables 2 and 3, and assessing the match with our four intuitions in Table 6, we have provided some detailed guidelines on how to choose an appropriate quality measure for a given SD task. Furthermore, we have demonstrated the relative effectiveness of each quality measure in an important biological application, giving more specific input as to the question of what measure to use. Here, we summarize some of the conclusions, and give some suggestions for the optimal choice of measure.

The first choice to consider is whether the provided (partial) ranking is derived from some continuous measure (such as the *score* in the neuroblastoma application), or whether the ranking is based on a discrete process (such as a series of binary comparisons), as is customary in for example sports. In the latter case, an ordinal setting is obvious, and the three ordinal measures are applicable: $\varphi_{roc}(s)$, $\varphi_{mwm}(s)$, and $\varphi_{mmad}(s)$. In the case where both a rank and a score are available, one has the choice to use either target, resulting in either a ordinal setting, or both an ordinal and regression setting, respectively. The difference between using the rank and the score is a question of distribution that may sometimes be hard to answer. In general, it is wise to choose the target that allows for the most clear distinction. For example in cycling, a final sprint may show only marginal differences in time (actually all runners in the same group will typically be assigned the same time), whereas the ranking is crucial. In this case, the ordinal setting is obvious, and only three measures remain.

A further reduction in choice of measure is governed by the presence of ties. If the ranking is partial, this immediately rules out $\varphi_{roc}(s)$. All remaining measures can naturally deal with partial rankings.

A third issue to be considered is whether the size of the subgroup should be a factor in evaluating the candidate subgroups (Intuition I1). If the size of the subgroup is not an issue, except for a minimum size constraint not included in the measure, there may be a preference for $\varphi_{roc}(s)$ (ordinal), $\varphi_{avg}(s)$, $\varphi_{\chi^2}(s)$, and to a lesser extent $\varphi_t(s)$ (regression). In the more obvious case where larger subgroups should be encouraged, the remaining four measures work well. In this case, a minimum subgroup size constraint may be dispensed with.

Finally, when dealing with two order-equivalent measures, the ranking of found subgroups is of course identical. For this reason, the choice between $\varphi_z(s)$ and $\varphi_{mt}(s)$ is arbitrary. Still, we suggest using $\varphi_z(s)$, as this produces 'normalized' values, and thus allows easy interpretation of scores across different domains. Note that this contradicts the traditional use of $\varphi_{mt}(s)$ in SD for regression.

In the neuroblastoma application, the use of $\varphi_{roc}(s)$ is not an option, due to the presence of ties. Encouraging Intuition I1 is really a matter of taste here. Larger subgroups, as produced by for example $\varphi_z(s)$ and $\varphi_{mmad}(s)$, may be interesting as they uncover larger trends in the gene ranking. However, the high-level findings such as 'cell cycle' and 'mitosis' are often met with skepticism by the domain experts as they are well-known in cancer research, and not specific to the disease under investigation. For this reason, size-ignorant measures may be more desirable and provide more detailed information (e.g. the CDK3/CDC2 combination). It should be noted that the obvious findings particularly common in gene set enrichment are to some degree a result of the Gene Ontology [4], with its hierarchy of functions. More surprising findings turn up in the ranking of subgroups by $\varphi_z(s)$, once the first say ten obvious results are skipped. The same happens when focusing on background knowledge that inherently leads to smaller subgroups, such as the gene2gene database.

7.1 Conclusion

Although thus far, there has been limited interest in the analysis of ranked data, or in numeric and ordinal Subgroup Discovery for that matter, we have argued and demonstrated that ranked SD is an import asset to the Data Miner's toolbox. Furthermore, we have explained the large range of application areas where the analysis of ranked data is required, with Bioinformatics being an important example.

The Gene Set Enrichment experiments on neuroblastoma show that ranked SD can perform well. Patterns that have been found doing extensive research, have also been found using the more inexpensive techniques described in this paper. Given that known good patterns have been reproduced, the equally scoring patterns not familiar to the domain experts can be expected to represent reliable biological knowledge also. In this sense, SD in general, and ranked SD specifically, can aid domain experts in their research, and produce hypotheses that can be easily validated using traditional biological techniques.

Acknowledgements

The research in this paper was supported by the European Union through the EETP consortium (project nr. 037260) and the Netherlands Consortium for Systems Biology. Ivica Slavkov from the Jožef Stefan Institute assisted with the preparation of expression data and gene rankings.

References

1. Martin Atzmueller. Subgroup discovery. *Künstliche Intelligenz*, (4):52–53, 2005.
2. Garrett M. Brodeur. Neuroblastoma: Biological insights into a clinical enigma. *Nature Reviews*, 3:203–216, 2003.
3. Morris H. DeGroot and Mark J. Schervish. *Probability and Statistics*, chapter 8,9. Addison-Wesley, 2002.
4. The gene ontology, 2009. `http://www.geneontology.org`.
5. Henrik Grosskreutz. Cascaded subgroups discovery with an application to regression. In *ECML PKDD '08: Proceedings of the 19th European Conference on Machine Learning and 12th European Symposium on Principles and Practice of Knowledge Discovery in Databases*, 2008.
6. David J. Hand and Robert J. Till. A simple generalisation of the area under the roc curve for multiple class classification problems. *Mach. Learn.*, 45(2):171–186, 2001.
7. Willi Klösgen. Explora: a multipattern and multistrategy discovery assistant. Advances in knowledge discovery and data mining, pp 249–271, 1996.
8. Arno Knobbe. *Multi-Relational Data Mining*. IOS Press, Amsterdam, 2006. `http://www.kiminkii.com/thesis.pdf`.
9. van de Koppel, E., Slavkov, I., Astrahantseff, K., Schramm, A., Schulte, J., Vandesompele, J., de Jong, E., Dzeroski, S., Knobbe, A. Knowledge Discovery in Neuroblastoma-related Biological Data. *Data Mining in Functional Genomics and Proteomics workshop*, 2007
10. Barbara F. I. Pieters. Subgroup discovery on numeric and ordinal targets, with an application to biological data aggregation. Technical report, Department of Information and Computing Sciences, Utrecht University, 2010.
11. Katleen De Preter, et al. Human fetal neuroblast and neuroblastoma transcriptome analysis confirms neuroblast origin and highlights neuroblastoma candidate genes. *Genome Biology*, 7:R84, 2006. `http://genomebiology.com/2006/7/9/R84`.
12. Robert D. Finn, et al. The pfam protein families database, 2008. `http://pfam.sanger.ac.uk`.
13. Aravind Subramanian, et al. *Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles*. PNAS October 25, 102:43, 2005.
14. Igor Trajkovski. *Functional Interpretation of Gene Expression Data*. PhD thesis, Jožef Stefan International Postgraduate School, 2007. `http://cs.nyus.edu.mk/trajkovski/data/phd_thesis.html`.
15. Trajkovski, Igor and Lavrač, Nada and Tolar, Jakub. *SEGS: Search for enriched gene sets in microarray data*, J. of Biomedical Informatics, 41:4, 2008.