Exceptional Model Mining

Duivesteijn, W.

Chapter 8

Exploiting False Discoveries – Validating Found Descriptions

In Exceptional Model Mining, just like in Subgroup Discovery, we explore a large search space to find subsets of the data that have a relatively high value for a designed or selected quality measure. As we addressed in Chapter 3, the magnitude of the candidate space is potentially exponential in the number of records. Therefore, the process suffers from the multiple comparisons problem [52], which roughly states that, when considering a large number of candidates for a statistical hypothesis, some candidates will inevitably be incorrectly labeled as passing the test. Hence one of the many practical problems in SD/EMM is that it is nontrivial to determine whether discovered descriptions are actual discoveries, or *false discoveries* caused by random artifacts.

In this chapter, we draw upon statistical theory to build a model for false discoveries. Using this model, a number of practical problems can be solved. When applying SD/EMM to a dataset, one is often faced with the nontrivial task of choosing the right parameters for the discovery algorithm, in order to obtain a reasonable collection of results. The problems we intend to address are related to these parameter-setting issues. First of all, with the gradually extending range of quality measures available, for 'classical' Subgroup Discovery [35, 68] but also for non-standard variants such as regression [46, 89] and Exceptional Model Mining [25, 71], the issue of selecting the right measure for the task at hand is often hard. Users of

discovery tools often choose the measure based on their personal familiarity, or simply proceed with the default choice. We aim to provide more objective guidelines for selecting the measure that is most likely to produce interesting and exceptional results, and present empirical results that indicate a partial order amongst quality measures.

A second algorithm-tuning question we intend to address is that of setting a minimum threshold for the selected quality measure. Different measures have different domains, and end-users find it hard to set a reasonable value. Ideally, one would like to choose a minimum quality, such that all descriptions exceeding this value are reliably exceptional, and do not include "random" results that stem from the potentially large search space and the multiple comparisons problem inherent to all discovery methods. In other words, given some desired significance level α (typically 5% to 1%), we would like to obtain the corresponding minimum quality for the measure and dataset in question. As a converse, but very related task, one would like to compute a p-value for each reported description, that indicates to what extent the result is statistically significant.

As mentioned, SD and EMM potentially suffer from the multiple comparison problem. The main contribution of this chapter is the introduction of a method that employs a randomization technique to build a statistical model for the false discoveries caused by the multiple comparisons problem. Using this statistical model, we can refute many insignificant results returned by the discovery algorithm, and thus identify a set of on average more interesting descriptions. Furthermore, we employ the statistical validation to provide an experimental comparison of measures, and propose a suitable choice of measure.

8.1 Problem Statement

As mentioned in the introduction, the main contribution of this chapter is a method that builds a statistical model for false discoveries. This model can be used to solve a plethora of practical problems, of which we will empirically illustrate the following two

1. given a dataset Ω, a quality measure φ and a set $\mathcal{S}$ of descriptions found with this measure through SD/EMM, determine the statistical significance of each description $D \in \mathcal{S}$;
2. given datasets $\Omega_1, \ldots, \Omega_t$, determine which of the given quality measures $\varphi_1, \ldots, \varphi_g$ are better in distinguishing the top-q descriptions found with that measure from a random baseline, for a given q.

8.2 Validation Method

Philosophically speaking, we deal with the multiple comparisons problem (MCP) by a strategy akin to the informal saying "if you can't beat 'em, join 'em". The MCP is caused by random artifacts generating false discoveries. Our plan is to manage this process, generating some false discoveries *of our own*. Then, we can put these *artificial false discoveries* to work, to distinguish the true from the false discoveries in real SD/EMM results.

Informally speaking, to solve problem 1 from the previous section, we perform the following actions. First, we generate a number of false discoveries, by running SD/EMM on datasets obtained using a randomization technique. Then, we build a global model over the qualities of these false discoveries. Finally, we compare the qualities of the descriptions found on the original dataset with this global model. Descriptions that perform substantially better than the model for artificial false discoveries, are deemed likely to be true discoveries.

To solve problem 2 from the previous section, we perform these preceding actions for each quality measure. Then we compare the significance assessments of the top-q descriptions between the quality measures φ_i, distilling a preferential ranking of the quality measures for a dataset Ω_j. We do this for all datasets $\Omega_1, \ldots, \Omega_t$, which allows us to draw conclusions on whether quality measures are consistently preferential over one another.

The solutions to the two problems can be incorporated into one encompassing validation method, consisting of three consecutive steps. Specific choices for each of these steps will be thoroughly examined in the subsequent three sections. For now, we express the validation method in the following, somewhat formal manner

Validation Method. *Suppose a dataset Ω, quality measures $\varphi_1, \ldots, \varphi_g$, and sets of descriptions $\mathcal{S}_1, \ldots, \mathcal{S}_g$ where $\forall_{i=1}^{g} : \mathcal{S}_i$ is found through SD/EMM using quality measure φ_i. The method consists of the following steps*

I. $\forall_{i=1}^{g}$: *use a randomization technique to generate baseline descriptions $B_1^i, \ldots, B_x^i \subseteq \Omega$ for arbitrarily large x;*

II. $\forall_{i=1}^{g}$: *build a statistical model for false discoveries based on the qualities $\varphi_i\left(B_1^i\right), \ldots, \varphi_i\left(B_x^i\right)$ of the baseline descriptions. Then determine for each $D \in \mathcal{S}_i$ how much $\varphi_i(D)$ deviates from the model;*

III. choose any positive integer q, and determine preference between the quality measures by comparing the deviations corresponding to the top-q descriptions in $\mathcal{S}_i$.

In order to solve problem 1, we need steps I and II of the method, for $g = 1$ (since there is only one quality measure). Solving problem 2 requires taking all three steps, and repeating them for each dataset $\Omega_1, \ldots, \Omega_t$.

Since we determine the statistical soundness of quality measures in terms of their ability to deviate from a random baseline, we could interpret this as a test to what extent a quality measure is also a measure for exceptionality. Notice that our method does not consider the coherence of a set $\mathcal{S}$ of descriptions: we do not solve the problem of redundancy within such a set, we do not solve the problem of selecting a small subset of jointly interesting descriptions in $\mathcal{S}$, we merely consider for every single description in $\mathcal{S}$ the likelihood that it is deemed interesting because of the multiple comparisons problem.

8.2.1 Randomization Techniques

There are several randomization techniques we can use to generate the baseline descriptions $B_1^i, \ldots, B_x^i$ (i.e. perform step I of the method). We will employ the randomization technique that is currently the most popular in data mining: swap randomization. Gionis et al. have published a paper detailing its use in data mining [43]. In its most radical form for zero-one matrices, swap randomization shuffles the elements of the dataset in such a way that all row and column sums are kept intact, which is what

the authors of [43] have done for tests involving itemset mining. Swap randomization is also frequently used for validating classifiers in a more moderate form: only the column containing the class labels is replaced by a random permutation of itself. For SD/EMM, it seems reasonable to use this moderate form of swap randomization: each target column is shuffled, independently (cf. Section 8.6) from the other target columns.

We generate the baseline descriptions in the following way. For each B_j^i to be generated, we create a swap-randomized version of the data, by keeping all descriptors intact but applying a random permutation to each target column. Then we run our SD/EMM algorithm on the resulting dataset using quality measure φ_i, and let B_j^i be the best description found.

The rationale behind this process is that by swap-randomizing each target column, we keep its distribution intact. However, we randomize all dependencies between the target columns and the descriptors, and we randomize all internal dependencies between the target columns. Hence the best description found on the swap-randomized data represents the best-quality discovery made while there is no connection between target column and other attributes, and no connection between multiple target columns, apart from connections caused by random artifacts. In other words, this best description represents a false discovery, and its quality is among the highest qualities a false discovery can have.

Another reason why a description found on the swap-randomized data is a good representation of a false discovery is the fact that its discovery has resulted from the same search process as employed while discovering actual descriptions on the original dataset. Alternatively one could easily choose a method to directly generate some random baseline description for use in step I of our method. However, a description found on swap-randomized data goes through the same motions of the SD/EMM algorithm as the actual descriptions found on the original dataset, i.e. the same hypothesis space is traversed, the traversal is performed in the same way, and the search is bounded by the same constraints. Hence the generated false discovery can reasonably be considered a false discovery of the search process.

8.2.2 Building a Statistical Model

When we have generated the baseline descriptions, there are plenty of ways to build a statistical model from them (i.e. perform step II of the method). The most straightforward technique, and the simplest in terms of statistical interpretability, is a direct application of the central limit theorem (CLT) [75]. Under the assumption that x (the number of baseline descriptions) is sufficiently large, according to the central limit theorem, the mean of $\varphi_i\left(B_1^i\right), \ldots, \varphi_i\left(B_x^i\right)$ follows a normal distribution, since these are independent and identically distributed random variables. We use the sample mean ($\hat{\mu}$) and sample standard deviation ($\hat{\sigma}$) as distribution parameters, as suggested by the method of moments [88]. We call this distribution, $\mathcal{N}(\hat{\mu}, \hat{\sigma})$, the *Distribution of False Discoveries* (DFD). Let $D \in \mathcal{S}$ be a description under consideration. We can now formulate the null hypothesis

$$H_0 : \varphi(D) \text{ is generated by the DFD}$$

We can compute a p-value corresponding to this null hypothesis for each $D \in \mathcal{S}$, and this p-value gives us the deviation required as result of step II of the method.

Notice that although the null hypothesis is fixed, its interpretation may vary depending on the randomization technique employed in step I of the method.

Using the DFD, we can not only validate a found description, but also compute threshold values for the quality measure at given significance levels, prior to the actual mining run. Such a threshold could be used as lower bound on the quality of a description in the SD/EMM process. This is a nontrivial contribution to the process, since it is generally not easy for an end-user to set a sensible lower bound for any given quality measure. Additionally, such sensible values for a lower bound depend heavily on the dataset at hand. Until now, it was common to use a default value for such a lower bound by lack of a better method; the DFD gives us more sensible threshold values.

8.2.3 Comparing Quality Measures

For performing step III, comparing the relative performance of the quality measures, we use a technique recently described by Demšar in an article [19] on statistical comparisons of classifiers over multiple datasets. First a Friedman test [31, 32] is performed to determine whether the quality measures all perform equivalently. This is a non-parametric version of the repeated-measures ANOVA. For each test case the quality measures are ranked by their performance; in case of ties we assign average ranks. Let r_i denote the average rank over all test cases for quality measure φ_i, $\forall_{i\in\{1,\ldots,g\}}$, and let T denote the number of test cases. The null hypothesis states that all measures perform similarly, hence their average ranks should be equal. Under this null hypothesis, the Friedman statistic

$$\chi_F^2 = \frac{12T}{g(g+1)} \cdot \sum_i \left(r_i - \frac{g+1}{2} \right)^2$$

follows a chi-squared distribution with $g-1$ degrees of freedom.[1]

If the null hypothesis of the Friedman test is rejected, we can determine which quality measures are significantly better than others with a post-hoc test. Following Demšar's proposal, we use the Nemenyi test [83], which is similar to the Tukey test for ANOVA. In this test a critical difference (CD) is computed

$$CD = q_\alpha \sqrt{\frac{g(g+1)}{6T}}$$

where the critical values q_α are based on the Studentized range statistic [94, pp. 451–452] divided by $\sqrt{2}$. If the difference between the average ranks of two quality measures surpasses this CD, then the better-ranked measure performs significantly better.

8.3 Experiments

To illustrate how our method performs, we run Subgroup Discovery experiments on several datasets. The bulk of our empirical evaluation will

[1] Careful readers may notice that this formula is not the one given by Demšar. It is, however, the one given by Friedman himself. Equivalence can be shown in four lines of math; the equation shown here is slightly easier to compute, and easier on the eye.

Table 8.1: UCI datasets used for the DFD experiments.

	Dataset	N	# descriptors		$\|\ell\|$
			discrete	numeric	
1.	*Adult*	48842	8	6	2
2.	*Balance-scale*	625	0	4	3
3.	*Car*	1728	6	0	4
4.	*CMC*	1473	7	2	3
5.	*Contact-lenses*	24	4	0	3
6.	*Credit-a*	690	9	6	2
7.	*Dermatology*	366	33	1	6
8.	*Glass*	214	0	9	6
9.	*Haberman*	306	1	2	2
10.	*Hayes-roth*	132	0	4	3
11.	*Ionosphere*	351	0	34	2
12.	*Iris*	150	0	4	3
13.	*Labor*	57	8	8	2
14.	*Mushroom*	8124	22	0	2
15.	*Pima-indians*	768	0	8	2
16.	*Soybean*	683	35	0	19
17.	*Tic-tac-toe*	958	9	0	2
18.	*Wisconsin*	699	0	9	2
19.	*Yeast*	1484	1	7	10
20.	*Zoo*	101	16	1	7

be done in terms of Subgroup Discovery with only one discrete target, but this is by no means essential to the method. In fact, it can be applied to any supervised Local Pattern Mining technique. We will briefly illustrate this by applying our method not only to traditional Subgroup Discovery, but also to the instance of Exceptional Model Mining with the correlation model class, as introduced in Chapter 4. Results of the experiments on traditional SD are described in Sections 8.3.1 and 8.3.2, and results on the EMM instance in Section 8.3.3.

We pick the following parameters for the beam search process. On each level, we select the $w = 25$ best descriptions, and refine these to create the candidate descriptions for the next level. To bound the complexity of the descriptions we use a search depth of $d = 3$. We let $minsup = \lfloor \frac{N}{10} \rfloor$, i.e. a description must be covered by at least 10% of the dataset. These parameter settings are somewhat arbitrary; we believe that this is not really relevant for the purpose of demonstrating our new method.

The 20 datasets we have used for our tests with traditional SD, can be found in the UCI Machine Learning Repository [3]. Table 8.1 contains details on the datasets considered. Here, $|\ell|$ denotes the number of distinct target values in the dataset. Notice that, rather unfortunately, this table features a second dataset named *Yeast*, after the one introduced in Table 6.1. The dataset considered in this chapter is the *Yeast* dataset that can be found through UCI [3], and not the *Yeast* dataset that was introduced in the paper by Elisseeff et al. [28].

Before experimenting with the method, let us empirically investigate a simpler solution: *empirical* p*-values*. Given a description D to be validated, one could simply assign as p-value the fraction of randomly generated results that outperform D. Though valid as a validation method for single descriptions, the empirical p-values lack the expressive power necessary to perform step III of our validation method, which is required when we want to validate quality measures. This is illustrated by the histogram (represented by the jagged line), displayed in Figure 8.1, of qualities of 1000 random subsets on the *CMC* dataset with target value *'no-use'*, normalized into Z-space (i.e. a subset has value one on the x-axis in the histogram when its quality is one standard deviation higher than the sample mean). The figure also contains our CLT-based normal distribution fitted to the

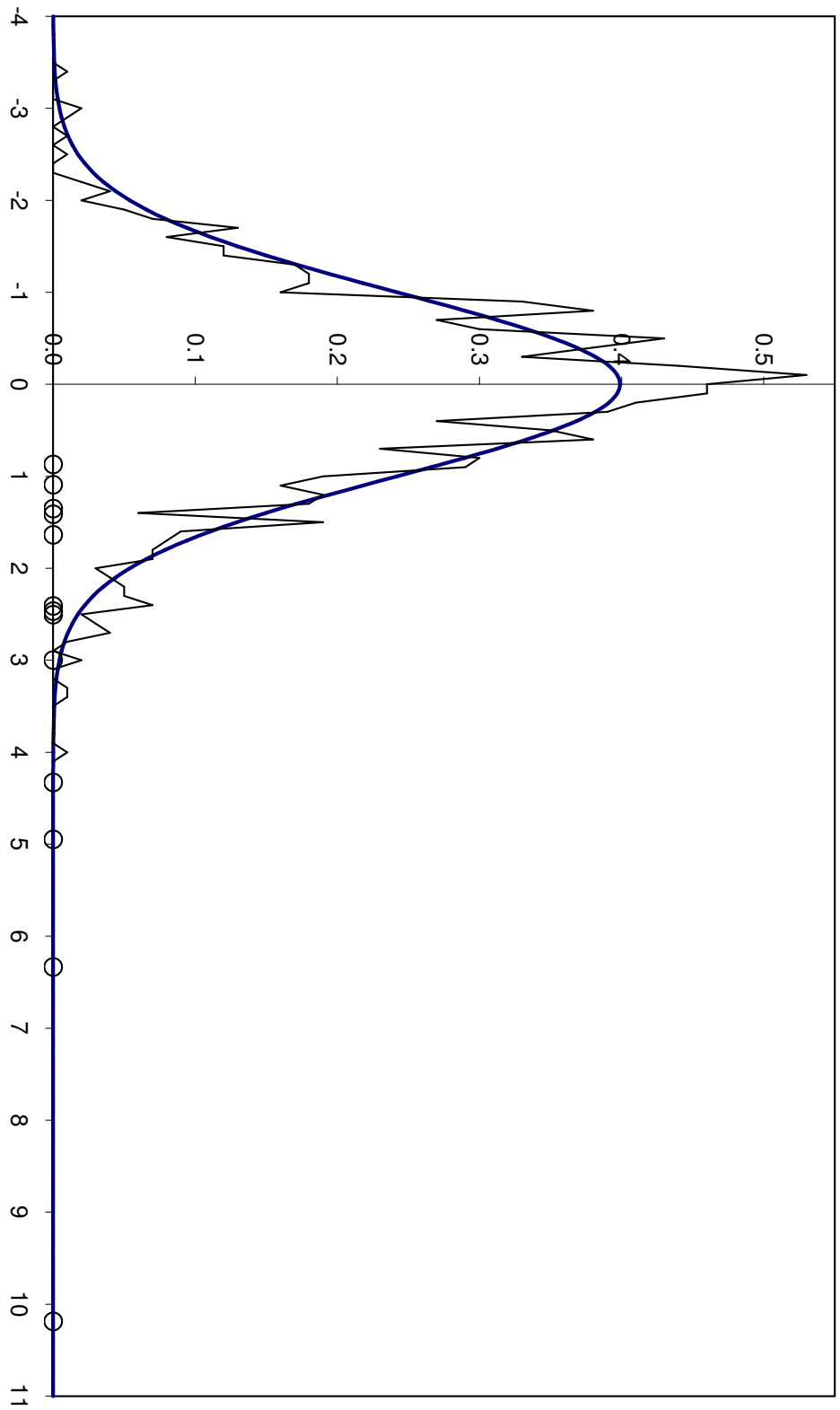

Figure 8.1: CLT-based model versus empirical p-values.

random qualities (represented by the smooth curve) and the 13 descriptions (represented by circles on the x-axis) found using a very shallow search of $d = 1$. The rightmost nonzero value of the histogram occurs at $x = 4$, hence all descriptions to the right of that point are indistinguishable by empirical p-values. The normal distribution never becomes zero, hence does not suffer from this problem. Hence, with the CLT-based model we can still compare the relative significance of the 4 best descriptions, whereas with empirical p-values we cannot.

8.3.1 Validating Descriptions

We will first illustrate how to use our method to solve problem 1 from Section 8.1: validating single descriptions. To this end, we only need the method's first two steps.

We consider just one quality measure: Weighted Relative Accuracy (WRAcc) [55], arguably the most popular quality measure in Subgroup Discovery. For each dataset, we perform an SD run for each target value, and report the 1000 best descriptions. We then run the first two steps of our method to determine how many of the found descriptions remain if insignificant descriptions are removed. We report the average fraction of descriptions that is retained per dataset for different significance levels in Table 8.2.

As stated in Section 8.2, one could also use the Distribution of False Discoveries to determine quality measure thresholds for given significance levels, a common practical issue with SD/EMM exercises. We illustrate this by determining thresholds on the *Contact-lenses* dataset with target value *'none'*. Notice that WRAcc can theoretically assume values between -0.25 and 0.25. We find that with significance level $\alpha = 10\%$ a subgroup needs to have a WRAcc of at least 0.054 to reject the null hypothesis that it is a false discovery, with $\alpha = 5\%$ a subgroup needs to have a WRAcc of at least 0.068, and with $\alpha = 1\%$ a value of at least 0.093. To provide context: the best description found on this dataset with this target value has a WRAcc of 0.188.

Table 8.2: Fraction of descriptions retained when removing insignificant descriptions.

Dataset	$\alpha = 10\%$	$\alpha = 5\%$	$\alpha = 1\%$
Adult	1.000	1.000	1.000
Balance-scale	0.561	0.554	0.548
Car	0.650	0.591	0.518
CMC	0.506	0.484	0.445
Contact-lenses	0.069	0.069	0.052
Credit-a	1.000	1.000	1.000
Dermatology	0.838	0.808	0.761
Glass	0.738	0.675	0.562
Haberman	0.427	0.392	0.327
Hayes-roth	0.388	0.313	0.210
Ionosphere	1.000	1.000	1.000
Iris	0.902	0.879	0.834
Labor	0.628	0.567	0.401
Mushroom	0.967	0.966	0.964
Pima-indians	1.000	1.000	1.000
Soybean	0.724	0.713	0.689
Tic-tac-toe	0.493	0.446	0.311
Wisconsin	1.000	1.000	1.000
Yeast	0.687	0.673	0.647
Zoo	0.600	0.582	0.524

8.3.2 Validating Quality Measures

We can build on the instantiation of our model that we used in the previous section to solve problem 2 from Section 8.1: validating quality measures. We select 12 quality measures for single discrete targets that are quite common in Subgroup Discovery, and test them against each other. The measures are WRAcc, |WRAcc|, χ^2, Confidence, Purity, Jaccard, Specificity, Sensitivity, Laplace, F-measure, G-measure, and Correlation. Details on these measures and their origins can be found in the paper by Fürnkranz and Flach [35].

Table 8.3: Average ranks of the quality measures.

	All datasets		Binary target	
Measure	$q = 1$	$q = 100$	$q = 1$	$q = 100$
χ^2	4.435	4.038	4.694	3.889
Jaccard	5.224	5.622	5.361	7.028
Correlation	5.235	4.679	5.361	4.667
|WRAcc|	5.288	4.571	5.306	4.333
G-measure	5.312	5.538	5.417	6.750
F-measure	5.582	5.718	5.250	6.778
WRAcc	5.800	5.027	5.417	4.722
Confidence	6.506	6.865	7.333	7.028
Laplace	6.553	6.654	7.278	6.139
Specificity	7.465	8.455	8.306	7.806
Purity	10.235	10.141	8.389	7.361
Sensitivity	10.365	10.692	9.889	11.500
χ^2_F $(\alpha = 1\%)$	261.916	292.001	40.674	57.618
CD $(\alpha = 1\%)$	2.069	2.160	4.496	4.496

For each dataset, we perform steps I and II of our method in the same way as in the previous section, with each of the 12 quality measures. We then compare the measures in step III by comparing the p-values of the q best descriptions, for both $q = 1$ and $q = 100$ (for $q = 100$ we take the average p-values over the top-100 descriptions). Hence for all measures we obtain for both choices of q one test score for each combination of dataset and target value within that dataset. For $q = 1$ this leads to a grand total of 85 test scores for each quality measure. On both the *Car* and the *Contact-lenses* dataset, no 100 descriptions are found that satisfy the *minsup* constraint. Hence there are no results on these datasets for $q = 100$, leaving a total of 78 test cases for $q = 100$.

The measures are subsequently ranked, where a lower test score (p-value) is better. The resulting average ranks can be found in the second and third columns of Table 8.3. This table also displays the results of the Friedman tests, the values for χ^2_F. With a significance level of $\alpha = 1\%$ we need χ^2_F to be at least 24.73 to reject the null hypothesis that all quality measures perform equally well. Hence we comfortably pass this test.

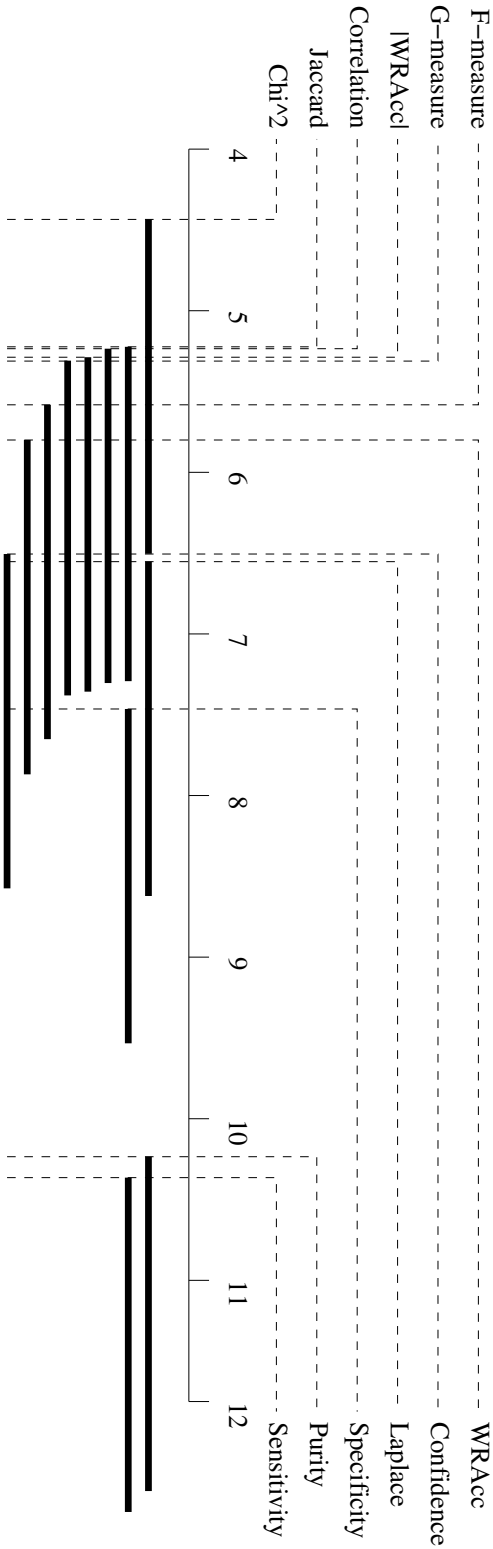

Figure 8.2: CD chart for $q = 1$ (CD $= 2.069$).

Since the Friedman test is passed, we can now perform Nemenyi tests to see which quality measures outperform others. For the $q = 1$ setting, the critical difference CD equals 2.069 with significance level $\alpha = 1\%$. For each pair of measures we compute from Table 8.3 whether their difference is larger than CD, and if so, the one with the smaller average rank is better than the other. The corresponding CD chart [19] can be found in Figure 8.2. Such a chart features a horizontal bar of length CD for each quality measure φ_i, starting at its average rank. Hence φ_i is significantly better than each quality measure whose bar starts to the right of the bar of φ_i. For instance, in Figure 8.2 we see that χ^2 is significantly better than Laplace, Specificity, Purity, and Sensitivity. Figure 8.3 displays the CD chart for the $q = 100$ setting.

When we have a dataset with many distinct target values, we repeatedly let one of the target values correspond to positive examples and the rest to negative examples. Hence the more distinct target values we have, the lower the average fraction of positive examples in the dataset. To see whether certain quality measures suffer from this effect, we also computed the average ranks considering only the 9 datasets with a binary target. The results can be found in the last two columns of Table 8.3. Again, the average ranks easily pass the Friedman test. Now that we have only 18 test cases, the critical difference for the Nemenyi test becomes $CD = 4.496$ with significance level $\alpha = 1\%$.

8.3.3 Validating EMM Results

So far we have illustrated our method with measures for Subgroup Discovery over a single discrete target. We now turn to the variant of EMM with the correlation between two targets as model class, as introduced in Chapter 4. In that chapter, we introduced three quality measures for the problem: φ_{abs}, φ_{ent}, and φ_{scd}. We validate these measures, together with some simpler measures that do not explicitly compare multiple models, but rather simply search for descriptions maximizing one particular quantity. For such quantities we consider a high positive correlation, by maximizing $\hat{r}$, a high negative correlation, by maximizing $-\hat{r}$, a high positive or negative correlation, by maximizing $\hat{r}^2$, and a near-zero correlation, by maximizing $-\hat{r}^2$. Notice that these four measures do not take the correlation on the

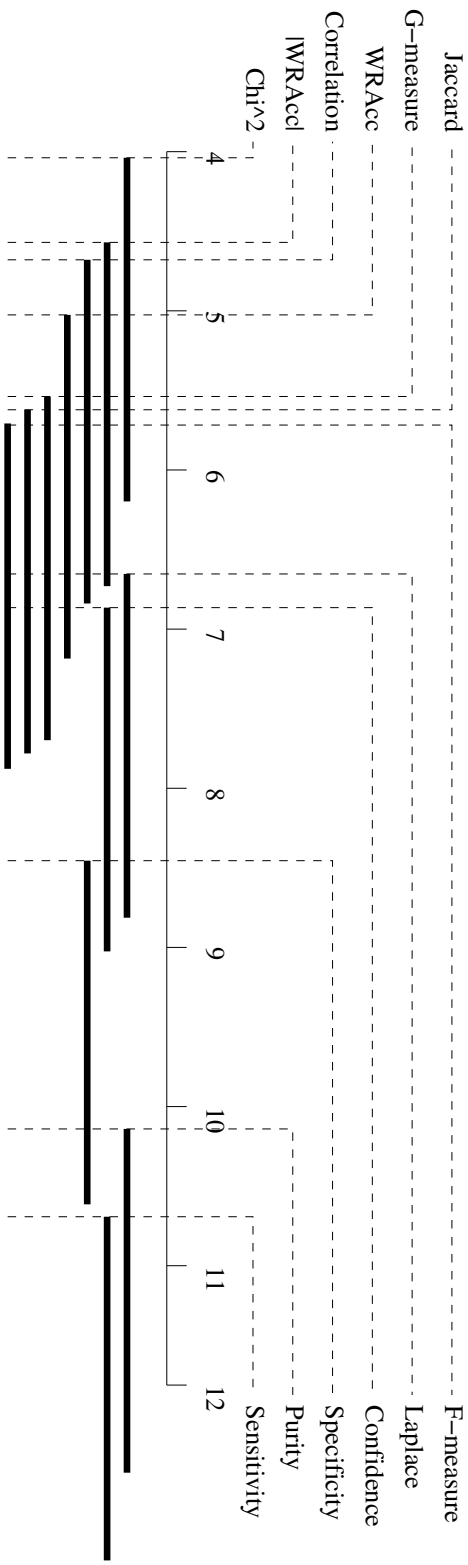

Figure 8.3: CD chart for $q = 100$ (CD $= 2.160$).

Table 8.4: Average ranks for correlation model measures.

Measure	Average rank
φ_{ent}	1.75
$\hat{r}$	3.00
$\hat{r}^2$	3.75
φ_{abs}	4.25
$-\hat{r}$	4.75
$-\hat{r}^2$	5.25
φ_{scd}	5.25
χ^2_F	21.96
CD $(\alpha = 1\%)$	4.114

whole dataset or the correlation on the complement into account. Hence they do not comply with the EMM quality measure design guidelines we outlined in Section 3.2.

We validate the 7 measures for this setting on the datasets and target concepts introduced in Table 4.1. The resulting average ranks over the 2 datasets — *Windsor Housing* and *Affymetrix* — can be found in Table 8.4. The Friedman test value for these ranks is $\chi^2_F = 21.96$, where 16.81 would be enough with 7 measures, so we can proceed with the Nemenyi test. The critical difference is CD $= 4.114$ with significance level $\alpha = 10\%$ when testing 7 measures on 4 test cases (aggregating over the results for $q = 1$ and $q = 100$). In these modest experiments, we find that no significant conclusions can be drawn.

8.4 Discussion

The previous section summarized the results experimentally obtained with our new method; in this section we will interpret them. We start with the results obtained by the technique for validating descriptions in a set $\mathcal{S}$ found through SD/EMM.

8.4.1 Validating Descriptions

From Table 8.2 we find that we cannot refute any descriptions from $\mathcal{S}$ in several datasets: *Adult*, *Credit-a*, *Ionosphere*, *Pima-indians*, and *Wisconsin*. To explain this result, we craft a metalearning dataset from Tables 8.1 and 8.2. We select the columns from Table 8.1 as descriptors of our metalearning dataset, and add three new columns, representing the total number of descriptors in the dataset, a boolean column representing whether the dataset has discrete descriptors, and a boolean column representing whether the dataset has numeric descriptors. As target column we add the last column of Table 8.2: the fraction of descriptions retained when insignificant descriptions are removed, with significance level $\alpha = 1\%$. On this metalearning dataset we perform a shallow (using search depth $d = 1$) but exhaustive Subgroup Discovery run, using Klösgen's mean test [55] as quality measure. The resulting metadescriptions should consist of those datasets with a relatively high fraction of kept descriptions.

The best metadescription is defined by the condition that the datasets have more than five numeric descriptors. The eight datasets covered by this metadescription are *Adult*, *Credit-a*, *Glass*, *Ionosphere*, *Labor*, *Pima-indians*, *Wisconsin*, and *Yeast*. This set encompasses all datasets for which we cannot refute any of the descriptions from $\mathcal{S}$. This makes sense, since for each dataset we have only considered the top-1000 descriptions, a fixed number independent of dataset characteristics. Numeric descriptors usually have many different values, resulting in a hypothesis space that is much larger than it would have been if the descriptors were discrete. Hence in datasets with relatively many numeric descriptors, it is more likely that the 1000 best descriptions represent relatively rare spikes in a quality distribution consisting mainly of low values. Therefore it is less likely that the random baseline incorporates some of these spikes, and thus the baseline is more likely to be relatively weak.

8.4.2 Validating Quality Measures

The results we obtained by the technique for validating quality measures show that χ^2 achieves the best performance of all quality measures in distinguishing the top-q descriptions from false discoveries. Many of the relations

between quality measures, however, are not significant. For $q = 1$, all other quality measures perform significantly better than Purity and Sensitivity. Additionally, Specificity performs significantly worse than Jaccard, Correlation, |WRAcc|, and the G-measure, and χ^2 significantly outperforms Laplace.

For $q = 100$, we see some slight changes: χ^2 and |WRAcc| now also perform significantly better than Confidence, and Specificity is now additionally outperformed by the F-measure and WRAcc while it no longer performs significantly better than Purity. Finally, Correlation significantly outperforms Confidence. Obviously, some measures might be better than others in distinguishing the top-q descriptions from false discoveries when $q = 1$, while others might be better when $q = 100$. The observed changes are not very dramatic, and we consider the selection of q a user-derived parameter in the method.

One of the significant relations that seems somewhat peculiar, is the result that for both $q = 1$ and $q = 100$, Confidence performs significantly better than Purity, while the latter is defined to be $\max\{\text{Confidence}, 1 - \text{Confidence}\}$. While there may be a good theoretical reason to consider the Purity of a description, we can see from the definition that Purity has a lower bound of 0.5, hence the random baseline will generate higher values with Purity than with Confidence. Apparently the quality of the descriptions found with Purity does not increase enough compared to those found with Confidence to compensate for this effect.

By comparing the second and third columns of Table 8.3 with the last two columns, we can see that |WRAcc|, WRAcc, and particularly Purity perform better when we restrict the tests to datasets with a binary target. These measures benefit from the fact that in these test cases we have a better balance between positive and negative examples in the data, compared to test cases on other datasets. We can also read from the table that we have fewer measures that are significantly better than others on datasets with a binary target. This is mainly because significance is hard to achieve in an experiment with only 18 test cases as opposed to 85 or 78 on all datasets. With 18 test cases, the critical difference for the Nemenyi test with significance level $\alpha = 1\%$ is $CD = 4.496$, rather than $CD = 2.069$ with 85 test cases. Since the average ranks range from 1 to 12, a critical

difference of 4.496 is substantial. More significant difference relations between the quality measures can be expected when experiments would be performed on more datasets with a binary target.

8.4.3 Validating EMM Results

The results for the EMM variant were generated on a modest number of test cases. As a result, the critical difference for the Nemenyi test is quite high, and one could not expect to find many significant results. Extensive experimentation may give a significant reason to prefer one measure over another in this setting. For now, what matters is that this illustrates that our method is applicable in more general settings than just traditional Subgroup Discovery.

Another noteworthy result, albeit non-significant, is the empirical illustration that a difference quantification is not always enough to make a good quality measure for EMM, as announced in Section 3.2.2. We had stated that deviations from the norm are easily achieved in small subsets of the data. Hence, quality measures that evaluate merely on the difference in model characteristics, such as φ_{abs}, will favor relatively small descriptions. It is relatively easy to find a small description with a high value for such a quality measure *on swap-randomized data as well*. This effect can be mitigated by incorporating an entropy term in the quality measure, as we have done with φ_{ent}. Recall that $\varphi_{\text{ent}} = \varphi_{\text{abs}} \cdot \varphi_{\text{ef}}$. As can be seen in Table 8.4, the incorporation of the entropy term can lift a quality measure with mediocre performance (φ_{abs}) in distinguishing real from false discoveries, to top-level performance (φ_{ent}).

8.5 Related Work

Statistical validation specifically tailored for Subgroup Discovery barely exists. Fortunately, many techniques for statistical validation in local pattern mining settings, which have been developed ever since association rules were invented, are applicable in Subgroup Discovery. Most of the recent approaches employ empirical p-values, as shortly introduced in Sec-

tion 8.3. This method has been applied in papers concerning significant query results on multi-relational databases [85] and swap randomization on high-dimensional 0/1 datasets [43]. In many circumstances, the use of empirical p-values is very appropriate. However, we attempt to validate descriptions with a high quality by comparing them to random descriptions that are expected to have a more moderate quality. Since we are trying to validate outliers in the quality measure distribution, in many cases we will find empirical p-values to be zero for many measures, hence they are not very useful for comparing the measures with each other.

A method that assigns nonempirical p-values to single association rules has been proposed by Megiddo and Srikant [79]. They use random approximation techniques to assign significance to single association rules and sets of associations. Unfortunately, their choice of underlying distribution is not motivated in any way.

Quality measures exist for Subgroup Discovery that directly implement a statistical significance test. For instance, one can show that Klösgen's mean test $\left(\sqrt{n}\,(p - p_0)\right)$ [55] is order-equivalent to a t-test. Also well suited for Subgroup Discovery is the chi-squared $\left(\chi^2\right)$ measure [98], originally defined for association rules. While such quality measures automatically statistically validate single descriptions, their application in Subgroup Discovery and hence use in a vast search space will invariably suffer from the multiple comparison problem, and hence the results will fall prey to the problem we attempt to solve in this chapter.

Tan et al. have developed a method [101] to compare quality measures on contingency tables by intrinsic properties. The results this method delivers are somewhat inconclusive, hence the method relies on experts to decide which measure is to be preferred. Also, the method seems not to be extendable beyond k-way contingency tables.

Finally, Webb devised a procedure to assign significance to individual descriptions [112]. He gives two different ways to perform a Bonferroni-style adjustment to the significance level: direct adjustment, and an approach that is very similar to the train-and-test-set procedure known from the determination of the predictive accuracy of a classifier. As is typical for Bonferroni correction, the adjustments may be a bit too strict. This especially holds when the search space becomes very large, for instance when

dealing with numeric descriptors. When applying a Bonferroni correction one assumes that the different hypotheses are independent, which in a Subgroup Discovery setting is not the case, leading to too strict adjustments to the significance level. Also, rather than being a method that assigns significance to descriptions, Webb's work is more a framework that can be used with any statistical hypothesis test.

8.6 Conclusions

We propose a method that deals with the multiple comparisons problem in SD/EMM, i.e. the problem that when exploring a vast search space one basically considers many candidates for a statistical hypothesis, hence one will inevitably incorrectly label some candidates as passing the test. Our method tackles this problem by building a statistical model for the false discoveries: the *Distribution of False Discoveries* (DFD). This distribution is generated by, given a dataset and quality measure, repeatedly running an SD/EMM algorithm on a swap-randomized version of the data. In this swap-randomized version, while the distribution of each target is maintained, the correlations with the descriptors and the correlations between targets are destroyed. Hence the best description discovered on this dataset represents a false discovery. The DFD is then determined by applying the central limit theorem to the qualities of these false discoveries.

Having determined the DFD, one can solve many practical problems prevalent in SD/EMM. For any discovered description, one can determine a p-value corresponding to the null hypothesis that it is generated by the DFD; refuting this null hypothesis implies that the description is not a false discovery. Given a set of quality measures, one can use the DFD to determine which quality measures are better than others in distinguishing the top-q descriptions from false discoveries. This gives an objective criterion for selecting a quality measure that is more likely to produce exceptional results. Finally, given some desired significance level α, one could extract from the DFD a minimum threshold for the quality measure at hand.

When validating single descriptions, we see that our method removes insignificant descriptions found on datasets that have few numeric descriptors. From metalearning we find that on large datasets, for instance with

more than five numeric descriptors, the random baseline is more likely to accept many descriptions. This is reasonable because of the associated larger hypothesis space. Table 8.2 shows that our method can remove insignificant descriptions on some of the datasets with more than five numeric descriptors, but not on all of them.

When we validate quality measures, we have outlined that the method we described determines the extent to which a quality measure is also an exceptionality measure. We have seen that of the twelve measures for Subgroup Discovery we tested, χ^2 is the best exceptionality measure, and Purity and Sensitivity are by far the worst. For the EMM correlation model variant no significant conclusions can be drawn from the modest experiments.

In this chapter we have presented a technique making extensive use of swap randomization. Notice that we do not by any means claim to have invented this particular randomization method. Also, its use in step I of the method we introduced in this chapter is not the only option available. We have extensively explained why using swap-randomized data leads to a good model for false discoveries, but it comes at a price: for every result of an SD/EMM run one wishes to validate, one has to run the same SD/EMM algorithm an additional x times, where x needs to be large enough to satisfy the constraints of the Central Limit Theorem. In the more traditional Subgroup Discovery setting, one can usually afford this extra computation time. For more complex settings, for instance the EMM variant using Bayesian networks introduced in Chapter 6, this becomes problematic. When computation time becomes an issue, one might consider different randomization techniques to generate $B_1, \ldots, B_x$, for instance by simply drawing a random sample from Ω of a certain size for each B_i. Before such a technique can be employed, its theoretical ramifications need to be explored. In future work, we also plan to empirically investigate the effect of certain parameters on the outcome of the method.

Another randomization-related point that is worthy of further investigation, is induced by the general applicability of the validation method beyond traditional Subgroup Discovery. Most of the experiments run in this chapter concern SD. In this setting, the exact implementation of swap-randomizing the target is clear-cut, as well as its philosophical implica-

tions: there is only one target to be permuted, and doing this breaks all connections between target and descriptors while keeping the target distribution intact. By contrast, in Exceptional Model Mining, there are multiple targets. Swap-randomizing these targets can be done in two straightforward ways, whose philosophical implications are unclear. In Section 8.2.1, we outlined how each target column is permuted *independently* from the other target columns. This ensures that connections between targets and descriptors are broken, while keeping the *marginal* distribution of each target intact. However, the *joint* distribution over the targets is broken. This last effect can be prevented by the design choice to swap-randomize the targets not independently, but jointly: we generate only one permutation, and apply that same permutation to every target column.

The goal of EMM is to measure unusual interactions between several targets. Hence, on first glance, it seems preferable to maintain the joint distribution over the targets when swap-randomizing. However, more specifically, EMM strives to find *subgroups* coinciding with unusual target interactions, where these interactions are gauged in terms of some kind of *modeling* over the targets. The subgroups are based on coherent descriptions: conditions on a few descriptive attributes of the dataset. Depending on the model class under consideration, the descriptive attributes may be representing latent variables that actually should have been present (for instance as a dummy variable) in our model. In fact, the subsequent chapter concerns an application of representing found subgroups in a global model. In this light, it becomes unclear to assess whether it still makes sense to simultaneously *maintain* the internal connections between the targets, and *break* the connections between targets and descriptors. Though breaking the joint target distribution, swap-randomizing targets independently at least has clear philosophical implications. Further study is needed regarding both the theoretical foundations and empirical consequences of choosing one of these swap randomization variants.