2011 11th IEEE International Conference on Data Mining

Exploiting False Discoveries – Statistical Validation of Patterns and Quality Measures in Subgroup Discovery

Wouter Duivesteijn, Arno Knobbe
LIACS
Leiden University
The Netherlands
$\{$*wouterd,knobbe*$\}$*@liacs.nl*

***Abstract*—Subgroup discovery suffers from the multiple comparisons problem: we search through a large space, hence whenever we report a set of discoveries, this set will generally contain false discoveries. We propose a method to compare subgroups found through subgroup discovery with a statistical model we build for these false discoveries. We determine how much the subgroups we find deviate from the model, and hence statistically validate the found subgroups. Furthermore we propose to use this subgroup validation to objectively compare quality measures used in subgroup discovery, by determining how much the top subgroups we find with each measure deviate from the statistical model generated with that measure. We thus aim to determine how good individual measures are in selecting significant findings. We invoke our method to experimentally compare popular quality measures in several subgroup discovery settings.**

***Keywords*-Statistical validation, subgroup discovery**

I. Introduction

Subgroup discovery [1], [2] is a data mining framework (closely related to Contrast Set Mining [3] and Emerging Pattern Mining [4]; see also [5]) concerned with discovering subgroups that satisfy certain user-specified constraints. In this process, we explore a large search space to find subsets of the data that have a relatively high value for a given quality measure. Because of the magnitude of the candidate set, the process suffers from the multiple comparisons problem [6], 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 subgroup discovery is that it is nontrivial to determine whether discovered subgroups are actual discoveries, or *false discoveries* caused by random artifacts.

In this paper, we draw upon statistical theory to build a model for false discoveries. Using this model, a number of practical problems in subgroup discovery can be solved. When applying subgroup discovery 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 [7], [8] but also for non-standard variants such as regression [9], [10] and Exceptional Model Mining [11], [12], 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 an 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 subgroups 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 [6] 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 subgroup, that indicates to what extent the result is statistically significant.

As mentioned, subgroup discovery potentially suffers from the multiple comparison problem. The main contribution of this paper 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 subgroups. Furthermore, we employ the statistical validation to provide an experimental comparison of measures, and propose a suitable choice of measure.

The article is organized as follows. In Section II, we recall basic subgroup discovery and give a more formal description of our problems. In Section III, we introduce the method used to compare found subgroups with a model for false discoveries. Section IV discusses currently used

1550-4786/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDM.2011.65

validation approaches and other related work. Our method is empirically illustrated in Section V, after which we interpret the results in Section VI. Section VII concludes the paper with a summary.

II. Preliminaries

Throughout this paper we assume a dataset Ω with N elements (*data points*) that are $(h+1)$-dimensional vectors of the form $x = \{a_1, \ldots, a_h, \ell\}$. Hence, we can view our dataset as an $N \times (h+1)$ matrix, where each data point is stored as a row $x^i \in \Omega$. We call $a^i = \left\{a_1^i, \ldots, a_h^i\right\}$ the *attributes* of x^i, and ℓ^i its *target*. Attributes are taken from an unspecified domain $\mathcal{A}$. At this point, we will assume that ℓ is a single value from a discrete domain, which is the case in traditional subgroup discovery. In Section V-C we will explore other possible target domains.

For our definition of subgroups we need to define *patterns*. Usually a pattern is created by building a conjunction of constraints on singular attributes, and a data point is covered by the pattern if all constraints are satisfied. Hence a pattern is an intensional description of a part of our dataset, and its extension is the subgroup of records covered by the pattern. For practical reasons, in this paper we will technically define patterns to be functions $p : \mathcal{A} \to \{0, 1\}$. A pattern p *covers* a data point x^i if and only if $p\left(a^i\right) = 1$. We denote the space of all patterns by $\mathcal{P}$.

Definition (Subgroup). A *subgroup* corresponding to a pattern p is the bag of data points $G_p \subseteq \Omega$ that p covers:

$$G_p = \left\{x^i \in \Omega \;\middle|\; p\left(a^i\right) = 1\right\}$$

From now on we omit the p if no confusion can arise, and refer to a subgroup as G. We write n for the size of G.

We can now formally define a quality measure:

Definition (Quality measure). A *quality measure* is a function $\varphi : 2^{\Omega} \to \mathbb{R}$ which assigns to each subset of the data exactly one numeric value.

Notice that this definition may seem trivial. However, it implies that we allow any kind of quality measure. No special properties such as antimonotonicity are required; any function assigning a quality value to a subset of Ω is suitable.

As mentioned in the introduction, subgroup discovery [1], [2] is a data mining framework concerned with discovering subgroups that satisfy certain user-specified constraints. These constraints usually include a lower bound on the quality of a subgroup $\varphi\left(G_p\right) \geq t$, as well as a minimum support threshold $n \geq minsup$ that guarantees a lower bound on the size of the corresponding subgroup. Further constraints may involve properties such as the complexity of the generating pattern p. In most cases, a subgroup discovery algorithm will traverse a search lattice of candidate patterns in a top-down, general-to-specific fashion. The structure of the lattice is determined by a *refinement operator* $\rho : \mathcal{P} \to 2^{\mathcal{P}}$, a syntactic operation which determines how simple patterns can be extended into more complex ones by atomic additions. In our application, the refinement operator is assumed to be a *specialisation operator*: $\forall q \in \rho(p) : p \succeq q$ (p is more general than q).

The actual search strategy used to consider candidates is a parameter of the algorithm. We have chosen the *beam search* strategy [13], because it nicely balances the benefits of a greedy method with the implicit parallel search resulting from the beam. Beam search effectively performs a level-wise search that is guided by the quality measure φ. On each level, the best-ranking w patterns are refined to form the candidates for the next level. This means that although the search will be targeted, it is less likely to get stuck in a local optimum, because at each level alternatives are being considered. The search is further bounded by complexity constraints and the $minsup$ constraint. The end-result is a ranked list of patterns (each one corresponding to a subgroup) that satisfy the inductive constraints.

Notice that while we choose to traverse the search space heuristically using beam search, algorithms do exist that provably find the global top k subgroups without resorting to a heuristic, given certain restrictions. Such restrictions usually compel the attribute domain to be nominal [1], [2], or impose an anti-monotonicity property on the quality measure which is then used to prune the search space [14]. We choose to free this paper from such restrictions, but notice that this is by no means essential to the described method. It would also work with an exhaustive search setting. Regardless of restrictions, it may very well be the case that the global top k contains subgroups that cannot be statistically validated, especially if the search space is not too large. Even if a run of a subgroup discovery algorithm has the goal to find the top k subgroups, there may not be k statistically valid subgroups to report, and this may be of interest to the end-user. Merely finding the top k is not a justification of the siginificance of these k subgroups, hence this is not enough.

A. Problem statement

As mentioned in the introduction, the main contribution of this paper 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 two:

1) given a dataset Ω, a quality measure φ and a set $\mathcal{S}$ of subgroups found with this measure through subgroup discovery, determine the statistical significance of each element of $\mathcal{S}$;
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 k subgroups found with that measure from a random baseline, for a given k.

III. Validation method

In order to deal with the aforementioned multiple comparisons problem, we introduce a method consisting of the following steps: first we generate a random baseline model for false discoveries. Then we consider a discovered subgroup statistically sound if its measure value is significantly better than the random baseline. Finally, a quality measure is statistically sound if the best subgroups found with it are statistically sound. We can express this method more formally:

Suppose a dataset Ω, quality measures $\varphi_1, \ldots, \varphi_g$, and sets of subgroups $\mathcal{S}_1, \ldots, \mathcal{S}_g$ where $\forall_{i=1}^{g} : \mathcal{S}_i$ is found through subgroup discovery using quality measure φ_i.

I. $\forall_{i=1}^{g}$: use a randomization technique to generate baseline subsets $R_1^i, \ldots, R_m^i \subseteq \Omega$ for arbitrarily large m;

II. $\forall_{i=1}^{g}$: build a statistical model for false discoveries based on $\varphi_i\left(R_1^i\right), \ldots, \varphi_i\left(R_m^i\right)$. Then determine for each $S \in \mathcal{S}_i$ how much $\varphi_i(S)$ deviates from the model;

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

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 subgroups: 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 subgroups in $\mathcal{S}$, we merely consider for every single subgroup in $\mathcal{S}$ the likelihood that it is deemed interesting because of the multiple comparisons problem. The rest of this section investigates techniques for each step in the method separately.

A. Randomization techniques

There are several randomization techniques we can use to generate the baseline subsets $R_1^i, \ldots, R_m^i$ (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. Recently, Gionis et al. have published a paper detailing its use [15]. In its most radical form for zero-one matrices, swap randomization shuffles the elements of the data matrix in such a way that all row and column sums are kept intact, which is what the authors of [15] 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 subgroup discovery, it seems reasonable to use this moderate form of swap randomization.

We generate the baseline subsets in the following way. For each R_j^i to be generated, we create a swap-randomized version of the data, by keeping all attributes intact but taking a random permutation of the target column. Then we run our subgroup discovery algorithm on the resulting dataset using quality measure φ_i, and let R_j^i be the best subgroup found.

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

Another reason why a subgroup 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 subgroups on the original dataset. Alternatively one could easily choose a method to directly generate some random baseline subsets for use in step I of our method. However, a subgroup found on swap-randomized data goes through the same motions of the subgroup discovery algorithm as the actual subgroups 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 be reasonably considered a false discovery of the search process.

B. Building a statistical model

When we have generated the baseline subsets, 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) [16]. Under the assumption that m, the number of baseline subsets, is sufficiently large, according to the central limit theorem the mean of $\varphi_i\left(R_1^i\right), \ldots, \varphi_i\left(R_m^i\right)$ follows a normal distribution, since these are independent and identically distributed random variables. We use the sample mean and standard deviation as parameters for this distribution, as suggested by the method of moments [17]. We call this distribution the *Distribution of False Discoveries* (DFD). Let $S \in \mathcal{S}$ be a subgroup under consideration. We can now formulate the null hypothesis

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

We can compute a p-value corresponding to this null hypothesis for each $S \in \mathcal{S}$, 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 subgroup, 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 subgroup in the subgroup discovery 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.

C. 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 [18] on statistical comparisons of classifiers over multiple data sets. First a Friedman test [19], [20] 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; if 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 [21], 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 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.

IV. 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*, i.e. a pattern to be validated is assigned as p-value the fraction of randomly generated results that outperform the pattern. This method has been applied in articles concerning significant query results on multi-relational databases [22] and swap randomization on high-dimensional 0/1 datasets [15]. In many circumstances, the use of empirical p-values is very appropriate. However, we attempt to validate subgroups with a high quality by comparing them to random descriptions/subsets that are expected to have 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 [23]. 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 $(\sqrt{n}\,(p - p_0))$ [24] is order-equivalent to a t-test. Also well suited for subgroup discovery is the chi-squared $\left(\chi^2\right)$ measure [25], originally defined for association rules. While such quality measure automatically statistically validate single subgroups, 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 paper.

Tan et al. have developed a method [26] to compare quality measures on contigency 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 subgroups [27]. 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 attributes. 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 a method that assigns significance to subgroups, Webb's work is more a framework that can be used with

[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 of the formulae can be shown in four lines of math.

any statistical hypothesis test.

V. Experiments

To illustrate how our method performs, we experimented on several UCI datasets. For the subgroup discovery process we use the parameterized implementation available in the *Cortana* discovery package [28]. This Java implementation is an open-source spin-off of the Safarii Data Mining system.

Although we have so far stated our problems and method in terms of subgroup discovery with only one discrete target, this is by no means essential to the method. In fact, it can be applied to any local pattern discovery technique. We illustrate this by applying our method not only to traditional subgroup discovery, but also to an instance of Exceptional Model Mining (EMM) [12], an extension of subgroup discovery incorporating more complex target concepts. Whereas in traditional subgroup discovery there is only one target attribute ℓ, in EMM the target concept may consist of multiple attributes. EMM is instantiated by choosing a model class over these attributes, and defining a quality measure for this model class. For instance, one could be interested in finding conditions under which there is an exceptional ratio between two designated features. Then the chosen model class would be a simple linear regression model, and the quality measure would be based on the slope of the fitted regression line. This particular EMM instance would, given a dataset detailing the sales price of houses and their lot size, allow one to find conditions under which a set of houses has a relatively high price per square meter.

Both EMM and subgroup discovery are supervised settings, but nothing in the method requires this, so we could also apply it in unsupervised settings such as association discovery. We test our method on traditional subgroup discovery in Sections V-A and V-B, and on the EMM variant in Section V-C.

We pick the following parameters for the beam search process. On each level, we select the $w = 25$ best patterns, and refine these to create the candidate patterns for the next level. This beam width w creates a balance between redundancy in the reported patterns and search efficiency on the one hand, and search extensiveness on the other hand. When refining a pattern by adding a constraint on a numeric attribute, we partition the domain into eight equal-height intervals and consider inequalities on these dynamically allocated split points as the refining constraints. To bound the complexity of the patterns we use a search depth of $d = 3$ (at most three refinements). We let $minsup = \left\lfloor \frac{N}{10} \right\rfloor$, i.e. a pattern 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.

Notice that the dynamic partition of a numeric attribute into eight intervals uses more information from the attribute than the information we would use if we would statically discretize the attribute in eight intervals. Suppose for the sake of argument that we only refine candidate subgroups by adding a constraint on one particular numeric attribute. On search level one, the dynamic partition amounts to static discretization. On search level two, however, we start with candidate subgroups that have seven different sizes (the eighth possibility is not a proper refinement). Each of these is refined by dynamically partitioning into eight intervals, leading to 49 different possibilities to start with on search level three, etcetera. Hence the used information is much more than the information used when statically discretizing.

Table I
UCI data sets used for the experiments.

	Dataset	N	# attributes		$\lvert\ell\rvert$
			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

The 20 datasets we have used for our tests with traditional subgroup discovery, in the following two sections, can be found in the UCI Machine Learning Repository [29]. Table I contains details on the datasets considered. Here, $|\ell|$ denotes the number of distinct target values in the dataset.

Before experimenting with the method, let us shortly substantiate the claim made in the previous section, that empirical p-values are not very suitable in our setting. Figure 1 displays a histogram (represented by the twitchy line) 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 model fitted to the random qualities (represented by the smooth curve) and the 13 subgroups (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 subgroups 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.

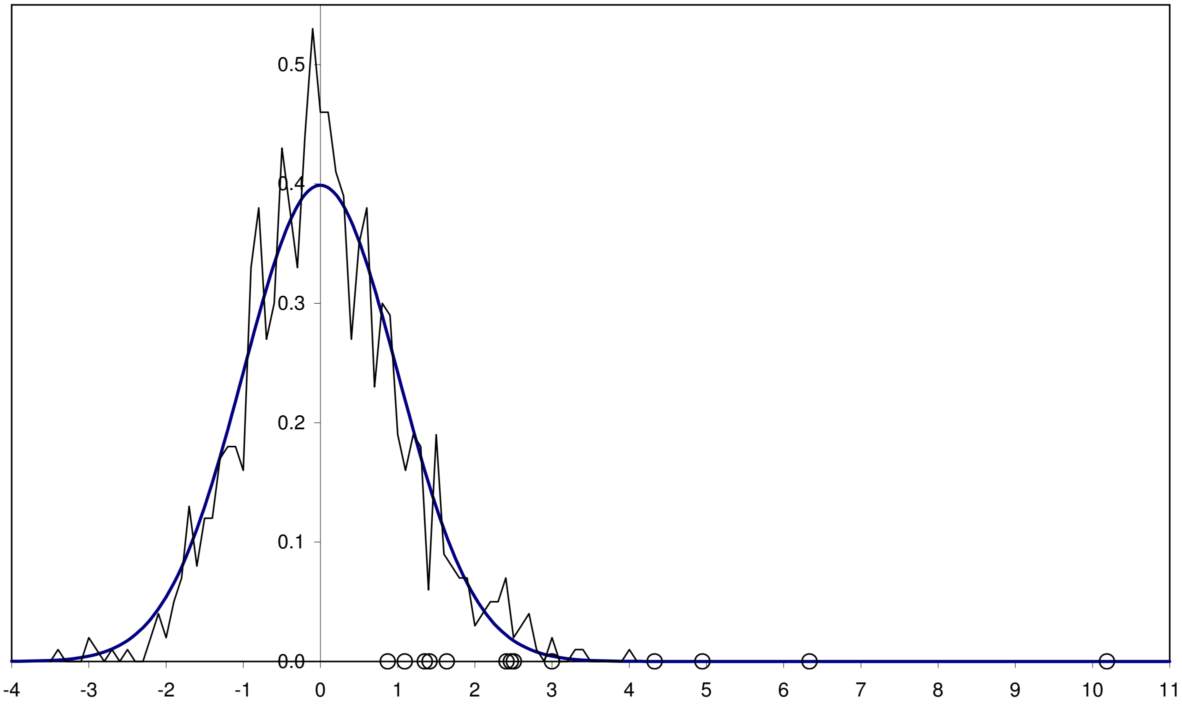

Figure 1. CLT-based model versus empirical p-values.

A. Validating subgroups

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

We consider just one quality measure: Weighted Relative Accuracy (WRAcc) [24], arguably the most popular quality measure in subgroup discovery. For each dataset, we perform a subgroup discovery run for each target value, and report the 1000 best subgroups. We then run the first two steps of our method to determine how many of these subgroups remain if insignificant subgroups are removed. We report the average fraction of subgroups that is retained per dataset for different significance levels in Table II.

As stated in Section III, one could also use the Distribution of False Discoveries to determine quality measure thresholds for given significance levels, a common practical issue with subgroup discovery 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. For illustration, the best subgroup found on this dataset with this target value has a WRAcc of 0.188.

Table II
FRACTION OF SUBGROUPS RETAINED WHEN REMOVING INSIGNIFICANT SUBGROUPS.

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

B. 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 II-A: 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,

Table III
Average ranks of the quality measures.

	All datasets		Binary datasets	
Measure	$k = 1$	$k = 100$	$k = 1$	$k = 100$
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

Table IV
Average ranks for correlation model measures.

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

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 [7].

For each dataset, we perform steps I and II of our method 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 k best subgroups, for both $k = 1$ and $k = 100$ (for $k = 100$ we take the average p-values over the top 100 groups). Hence for all measures we obtain for both choices of k one test score for each combination of dataset and target value within that dataset. For $k = 1$ 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 subgroups are found that satisfy the *minsup* constraint. Hence there are no results on these datasets for $k = 100$, leaving a total of 78 test cases for $k = 100$.

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 III. 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 good. Hence we comfortably pass this test.

Since the Friedman test is passed, we can now perform Nemenyi tests to see which quality measures outperform others. For the $k = 1$ setting, the critical difference CD equals 2.069 with significance level $\alpha = 1\%$. For each pair of measures we compute from Table III 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 [18] can be found in Figure 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 2 we see that χ^2 is significantly better than Laplace, Specificity, Purity, and Sensitivity. Figure 3 displays the CD chart for the $k = 100$ 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 III. 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 $\alpha = 1\%$.

C. Beyond subgroup discovery

So far we have illustrated our method with measures for subgroup discovery over a single discrete target. We now turn to a variant of EMM [11], [12], an extension of subgroup discovery incorporating complex target concepts. This variant strives to find subgroups for which the correlation between two attributes is significantly different from their correlation on the whole dataset. Several quality measures have been proposed for this problem [12]. We validate measures for this setting on the datasets and target concepts used in the original paper. The resulting average ranks over the two datasets — Windsor Housing and Gene Expression — can be found in Table IV. 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 $k = 1$ and $k = 100$). In these modest experiments we find that no significant conclusions can be drawn.

VI. Discussion

The previous section displayed 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 subgroups in a set $\mathcal{S}$ found through subgroup discovery.

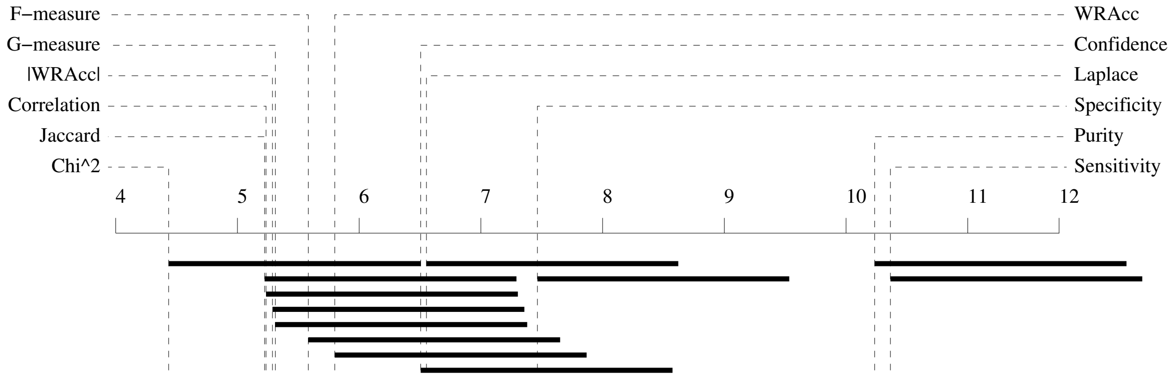

Figure 2. CD chart for $k = 1$ ($CD = 2.069$).

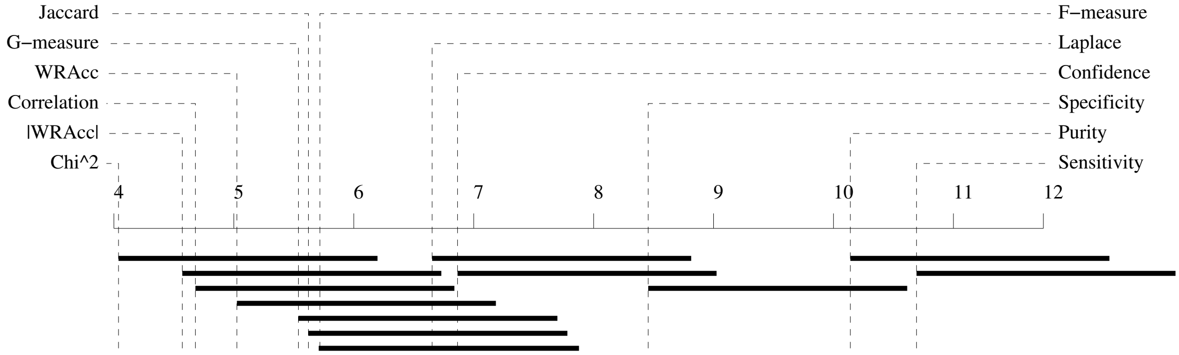

Figure 3. CD chart for $k = 100$ ($CD = 2.160$).

A. Validating subgroups

From Table II we find that we cannot refute any subgroups from $\mathcal{S}$ in several datasets: Adult, Credit-a, Ionosphere, Pima-indians, and Wisconsin. To explain this result, we crafted a metalearning dataset from Tables I and II. We selected the columns from Table I as attributes of our metalearning dataset, and added three new columns, representing the total number of attributes in the dataset, a boolean column representing whether the dataset has discrete attributes, and a boolean column representing whether the dataset has numeric attributes. As target column we added the last column of Table II: the fraction of subgroups retained when insignificant subgroups are removed, with significance level $\alpha = 1\%$. On this metalearning dataset we performed a shallow (using search depth $d = 1$) but exhaustive subgroup discovery run, using Klösgen's mean test as quality measure. The resulting metasubgroups should consist of those datasets with a relatively high fraction of kept subgroups.

The best metasubgroup is defined by the condition that the datasets have more than five numeric attributes. The eight datasets belonging to this metasubgroup are Adult, Credit-a, Glass, Ionosphere, Labor, Pima-indians, Wisconsin, and Yeast. This set includes all datasets for which we cannot refute any of the subgroups from $\mathcal{S}$. This makes sense, since for each dataset we have only considered the top 1000 subgroups, a fixed number independent of dataset characteristics. Numeric attributes usually have many different values, resulting in a hypothesis space that is much larger than it would have been if the attributes would have been discrete. Hence in datasets with relatively many numeric attributes, it is more likely that the 1000 best subgroups 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.

B. 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 k subgroups from false discoveries. Many of the relations between quality measures, however, are not significant. For $k = 1$, 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 $k = 100$, we see some slight changes: χ^2 and

$|\text{WRAcc}|$ 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 k subgroups from false discoveries when $k = 1$, while others might be better when $k = 100$. The observed changes are not very dramatic, and we consider the selection of k a user-derived parameter in the method.

One of the significant relations that seems somewhat peculiar, is the result that for both $k = 1$ and $k = 100$, Confidence performs significantly better than Purity, while the latter is defined to be $\max\{\text{Confidence}, 1-\text{Confidence}\}$. While there is a good theoretical reason to consider the Purity of a subgroup, 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 subgroups 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 III with the last two columns, we can see that $|\text{WRAcc}|$, 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 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 differences between the quality measures can be expected when tested on more datasets with a binary target.

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. Expensive 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.

VII. Conclusions

We propose a method that deals with the multiple comparisons problem in subgroup discovery, 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 a subgroup discovery algorithm on a swap-randomized version of the data. In this swap-randomized version, while the distribution of the target variable is maintained, its correlation with the attributes is destroyed. Hence the best subgroup 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 subgroup discovery. For any given discovered subgroup, 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 subgroup 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 k subgroups 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 subgroups, we see that our method removes insignificant subgroups found on datasets that have few numeric attributes. From metalearning we find that on large datasets, for instance with more than five numeric attributes, the random baseline is more likely to accept many patterns. This is reasonable because of the associated larger hypothesis space. Table II shows that our method can remove insiginficant subgroups on some of the datasets with more than five numeric attributes, 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 paper 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 paper 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 a subgroup discovery run one wishes to validate, one has to run the same subgroup discovery algorithm an additional m times, where m 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 [11], this becomes problematic. When computation time becomes an issue, one might consider different randomization techniques to generate $R_1, \ldots, R_m$, for instance by simply drawing a random sample from Ω of a certain size for each R_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.

Acknowledgments

This research is financially supported by the Netherlands Organisation for Scientific Research (NWO) under project number 612.065.822 (Exceptional Model Mining).

References

[1] J. Friedman, N. Fisher, Bump-Hunting in High-Dimensional Data, Statistics and Computing 9(2), pp. 123–143, 1999.

[2] W. Klösgen, Subgroup Discovery, Handbook of Data Mining and Knowledge Discovery, ch. 16.3, Oxford University Press, New York, 2002.

[3] S. D. Bay, M. J. Pazzani, Detecting group differences: Mining contrast sets, Data Mining and Knowledge Discovery, 5(2), pp. 213–246, 2001.

[4] G. Dong, J. Li, Efficient mining of emerging patterns: Discovering trends and differences, Proc. KDD, pp. 43–52, 1999.

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

[6] Y. Hochberg, A. Tamhane, Multiple Comparison Procedures, Wiley, New York, 1987.

[7] J. Fürnkranz, P. A. Flach, ROC 'n' Rule Learning – Towards a Better Understanding of Covering Algorithms, Machine Learning 58 (1), pp. 39-77, 2005.

[8] N. Lavrač, P. Flach, B. Kavšek and L. Todorovski, Rule induction for subgroup discovery with CN2-SD, Proc. ECML/PKDD, 2002.

[9] H. Grosskreutz, Cascaded subgroups discovery with an application to regression, Proc. ECML/PKDD, 2008.

[10] B. F. I. Pieters, A. Knobbe, S. Džeroski, Subgroup Discovery in Ranked Data, with an Application to Gene Set Enrichment, Proc. Preference Learning workshop (PL2010) at ECML PKDD, 2010.

[11] W. Duivesteijn, A. Knobbe, A. Feelders, M. van Leeuwen, Subgroup Discovery meets Bayesian networks - an Exceptional Model Mining approach, Proc. ICDM, pp. 158–167, 2010.

[12] D. Leman, A. Feelders, A. Knobbe, Exceptional Model Mining, Proc. ECML/PKDD, Part II, pp. 1–16, 2008.

[13] Y. H. Xu, A. Fern, On Learning Linear Ranking Functions for Beam Search, Proc. ICML 2007, ACM International Conference Proceeding Series vol. 227, pp. 1047–1054, ACM, New York.

[14] H. Grosskreutz, S. Rüping, On Subgroup Discovery in Numerical Domains, Data Mining and Knowledge Discovery 19(2), pp. 210–226, 2009.

[15] A. Gionis, H. Mannila, T. Mielikäinen, P. Tsarapas, Assessing data mining results via swap randomization, Proc. KDD, pp. 167–176 , 2006.

[16] A. M. Lyapunov, Nouvelle forme du théorème sur la limite de probabilité, St. Petersburg, 1901.

[17] K. Pearson, L. Filon, Mathematical contributions to the theory of evolution, iv. on the probable errors of frequency constants and on the influence of random selection on variation and correlation, Phil. Trans. A. 191, pp. 229–311, 1898.

[18] J. Demšar, Statistical Comparisons of Classifiers over Multiple Data Sets, Journal of Machine Learning Research 7, pp. 1–30, 2006.

[19] M. Friedman, The use of ranks to avoid the assumption of normality implicit in the analysis of variance, Journal of the American Statistical Association 32, pp. 675–701, 1937.

[20] M. Friedman, A comparison of alternative tests of significance for the problem of m rankings, Annals of Mathematical Statistics 11, pp. 86–92, 1940.

[21] P. B. Nemenyi, Distribution-free multiple comparisons, PhD thesis, Princeton University, 1963.

[22] M. Ojala, G. C. Garriga, A. Gionis, H. Mannila, Evaluating Query Result Significance in Databases via Randomizations, Proc. SDM, pp. 906–917, 2010.

[23] N. Megiddo, R. Srikant, Discovering Predictive Association Rules, Proc. KDD, pp. 274–278, 1998.

[24] W. Klösgen, Explora: A multipattern and multistrategy discovery assistent, Advances in Knowledge Discovery and Data Mining, pp. 249-271, 1996.

[25] C. Silverstein, S. Brin, R. Motwani, Beyond Market Baskets: Generalizing Association Rules to Dependence Rules, Data Min. Knowl. Discov. 2 (1), pp. 39–68, 1998.

[26] P.-N. Tan, V. Kumar, J. Srivastava, Selecting the right interestingness measure for association patterns, Proc. KDD, pp. 32–41, 2002.

[27] G. I. Webb, Discovering Significant Patterns, Machine Learning 68 (1), pp. 1–33, 2007.

[28] M. Meeng, A. J. Knobbe, Flexible Enrichment with Cortana – Software Demo, Proc. Benelearn, pp. 117–119, 2011.

[29] D. Newman, S. Hettich, C. Blake, C. Merz, UCI repository of machine learning databases, 1998.