Subgroup Discovery for Test Selection: A Novel Approach and Its Application to Breast Cancer Diagnosis

Marianne Mueller[1], Rómer Rosales[2], Harald Steck[2],
Sriram Krishnan[2], Bharat Rao[2], and Stefan Kramer[1]

[1] Technische Universität München, Institut für Informatik, 85748 Garching, Germany
[2] IKM CAD and Knowledge Solutions, Siemens Healthcare, Malvern PA 19335, USA

Abstract. We propose a new approach to test selection based on the discovery of subgroups of patients sharing the same optimal test, and present its application to breast cancer diagnosis. Subgroups are defined in terms of background information about the patient. We automatically determine the best t subgroups a patient belongs to, and decide for the test proposed by their majority. We introduce the concept of prediction quality to measure how accurate the test outcome is regarding the disease status. The quality of a subgroup is then the best mean prediction quality of its members (choosing the same test for all). Incorporating the quality computation in the search heuristic enables a significant reduction of the search space. In experiments on breast cancer diagnosis data we showed that it is faster than the baseline algorithm APRIORI-SD while preserving its accuracy.

1 Introduction

Diagnosis is the art or act of identifying a disease from its signs and symptoms. This implies that the more information is available about a patient, the easier it is to pose an accurate diagnosis. Information can be obtained by a variety of tests including questioning the patient, physical examinations, imaging modalities, or laboratory tests. However, due to costs, time, and risks for the patient, in clinical routine it is often preferable for patients to undergo as few tests as needed. Consequently, there is a trade-off between the costs (and number) of tests and the accuracy of the diagnosis. Therefore, optimal test selection plays a key role for diagnosis. The goal of this paper is to find the optimal set of tests to choose for a patient in a given situation, where the definition of optimality is also provided in this paper. Existing work on test selection [1,2] mostly addresses the problem of finding global solutions for all patients. However, it is not likely that for each patient the same test is the most informative one. Therefore, we believe that it is a better approach to concentrate on the task of identifying subgroups of patients for which the optimal test is the same. In this paper, we present a novel solution to this problem based on subgroup discovery (SD) [3,4], a family of data mining algorithms. Subgroup discovery methods compute all subgroups of a population

that are statistically most interesting with respect to a specified property of interest. Consider, for instance, a population described by general demographic attributes and a target variable (property/attribute of interest) representing a disease status (disease, non-disease). Lets assume a distribution of 43% disease and 57% non-disease in the entire population. Then, an SD algorithm might come up with a subgroup identified by two conditions, $age > 75$ and $gender = female$ in which the distribution is 85% disease and 15% non-disease. Here, the subgroup description consists of two attribute-value tests, and it selects a set of persons with a particularly high prevalence of the disease (85% instead of 43% in the entire population). Standard SD approaches are designed for a single target variable. However, in the setting of test selection, a single variable seems not sufficient. In fact, we want to target the relation between two variables: the outcome of a test and the actual state of disease. Therefore, the quality of a subgroup should correspond to the value of the result of a selected test with respect to the actual state of disease, e.g., a biopsy result. To quantify the value of a test result, we define a so-called *prediction quality* function in Section 2.1. The function gives high scores to a pair of a subgroup and a test if the result is close to the actual state of disease, and therefore leads to an accurate diagnosis. Since standard SD does not take into account complex scenarios like this, including benefits or costs of subgroups, we developed a new, cost-sensitive variant.Throughout the paper, we will use the term *prediction quality*, which corresponds to the *benefits* of a prediction rather than to its *costs*. However, as it is easy to transform one into the other, we can also speak of *cost-sensitive subgroup discovery*. The algorithm outputs subgroup descriptions consisting of background information about the patients. The overall goal is to compute an optimal test selection for a new patient. More precisely, our proposed solution is to identify subgroups of the data for which the same test is the optimal selection, to arrive at a correct diagnosis. In a second step, analyzing the subgroups will help to find out which features determine the performance of the tests. Hence, it will be possible to decide for a new patient, given its features, which test is the best to choose. We apply and validate this approach on a data set from breast cancer diagnosis, where for each patient four different tests are possible.

2   Background and Data

Our study was conducted in the area of breast cancer diagnosis. In breast cancer diagnosis, different imaging modalities are used routinely, in particular, Film Mammography (FMAM), Digital Mammography (DMAM), Ultrasound (USND), and Magnetic Resonance Imaging (MRI). Each modality has its own specific characteristics. When a patient is under scrutiny for breast cancer, it is often not clear which of these modalities is best suited to answer the basic question to whether the patient has or does not have cancer. The choice of a modality usually requires considerable experience of the health care workers. In this paper we show how to support the optimal test selection for a new patient

Table 1. Prediction score for agreement between the overall assessment ($OA_m =$ BIRADS) of a modality m and the biopsy finding BIO

pscr		OA_m					
		0	1	2	3	4	5
BIO	Malignant	75	0	0	25	100	100
	Atypia	75	75	90	90	90	75
	Benign	75	100	100	100	75	50

by retrospectively analyzing the performance of the tests on subgroups of previously examined patients with similar features. The basis of our work is a dataset collected in a breast cancer study of a large University Hospital, which comprises patients that had a suspicious finding in a screening. The study gathers patient specific information like medical history, demographic information, and a breast cancer risk summary. Each patient in the study underwent all four above mentioned modality tests. Each of these tests was independently analyzed by the appropriate specialist to judge for the occurrence of breast cancer. For each lesion detected, the specialist determines in which category it falls. The categories are called BIRADS score and range from 0 to 5: The higher the BIRADS, the higher the probability (assessed by the medical expert) for the lesion to be malignant. (0 = incomplete, i.e., needs additional imaging evaluation, 1 = no finding, 2 = benign finding, 3 = probably benign, 4 = suspicious abnormality, 5 = highly suggestive of malignancy) [5]. To obtain evidence of the initial assessments, a biopsy has to be performed. A pathologic examination of a biopsy determines whether the lesion is benign, atypia benign, or malignant. In this study at least one lesion for each patient is subject to biopsy.

2.1 Definition of Prediction Quality

To quantify the accuracy of a diagnosis, we propose a measure of prediction quality. Each test m results for each lesion l in an overall assessment $OA_m(l)$ of the physician. This is defined by the BIRADS score (see above). Each lesion has a biopsy $BIO(l)$ proving the status of the lesion (malignant, benign or atypia benign). The prediction quality expresses how close the assessment comes to the biopsy finding. Therefore, we define a prediction score *pscr* that evaluates the performance of a test for a single lesion. Table 1 gives the *pscr* for each pair (Overall Assessment, Biopsy). The values in the table were proposed by a domain expert in the field of breast cancer diagnosis.[1] The higher the prediction score, the more accurate is the prediction.

Having defined *pscr* for a single lesion l, we can easily obtain the prediction quality $pq(S, m)$ of a modality m for an example set S by averaging over the prediction scores of m and all lesions in S:

---

[1] However, they are certainly not the only possible values for the prediction score. For instance, as not all types of malignant findings are equally harmful, it might be more accurate to distinguish between invasive and non-invasive types of cancer.

$$pq(S, m) = \frac{1}{|S|} \sum_{l \in S} \cdot pscr(OA_m(l), BIO(l))$$

In our data set we have 138 lesions (of 72 patients) with biopsy and four modalities (Digital Mammography (DMAM), Film Mammography (FMAM), Magnet Resonance Imaging (MRI), and Ultrasound (USND)) to choose from. The prediction quality for the entire dataset separated for each modality is 77.9 for DMAM, 78.0 for FMAM, 78.4 for MRI, and 80.2 for USND. It shows that the prediction qualities of the different modalities over all lesions are quite similar (Entropy = 1.999 bits of a maximum 2 bits), with USND performing slightly better. By considering subgroups of patients we expect to increase the prediction quality for at least one modality per subgroup. Then, we apply this modality to all lesions in the subgroup to obtain the most accurate diagnosis.

3   Method

The general idea is to determine subgroups of lesions with an unusual modality performance. Let X be a training set of observed examples and n the number of tests $\{m_1, \ldots, m_n\}$ that can be performed. For each group[2] of lesions $S \subseteq X$ we consider the prediction qualities $pq(S, m_i)$ of the possible modalities and decide for the modality $m^*(S)$ with the highest pq-value[3]: $m^*(S) = \mathrm{argmax}_m\, pq(S, m)$. The optimal prediction quality of S is then defined as $pq^*(S) = \max_m pq(S, m)$.

We introduce an algorithm called $SD4TS$ (Subgroup Discovery for Test Selection). The task of the algorithm is defined in the following way:

Given: X, n, $minsupport$, the minimal number of examples that have to be covered by a subgroup description, t, the number of best subgroups we want to obtain from the algorithm, and a set of pq-values $\{pscr(s, m_i)|s \in X, m_i \in tests\}$ (in a more general setting a set of cost/benefit values).

Find: The t subgroups with the highest pq^* values (best costs/benefit) and at least $minsupport$ examples.

We base our algorithm on APRIORI-SD [3], an adaptation of the association rule learning algorithm APRIORI [6] to subgroup discovery. APRIORI-SD starts with generating subgroups described by a single attribute-value-pair. Subsequently, it generates subgroups with longer (and thus more specific) descriptions. Subgroups are only kept if they contain more examples than $minsupport$. All smaller subgroups are pruned, and no subgroups more specific than these are generated. For our task, we are interested in the t subgroups that are cost-efficient for at least one modality. Therefore, we can prune the search space even further, namely in a way that only the promising subgroups are kept. That means, during

---

[2] As in other SD algorithms we consider only groups that can be described by a conjunction of attribute-value pairs.

[3] In a more general setting, instead of pq we can assume any types of costs (where max should be replaced by min) or benefits that rate the performance of the tests.

the generation of subgroups, candidates are immediately evaluated and checked whether they have the potential to lead to improved costs.

3.1   Quality Pruning

Pruning is possible when a subgroup and all specializations of the subgroup will not outperform the quality of the already discovered subgroups. Specialization of a subgroup means adding an attribute-value pair to the subgroup description of a subgroup sg. This can cause changes of both the frequency and the quality. The frequency can only decrease. The defined quality, however, can change in both directions.

The critical point is hence to recognize when the quality of subgroup sg can not outperform at least one of the best t subgroups found so far. Thus, it is not enough to consider the actual $pq(sg)$ to determine if sg can be pruned. Furthermore, it is necessary to consider what we call the *coreGroup* of sg. The *coreGroup* is a group consisting of the *minsupport* examples covered by sg with the highest quality. The cost of the *coreGroup* upperbounds the costs of all possible specializations of sg, because the overall score is defined as an average of the elements of the group.

The example in Figure 1 demonstrates the discussed characteristics. The seven dots represent the generated subgroup sg, with $pq(sg) = 0.5$. Assume we have generated already a subgroup sg_{best} with $pq(sg_{best}) = 0.6$. In this case, sg has a worse pq value and seems to be not promising. However, pruning sg will inhibit finding an optimal subgroup sg_{new} (the four black dots) contained in sg with $pq(sg_{new}) = 0.625$.

Considering the pq-value of the *coreGroup* of sg will circumvent this mistake by providing the upper bound of the pq-values of any specialization of sg: For the given example, we assume a *minsupport* of 3. Then $pq(coreGroup(sg)) = 1$. Since $pq(coreGroup(sg)) > pq(sg_{best})$, sg is not pruned and keeps the option of generating the improved subgroup sg_{new} in a later iteration of the algorithm.

3.2   The $SD4TS$ Algorithm

The pseudo-code of the algorithm is shown in Algorithm 1. $SD4TS$ starts with generating 1-itemset candidates (described by one attribute-value-pair).
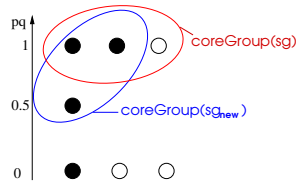

Fig. 1. Simple pruning fails. All dots are examples in sg. The black dots are also covered by sg_{new}. The y-direction corresponds to the pq-value of each example.

Candidates are pruned if they are not frequent or if they can not outperform (not even through specialization) one of the the best t subgroups generated so far. All remaining candidates are stored in *optimizableCandidates*. The best t subgroups are stored in *topCandidates*. For efficiency, we store all created subgroups (including the list of transactions that are covered, costs, *bestPossibleCosts*, support and a list *still2test* of 1-itemsets that have not been tested yet to specialize the subgroup) in an array *allSubgroups* in the order they were created. The sorted lists *topCandidates* and *optimizableCandidates* contain only pointers (the indices of the array) to the subgroups stored in *allSubgroups*. The list *topCandidates* is sorted according to the actual costs of the subgroups. This facilitates removing the worst subgroup, whenever a newly generated subgroup has better costs. The list *optimizableCandidates* is sorted according to the *bestPossibleCosts* a *coreGroup* can achieve. In that way, we explore always the subgroup with the highest potential first. That means specializing this subgroup

---

Algorithm 1. $SD4TS$ (subgroup discovery for test selection)

Input: Training set, set of costs, *minsupport*, number t of best subgroups to be produced

Output: list of *topCandidates*, including the proposed test(s) for each candidate

---

1: $optimizableCandidates = \{c | c$ frequent subgroup defined by 1 attribute-value-pair$\}$;
2: $topCandidates = \{c | c$ belongs to the t best candidates in *optimizableCandidates*$\}$;
3: $minpq$ = worst quality of *topCandidates*;
4: remove all c from *optimizableCandidates* with $c.bestPossibleCosts < minpq$
5: **while** *optimizableCandidates* not empty **do**
6:      $c_1 = optimizableCandidates$.removeFirst();//candidate with bestPossibleCosts
7:      **for all** $c_2 \in c_1.still2test$ **do**
8:          remove c_2 from $c_1.still2test$
9:          c_{new} = generate_new_candidates(c_1, c_2);
10:         **if** c_{new} frequent and $c_{new}.bestPossibleCosts > minpq$ **then**
11:             add c_{new} to *optimizableCandidates*
12:             **if** c_{new} better than worst c_t of *topCandidates* **then**
13:                 add c_{new} to *topCandidates*
14:                 **if** $size(topCandidates) > t$ **then**
15:                     remove worst c of *topCandidates*
16:                 **end if**
17:                 $minpq$ = worst quality of *topCandidates*;
18:                 remove all c from *optimizableCandidates* with $c.bestPossibleCosts < minpq$
19:             **end if**
20:         **end if**
21:     **end for**
22: **end while**
23: **return** *topCandidates*

---

is likely to lead to a subgroup that falls into the top t candidates and therefore raises $minpq$ which reduces the search space.

Safe pruning: A new subgroup candidate sg_{new} is only accepted if sg_{new} is frequent and at least one of the following holds:

1. there are less than t subgroups stored in $topCandidates$, or
2. sg_{new} has a better performance than the worst subgroup of $topCandidates$, or
3. at least one frequent subset of examples in sg_{new} (i.e., $coreGroup(sg_{new})$) leads to a better performance than the worst subgroup of $topCandidates$. This can be tested in $O(n * |sg_{new}| \log |sg_{new}|)$: For each test m determine the set CG_m of $minsupport$ examples covered by sg_{new} that have the best costs. $sg_{new}.BestPossibleCosts = \max_{m} pq(CG_m, m)$

In all cases we add the candidate to $optimizableCandidates$. In case 1 and 2 we also add the candidate to $topCandidates$. In the second case we additionally remove the worst stored subgroup from $topCandidates$.

3.3   Analysis of Runtime and Search Space

Figure 2 shows how the search space (i.e., the number of search nodes) depends on the parameters $minsupport$ and t. The higher $minsupport$, the smaller the search space. This is caused by the frequency pruning. We also see that a low t-value results in a small search space, which is the expected effect of quality pruning. For small values of t fewer subgroups are kept in $topCandidates$, which increases the threshold of costs below which subgroups are pruned. The right diagram in Figure 2 displays the runtime of the two algorithms. For $minsupport$ values below 25, $SD4TS$ is faster than APRIORI-SD, as frequency pruning is only effective for larger $minsupport$ values.
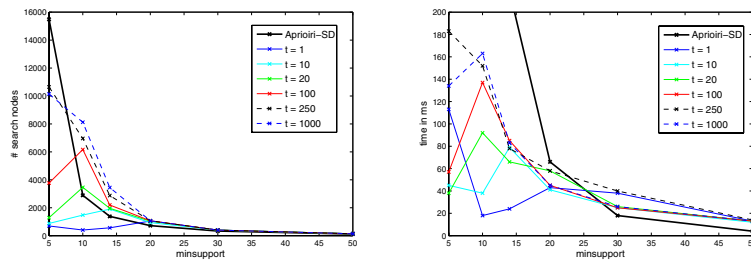

Fig. 2. Complexity of $SD4TS$. The left (right) diagram shows how the search space (runtime) depends on $minsupport$ (x-axis) for different t-values. In comparison, the black solid line shows the search space of APRIORI-SD.

4   Validation and Results

To evaluate the approach, we tested it in a predictive setting[4], more specifically, in a leave-one-out cross-validation. For each test lesion l, we generate only subgroups with attribute-value pairs contained in l. Table 2 shows the best $t = 5$ subgroups for 3 example lesions. From the resulting best t subgroups, we decide for the test proposed by the majority of the identified subgroups (for test lesion 9 it is USND). A test is proposed if it has the best costs averaged over all examples in the subgroup (for subgroup S1 it is USND). If more than one test has optimal costs, all of them are proposed (for subgroup S9 it is DMAM and USND). If more than one test is proposed most often by the subgroups, the cost for the test lesion l is determined by the mean of their costs.

4.1   Analysis of Performance Compared to Random Selection

For each lesion l, a vector of prediction qualities is given by

$$\overrightarrow{pq}(l) = (pq(l, FMAM), pq(l, DMAM), pq(l, MRI), pq(l, USND)).$$

This can be interpreted as a ranking of the modalities for each lesion. For instance, $\overrightarrow{pq}(l) = (0.8, 0.7, 0.9, 0.8)$ leads to the ranking $MRI > USND = DMAM > FMAM$. We encode this ranking as 1224. There is one modality ranked first, followed by two modalities ranked second, and one modality ranked fourth. In total, there are seven possible encodings: from 1111 (all modalities have the same prediction quality) to 1234 (all modalities have different prediction qualities).

Table 3 shows the distribution of codes of our dataset. It is remarkable that in 29% of the cases all modalities perform equally well. This implies that for those cases a random choice is as effective as a more informed choice. To have fairer conditions, we additionally validated the algorithm on two restricted subsets of test lesions (results are shown in Table 6). Set 1 consists of all 138 lesions. Set 2 is a subset of Set 1 containing only the 98 lesions whose costs are not the same over all modalities (all codes except 1111). Set 3 comprises 32 lesions, where one modality outperforms the other modalities (code 1222, 1224, and 1234). Note that the differences between the best and the worst, and between the best and the random choice improve significantly from Set 1 to Set 3.

4.2   Results

The results in Table 4 (column Set 1) show that the algorithm achieves in general better costs than picking a modality at random or picking always the same modality (compare with Table 3). It also becomes clear that the best results are achieved with low *minsupport* and high t values (considering even small subgroups), or, vice versa, high *minsupport* (50) and low t values (few large subgroups).

---

[4] Note that prediction is not our main goal. Additionally, we are interested in the discovery of new medical knowledge. Therefore, we prefer subgroup discovery over standard classifiers.

Table 2. Example of best 5 subgroups for 3 selected input test lesions. The shaded rows indicate the actual prediction scores of the modalities for the current input test lesion. The bold prediction qualities indicate the image modality proposed by $SD4TS$. For example, test lesion 9 will be assessed best by USND (pscr =100), the other three modalities fail to assess the lesion correctly (pscr = 0).

	top 5 subgroups for selected test lesions	subgr. size	DMAM	FMAM	MRI	USND
	test lesion 9		0	0	0	**100**
S1	Highschool or less + has past Mammo	17	76.5	76.5	57.4	**98.5**
S2	Highschool or less + has past Mammo + no relatives with cancer	16	75.0	75.0	56.3	**98.4**
S3	Highschool or less + has past Mammo + has past breast USND	14	85.7	78.6	62.5	**98.2**
S4	Highschool or less + has past Mammo + Race = white	14	85.7	78.6	66.1	**98.2**
S5	age 40-59 + no relatives with cancer + pre menopausal + Race=white	28	76.8	72.3	76.8	**93.8**
	test lesion 19		75	100	100	**100**
S6	Graduate School after college + age 40-59 + no relatives with cancer + Race=white	14	76.8	75.0	**98.2**	82.1
S7	Graduate School after college + has no breast USND + no relatives with cancer	21	94.1	82.1	92.9	**96.4**
S8	Graduate School after college + has no breast USND + no relatives with cancer + Race = white	19	93.4	80.3	92.1	**96.1**
S9	Graduate School after college + age 40-59 + no relatives with cancer + has no breast USND	18	**95.8**	81.9	91.7	**95.8**
S10	Graduate School after college + has no breast USND+ no relatives with cancer + Race = white + age 40-59	16	**95.3**	79.7	90.6	**95.3**
	test lesion 23		**100**	100	75	100
S11	no relatives with cancer + age ≥60	14	**94.6**	87.5	87.5	76.8
S12	Graduated from College + post menopausal status	16	78.1	82.8	**93.8**	70.3
S13	post menopausal status + age ≥ 60	15	**93.3**	86.7	86.7	76.7
S14	age ≥60	15	**93.3**	86.7	86.7	76.7
S15	Graduated form College + no relatives with cancer + post menopausal status	15	76.7	81.7	**93.3**	70.0

Table 3. Left: Distribution of lesions according to the ranking of modality performances. Right side shows the costs that arise for always picking the same modality (separated in always picking DMAM, always picking FMAM, etc), or for always picking the modality with the best costs, the worst costs, or picking one modality at random. The choice of parameters is good, if the costs are greater than the costs for random selection.

# best modalities	# cases	code	# cases		
4 or 0	40 (29%)	1111	40		
3	45 (33%)	1114	45		Set 1
2	21 (18%)	1133	16		
		1134	5	Set 2	
1	32 (23%)	1222	27		
		1224	3	Set 3	
		1234	2		

costs	Set 1	Set 2	Set 3
Best	99.45	99.49	99.22
Random	**78.6**	**70.22**	**40.63**
Worst	58.03	41.33	16.41
DMAM	77.92	69.13	30.47
FMAM	78.1	69.39	41.41
MRI	78.28	69.9	44.53
USND	80.11	72.45	46.09

Table 4. Results of leave-one-out cross-validation of $SD4TS$ with varying minsupport and t parameters over three different test sets (best parameter settings in **bold**). The displayed costs are averaged over all test lesions. A cost for a single lesion is derived by taking the costs of the test proposed by the majority of the returned subgroups. If two or more tests are proposed equally often, we take the average of their costs.

|	costs	Set 1					Set 2					Set 3				
	min-s.	5	10	20	30	50	5	10	20	30	50	5	10	20	30	50
	1	80.7	**82.6**	77.8	79.2	**81.5**	73.1	**75.8**	69.1	70.9	**74.2**	45.6	48.4	37.9	42.6	47.3
	5	79.5	81.3	76.6	78.2	**82.7**	71.4	73.9	67.4	70.2	**75.9**	42.8	42.1	35.6	43.0	**51.6**
	10	**82.7**	80.7	76.1	79.0	80.1	**75.9**	73.1	66.6	70.7	72.2	47.3	41.8	37.1	46.9	42.5
t	20	**82.6**	79.6	77.8	80.4	80.5	**75.8**	71.5	69.0	72.7	72.8	44.5	42.6	42.6	49.6	47.1
	30	**82.5**	79.3	80.1	79.2	80.5	**75.6**	71.1	72.2	70.9	72.8	49.6	46.1	49.2	41.4	47.1
	100	**82.8**	80.3	79.7	79.1	80.5	**76.0**	72.6	71.7	70.8	72.8	**51.0**	47.7	46.1	41.4	47.1
	250	**82.2**	80.7	79.7	79.1	80.5	**75.2**	73.1	71.7	70.8	72.8	**51.0**	47.7	46.1	41.4	47.1
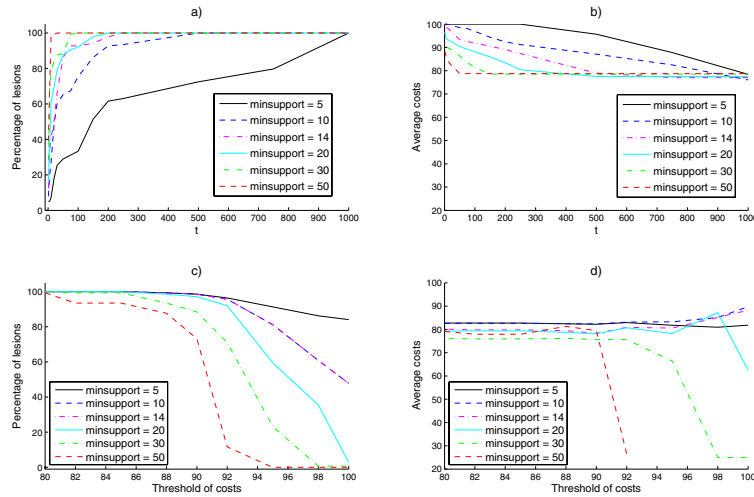
Fig. 3. a) Percentage of lesions that are covered by at least one subgroup for different values of t (x-axis) and *minsupport*. b) Average costs (prediction quality) of lesions when choosing the modality proposed by the best subgroup ignoring lesions that are not covered by any subgroup. c) Percentage of lesions covered by at least one subgroup with pq above a certain threshold (x-axis). d) Average quality of lesions when choosing modality proposed by the majority of the (maximal 10) best subgroups above the threshold. Lesions covered by no subgroup above the threshold are ignored.

We further validate the overall coverage of the lesions by the generated t subgroups. Figure 3a shows the percentage of lesions that are covered by at least one subgroup. With an increasing number of generated subgroups t, the coverage increases and the average quality (Figure 3b) decreases. It also shows that a higher minsupport induces a higher coverage, even for low values of t. Also for those cases the quality decreases. Figure 3c shows the behavior of the

proportion of lesions covered by a subgroup when introducing a threshold, which has to be overcome by the prediction quality of the subgroups. Subgroups with lower prediction qualities are ignored in this setting. Of course with raising the threshold the number of uncovered lesions increases. Small *minsupport* allows more lesions to be covered. The average quality increases with a higher threshold for low *minsupport* and decreases for high *minsupport* and a higher threshold. The larger subgroups seem to be not specific enough.

5 Related Work

Test selection has been investigated extensively in the medical and the statistical literature. Andreassen [1] presents a valid framework for test selection based on conditional probabilistic networks. However, it does not allow identifying all subgroups with optimal costs. Doubilet [2] offers a mathematical approach to test selection, however, it assumes that prior probabilities can be computed or estimated, which is problematic for small training sets. Furthermore it is not clear how background knowledge can be incorporated. It proposes only models for the entire population instead of individual models for smaller subgroups. Other subgroup discovery algorithms [3,4,7,8,9] mostly focus on finding subgroups that are interesting or unusual with respect to a single target variable (mostly class membership; for numerical variables see [7]). In our problem setting we need a more complex target variable that expresses the relation between the test outcome and the biopsy. Exceptional Model Mining [10] provides an approach that is able to discover subgroups with a more complex target concept: a model and its fitting to a subgroup. It performs a level-wise beam search and explores the best t subgroups of each level. In contrast, $SD4TS$ does not require the definition of a model and is guaranteed to find the globally optimal subgroup. While subgroup discovery usually aims for a descriptive exploration of the entire population, we discover for each patient only subgroups that are supported by the patients features. Therefore, we do not need a covering heuristic. With the introduction of prediction quality we have a measure that enables quality pruning of the search space (comparable to optimistic estimate pruning [9]), whereas existing algorithms quite often only offer pruning according to frequency [3]. While test selection and subgroup discovery are well-investigated areas of research, their combination has not yet been considered in the literature.

6 Conclusion

Many questions in medical research are related to the discovery of statistically interesting subgroups of patients. However, subgroup discovery with a single target variable is rarely sufficient in practice. Rather, more complex variants, e.g., handling costs, are required. In this study, we considered such variants of subgroup discovery in the context of the clinical task of test selection and diagnosis: For our breast cancer diagnosis scenario, the task is to detect subgroups for which single modalities should be given priority over others, as indicated by

a cost function. We designed an algorithm that handles costs and finds the most cost-efficient subgroups of a population. By limiting the output size to the best t subgroups, it is possible to prune the search space considerably, especially for lower values of the minimum frequency (i.e., support) parameter. Consequently, the proposed algorithm clearly outperforms the baseline algorithm used for comparison (APRIORI-SD) in our experiments. The main problems and limitations in the context of our study were caused by the small sample size (138 examples, i.e., lesions) and the non-unique solution for optimal test selection. In other words, for many cases, two or more tests perform equally well in practice. In future work we will investigate how to solve our task by applying slightly modified methods from Section 5. However, we expect this to result in longer runtimes. Moreover, we are planning to compare the method with a recently poposed information-theoretic approach of test selection [11]. Overall, we showed that subgroup discovery can be adapted for test selection. We believe that similar techniques should be applicable successfully in other areas as well.

References

1. Andreassen, S.: Planning of therapy and tests in causal probabilistic networks. Artifical Intelligence in Medicine 4, 227–241 (1992)
2. Doubilet, P.: A mathematical approach to interpretation and selection of diagnostic tests. Medical Decision Making 3, 177–195 (1983)
3. Kavšek, B., Lavrač, N.: APRIORI-SD: Adapting association rule learning to subgroup discovery. Applied Artificial Intelligence 20(7), 543–583 (2006)
4. Atzmüller, M., Puppe, F.: SD-map – A fast algorithm for exhaustive subgroup discovery. In: Fürnkranz, J., Scheffer, T., Spiliopoulou, M. (eds.) PKDD 2006. LNCS (LNAI), vol. 4213, pp. 6–17. Springer, Heidelberg (2006)
5. BI-RADS Breast Imaging Reporting and Data System, Breast Imaging Atlas. 4th edn. American College of Radiology (2003)
6. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: Proceedings of the 20th VLDB Conference, pp. 487–499 (1994)
7. Klösgen, W.: Explora: a multipattern and multistrategy discovery assistant, 249–271 (1996)
8. Lavrač, N., Kavšek, B., Flach, P., Todorovski, L.: Subgroup discovery with CN2-SD. Journal of Machine Learning Research (2004)
9. Wrobel, S.: An algorithm for multi-relational discovery of subgroups. In: Komorowski, J., Żytkow, J.M. (eds.) PKDD 1997. LNCS, vol. 1263, pp. 78–87. Springer, Heidelberg (1997)
10. Leman, D., Feelders, A., Knobbe, A.J.: Exceptional model mining. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008, Part II. LNCS (LNAI), vol. 5212, pp. 1–16. Springer, Heidelberg (2008)
11. Mueller, M., Rosales, R., Steck, H., Krishnan, S., Rao, B., Kramer, S.: Data-efficient information-theoretic test selection. In: Proceedings of the 12th Conference on Artificial Intelligence in Medicine (AIME 2009), pp. 410–415 (2009)