Quality Measures for Semi-Automatic Learning of Simple Diagnostic Rule Bases

Martin Atzmueller, Joachim Baumeister, Frank Puppe

University of Würzburg, 97074 Würzburg, Germany
Department of Computer Science
Phone: +49 931 888-6739, Fax: +49 931 888-6732
email: {atzmueller, baumeister, puppe}@informatik.uni-wuerzburg.de

Abstract. Semi-automatic data mining approaches often yield better results than plain automatic methods, due to early integration of the user's goals. For example in the medical domain, experts are likely to favor simpler models instead of more complex models. Then, the accuracy of discovered patterns is often not the only criterion to consider. Instead, the simplicity of the discovered knowledge is of prime importance, since this relates directly to the understandability and the interpretability of the learned knowledge.
In this paper, we present quality measures considering the understandability and the accuracy of (learned) rule bases. We describe an unifying quality measure, which can trade-off small losses concerning accuracy vs. increased simplicity. Furthermore, we introduce a semi-automatic data mining method for learning understandable and accurate rule bases. The presented work is evaluated using cases from a real world application in the medical domain.

1 Introduction

Automatic methods for learning rules commonly perform well, concerning the classification accuracy of the learned models. However, often the understandability of the learned patterns is poor, which is problematic if the learned knowledge should be manually processed in further steps. Semi-automatic approaches often yield better results than plain automatic methods, due to early integration of the user's goals. In such semi-automatic scenarios, the learned knowledge is not used as a black-box reasoning engine, but can be refined incrementally by other techniques, e.g., human interpretation. Furthermore, semi-automatic learning methods can incorporate additional background knowledge for further quality improvements. When guiding the knowledge discovery process, it often turns out that user interests concerning the accuracy of the learned knowledge are not related to other aspects, e.g., simplicity of the patterns [1,2]. So, the knowledge discovery method should take both accuracy and simplicity of the learned knowledge into account.

In this paper, we present quality measures for rating the simplicity of a learned rule base, and we will briefly introduce a semi-automatic learning method for simple scoring rules. Besides the discussed quality measures we propose an unifying quality measure balancing the accuracy and the understandability of a given rule base. It is worth noticing, that the presented measures are not only applicable to scoring rules

but can be easily generalized to other rule-based approaches, e.g., association rules. However, in this paper we will focus on the application of scoring rules.

The rest of the paper is organized as follows: In Section 2 we define the basic notions used in this paper, and we introduce diagnostic scores implemented by scoring rules as an intuitive concept for representing diagnostic knowledge. In Section 3 we present simplicity measures for diagnostic scores and scoring rules. These measures are used to determine the understandability of the learned knowledge. We present an unifying quality measure taking both the simplicity and the accuracy of the rule base into account. In Section 4 we outline a method for learning diagnostic scores, and discuss additional background knowledge that is applicable to the learning task. An evaluation using a real-world case base is given in Section 5. We conclude the paper in Section 6 discussing the presented work, and we show promising directions for future work.

2 Diagnostic Scores using Scoring Rules – an Overview

Before describing diagnostic scores, we first define the knowledge representation schema. Let Ω_Q be the universe set of all questions available in the problem domain. In the context of machine learning methods, questions are commonly called *attributes*. A value $v \in dom(q)$ assigned to a question $q \in \Omega_Q$ is called a *finding*, and we call $\Omega_\mathcal{F}$ the set of all possible findings in the given problem domain. A finding $f \in \Omega_\mathcal{F}$ is denoted by $q{:}v$ for $q \in \Omega_Q$ and $v \in dom(q)$. The set $F_q \subseteq \Omega_\mathcal{F}$ of possible findings for a given question q is defined as $F_q = \{f \in \Omega_\mathcal{F} \mid f = q{:}v \wedge v \in dom(q)\}$. Each finding $f \in \Omega_\mathcal{F}$ is defined as a possible input of a diagnostic knowledge system.

Let d be a *diagnosis* representing a possible output, of a diagnostic knowledge system. We define $\Omega_\mathcal{D}$ to be the universe of all possible diagnoses for a given problem domain. With respect to a given problem, a diagnosis $d \in \Omega_\mathcal{D}$ is assigned to a symbolic state $dom(d) = \{unprobable, undefined, probable\}$.

A *case* c is defined as a tuple $c = (\,\mathcal{F}_c, \mathcal{D}_c, \mathcal{I}_c\,)$, where $\mathcal{F}_c \subset \Omega_\mathcal{F}$ is the set of *observed findings* for the given case. The set $\mathcal{D}_c \subseteq \Omega_\mathcal{D}$ contains the diagnoses describing the solution of the case c, and $\mathcal{I}_c$ contains additional (meta-) information describing the case c in more detail. The set of all possible cases for a given problem domain is denoted by Ω_C. For the learning task, we consider a case base $CB \subseteq \Omega_C$ containing all available cases that have been solved previously.

Diagnostic scores, e.g., [3,4] are a rather wide spread formalism for medical decision making. For inferring a diagnosis, a limited number of findings is used in a regular and simple to interpret manner. In its simplest form, each observed finding individually contributes one point to an account. If the total score of the account exceeds a given threshold, then the diagnosis is established. Diagnostic scores are commonly implemented using scoring rules, which are used to infer a specific diagnosis. A *simple scoring rule* r is denoted by $r = f \xrightarrow{s} d$, where $f \in \Omega_\mathcal{F}$ is a finding, and $d \in \Omega_\mathcal{D}$ is the target diagnosis. For each rule a symbolic confirmation category $s \in \Omega_{scr}$ is attached with $\Omega_{scr} \in \{\, S_3, S_2, S_1, 0, S_{-1}, S_{-2}, S_{-3} \,\}$. Formally, a diagnostic score $DS(d)$ for a diagnosis $d \in \Omega_\mathcal{D}$ is defined as the set of scoring rules $r \in \mathcal{R}$ that contain d in their rule action, i.e., $DS(d) = \{r \in \mathcal{R} \mid r = f \xrightarrow{s} d \wedge f \in \Omega_\mathcal{F}\}$. Let Ω_R be the universe of

all possible rules for the sets $\Omega_{\mathcal{F}}$, $\Omega_{\mathcal{D}}$ and Ω_{scr}. Then, we call $\mathcal{R} \subseteq \Omega_R$ the *rule base* containing the inferential knowledge of the problem domain.

Confirmation categories of scoring rules are used to represent a qualitative degree of uncertainty. In contrast to quantitative approaches, e.g., Bayesian methods, symbolic categories state the degree of confirmation or disconfirmation for a diagnosis. In this way, a symbolic category s expresses the uncertainty for which the observation of finding f will confirm/disconfirm the diagnosis d. Whereas $s \in \{S_1, S_2, S_3\}$ stand for confirming symbolic categories in ascending order, the categories $s \in \{S_{\text{-}1}, S_{\text{-}2}, S_{\text{-}3}\}$ are ascending categories for disconfirming a diagnosis. A rule with category 0 has no effect on the diagnosis' state, and therefore is usually omitted from the rule base. It is worth noticing, that the value range Ω_{scr} of the possible symbolic categories is not fixed. For a more detailed (or coarse) representation of confirmation the value range may be extended (or reduced).

For a given case $c \in \Omega_C$ the final state of each diagnosis $d \in \Omega_{\mathcal{D}}$ is determined by evaluating the available scoring rules $r \in \mathcal{R}$ targeting d. Thus, rules $r = f \xrightarrow{s} d$ contained in $\mathcal{R}$ are activated, iff f is observed in case c, i.e., $f \in \mathcal{F}_c$. The symbolic categories of the activated rules are aggregated by adding the categories in a way, so that four equal categories result in the next higher category (e.g., $S_1 + S_1 + S_1 + S_1 = S_2$), and so that two equal categories with opposite sign nullify (e.g., $S_1 + S_{\text{-}1} = 0$). For a more detailed or coarse definition of Ω_{scr} the aggregation rules may be adapted. A diagnosis is assumed to be *probable* (i.e., part of the final solution of the case), if the aggregated score is greater or equal than the symbolic category S_3. Analogously, a diagnosis is assumed to be *unprobable*, if the aggregated score is less or equal than the symbolic category $S_{\text{-}3}$.

Related Work. Scoring rules have proven to be useful in large medical knowledge bases, e.g., in the INTERNIST/QMR project [5]. In our own work with the shell-kit D3, scores have been applied successfully in many (large) knowledge system projects, e.g., in a geo-ecological application [6] or in medical domains and technical domains [3] using generalized scores.

3 Quality Measures for Diagnostic Rule Bases

When we consider the quality of the learned knowledge in the semi-automatic setting, then we are not only interested in classification accuracy, but also in understandability of the learned patterns. The understandability of the learned scores is typically defined by its simplicity, which can be measured with respect to the learned scoring rules in the rule base $\mathcal{R} \subseteq \Omega_R$. If the learned rules have a low complexity, then it is easier for the expert/user to understand the corresponding scores.

In general, a score is considered to be the more complex, the more findings it contains. This directly corresponds to the number of learned rules per diagnosis. An overall impression of the simplicity of the learned scores is given by the total number of learned rules. Furthermore, as a global simplicity measure we count the total number of findings used in scoring rules of the rule base. Usually a moderate number of findings is considered more comprehensible than a huge number of findings. In the following we discuss simplicity measures and accuracy measures for diagnostic scores.

Simplicity Measures. It is difficult to determine the simplicity of a rule base by only one measure. For defining *global* simplicity measures, we consider the rule base as a whole. *Local* variants, considering a specific knowledge item, i.e., a diagnostic score, can be defined accordingly. In contrast to the local simplicity measures the use of global measures is appropriate for comparing the understandability of different rule bases. We consider the following issues and define corresponding functions applied on scoring rule bases.

- APPLIED FINDINGS: Number of findings used in the rule base; the rule base is much simpler to survey, if fewer findings are used to describe the scores.
- RULE BASE SIZE: Overall number of learned scoring rules; obviously the number of scoring rules is a direct measure for the complexity of the learned knowledge. However, for a more detailed analysis of the rule base complexity the applied classes of confirmation categories should be considered. Thus, the interpretation of scoring rules categorically establishing or excluding a diagnosis, i.e., S_3, $S_{\text{-}3}$, is very simple, when compared to scoring rules with less certain confirmation categories, e.g., S_1, $S_{\text{-}1}$.

  Therefore, it is suggestive to define a weighting function $w : \Omega_{scr} \rightarrow I\!N$ for confirmation categories. In the context of our work we defined $w(s) = 1$ for $s \in \{S_3, S_{\text{-}3}\}$, and $w(s) = 2$ otherwise. Thus, we define a category $s \in \Omega_{scr} \setminus \{S_3, S_{\text{-}3}\}$ to be as double complex as the categories $S_3, S_{\text{-}3}$, which are categorically (de)establishing a diagnosis.

  In summary, the measure RULE BASE SIZE is simply defined by the count of the rules contained in the rule base. A more refined measure RULE BASE SIZE for a rule base $\mathcal{R} \subseteq \Omega_R$ is defined as follows

$$\text{RULE BASE SIZE}(\mathcal{R}) = \sum_{r \in \mathcal{R}} w\big(category(r)\big)\,. \tag{1}$$

- MEAN SCORING RULES: This measure gives the mean number of rules for scoring a single diagnosis, and can be derived from RULE BASE SIZE. Obviously, less scoring rules for a diagnosis are much simpler to understand than more rules. Similar to the measure RULE BASE SIZE, it can be computed using the weighted categories or by directly counting the number of rules.
- MEAN SCORE CATEGORIES: Mean number of different confirmation categories applied for scoring a single diagnosis. A smaller number of distinct categories allow for a much simpler interpretation of the diagnosis score, since confirmation strengths of the findings contributing to a score are less distributed. This measure is indirectly dependent on the global number of different confirmation categories defined, i.e., $|\Omega_{scr}|$. A small universe of possible confirmation categories allows for a simpler distinction between the single categories.

In addition to the simplicity measures, the second part of the quality measures for the semi-automatic learning task are measures concerning the accuracy.

Accuracy Measures. There exists a variety of methods for assessing the accuracy of individual rules, or the whole rule base. Several factors need to be considered. For a two-class prediction problem, e.g., predicting a single diagnosis, we have to consider four possible outcomes, shown in the following table.

	Predicted Class = YES	Predicted Class = NO
Actual Class = YES	*True Positive*	*False Negative*
Actual Class = NO	*False Positive*	*True Negative*

The true positives and true negatives are correct classifications. If the class is incorrectly predicted as 'YES' while it is in fact 'NO', then we have a false positive. Likewise, if the class is incorrectly predicted as negative while it is in fact positive, then we have a false negative. For the different measures the trade-off between these classification alternatives has to be taken into account. In the following, TP, FP, TN, FN denote the number of true positives, false positives, true negatives, and false negatives, respectively.

For measuring the different tradeoffs between correct and false classifications, there exist several measures like *sensitivity* ($TP/(TP+FN)$), *specificity* ($TN/(TN+FP)$) from diagnosis, or likewise *precision* ($TP/(TP+FP)$) and *recall* (same as sensitivity) from information theory. The *success rate*, or *efficiency*, is a widely used measure: $(TP+TN)/(TP+TN+FP+FN)$. However, a single diagnosis, which is not predicted very frequently, and which also does not occur very frequently as the correct diagnosis of a case, might get a better rating, than a diagnosis which occurs more frequently. This is especially relevant, if we apply a case base with multiple disorders, as experienced in our evaluation setting. Therefore we used the *F-measure*, for all applied diagnoses $d \in \Omega_{\mathcal{D}}$. The *F-measure*, the harmonic mean between recall and precision, is defined as follows:

$$F(\mathcal{D}_c, \mathcal{D}_s) = \frac{(\beta^2+1) \cdot prec(\mathcal{D}_c, \mathcal{D}_i) \cdot recall(\mathcal{D}_c, \mathcal{D}_p)}{\beta^2 \cdot prec(\mathcal{D}_1, \mathcal{D}_2) + recall(\mathcal{D}_1, \mathcal{D}_2)}\,, \tag{2}$$

where D_c is the correct solution and D_p specifies the proposed, inferred solution. β denotes a constant weight for the precision, where usually a default of $\beta = 1$ is used.

An Unifying Quality Measure for Semi-Automatic Learning Methods. In a semi-automatic scenario, the user wants to obtain an overview of the quality of the learned knowledge. Concerning accuracy and simplicity of the learned knowledge, often there is a trade-off between these two measures. Then, quite accurate learned models are quite complex, while simpler ones lack performance. So the two measures need to be balanced. Also, user quality standards need to be taken into account regarding simplicity, since simplicity is subjective to the user's goals and is also dependent on the applied domain. We combine a normalized simplicity measure and the accuracy measure into a single quality measure. For the simplicity measure we first define a local simplicity measure SCORING RULES, which gives the absolute number of scoring rules for scoring a single specific diagnosis. Then, the function SCORING RULES$(DS(d))$ returns the number of scoring rules learned for the specified diagnostic score $DS(d)$.

For the definition of the unifying quality measure QM, we first introduce a normalized simplicity measure concerning a single diagnostic score $DS(d)$.

$$NSM\big(DS(d)\big) = 1 - \frac{\text{SCORING RULES}\big(DS(d)\big) - 1}{\text{SCORING RULES}\big(DS(d)\big) + \gamma}\,, \tag{3}$$

where γ is a generalization parameter, with default value $\gamma = 1$. If γ is set to larger values, then larger scores will get an increased simplicity value. Since $NSM(DS(d)) \in$

$[0;1]$, the maximum value $NSM(DS(d)) = 1$ is obtained, if a diagnosis d is predicted with a single rule, i.e., if the score has the size one.

We propose to combine this measure with the accuracy in a term similar to the F-measure balancing both measures. Then, the unifying quality measure $QM : 2^{\Omega_R} \rightarrow [0;1]$ for a rule base $\mathcal{R}$ is defined as follows,

$$QM(\mathcal{R}) = \frac{1}{|\Omega_{\mathcal{D}}|} \sum_{d \in \Omega_{\mathcal{D}}} \frac{(\alpha^2 + 1) \cdot NSM\big(DS(d)\big) \cdot ACC\big(DS(d)\big)}{\alpha^2 \cdot NSM\big(DS(d)\big) + ACC\big(DS(d)\big)}\,, \tag{4}$$

where the function $ACC\big(DS(d)\big)$ calculates the accuracy of the specified diagnostic score $DS(d)$ using the F-measure. The factor α is a weight balancing simplicity vs. accuracy. We used $\alpha = 1$ for our experiments.

Related Work. Favoring simple rules is in line with a classic principle of inductive learning methods called Ockham's Razor [7]. Existing interestingness measures applying this principle generate compact rules [8], for example, which takes the number of rules, the number of conditions in a rule, and the classification accuracy of a rule into account. A general measure discussed by [9] takes the size of the disjuncts of a rule into account. Due to the fact that we only consider simple scoring rules not containing disjuncts, this measure is not applicable to diagnostic scores. We purely concentrate on the syntactic elements contained in the rule base $\mathcal{R}$. For a localized evaluation, we propose an unifying quality measure, which combines both aspects, i.e., the simplicity and the accuracy. This measure with fixed upper and lower bounds provides a first intuitive evaluation for the user.

4 Learning Diagnostic Scores

In the following we outline the method for learning diagnostic scores. Due to the limited space we refer to a previous paper [10] for an in-depth discussion of the algorithm.

For learning diagnostic scores we first have to identify dependencies between findings and diagnoses. In general, all possible combinations between diagnoses and findings have to be taken into account. However, to reduce the search space, we only consider the findings occurring most frequently with the diagnosis. In summary, we basically apply three steps for learning a diagnostic scoring rule:

1. Identify a dependency between a finding $f \in \Omega_{\mathcal{F}}$ and a diagnosis $d \in \Omega_{\mathcal{D}}$
2. Rate this dependency and map it to a symbolic category $s \in \Omega_{scr}$
3. Finally, construct a diagnostic rule: $r = f \xrightarrow{s} d$

Identify Dependencies. For each diagnosis $d \in \Omega_{\mathcal{D}}$, we create a diagnostic profile containing the most frequent findings occurring with the diagnosis. We consider all attributes (questions) in the profile selecting the findings which are observed in the case base. For each finding $f = q{:}v$ we apply the χ^2*-test for independence* for binary variables, i.e., variable D for diagnosis d and variable F for finding f, respectively. D and F measure if d and f occur in a case. If they occur the respective variable is true and false otherwise.

For all dependent tuples (F, D) we derive the quality of the dependency, i.e., the strength of the association using the ϕ-coefficient, $\phi(F, D) \in [-1; 1]$, which is a correlation measure between two binary variables. We use it to discover positive or negative dependencies. A positive value of $\phi(F, D)$ signifies a positive association, whereas a negative value signifies a negative one. If the absolute value of $\phi(F, D)$ is less than a certain threshold $threshold_c$, i.e., $|\phi(F, D)| < threshold_c$, then we do not consider this weak dependency for rule generation. For the remaining dependencies we generate rules described as follows: If $\phi(F, D) < 0$, then we obtain a negative association between the two variables, and we generate a rule $f \xrightarrow{s} d$ with a negative category s. If $\phi(F, D) > 0$, then we construct a rule $f \xrightarrow{s} d$ with a positive category s.

Mapping Dependencies. For determining the exact symbolic confirmation category of the remaining rules r, we utilize two measures used in diagnosis: ***precision*** and the ***false alarm rate (FAR)***. The precision of a rule r is defined as $prec(r) = TP/(TP + FP)$, whereas the false alarm rate *FAR* for a rule r is defined as $FAR(r) = FP/(FP+TN)$.

To score the dependency, we first compute a *quasi probabilistic score (qps)* which we then map to a symbolic category. The numeric qps score for a rule r is computed as follows $qps(r) = sgn\big(\phi(D, F)\big) * prec(r)\big(1 - FAR(r)\big)$. We achieve a tradeoff between the accuracy of the diagnostic scoring rule to predict a disease measured against all predictions and the proportion of false predictions. The qps-scores are mapped to the symbolic categories according to the following conversion table ($\varepsilon \approx 0$):

qps(r)	*category(r)*	*qps(r)*	*category(r)*
[-1.0, -0.9)	$\rightharpoonup S_{-3}$	(ε, 0.5)	$\rightharpoonup S_1$
[-0.9, -0.5)	$\rightharpoonup S_{-2}$	[0.5, 0.9)	$\rightharpoonup S_2$
[-0.5, $-\varepsilon$)	$\rightharpoonup S_{-1}$	[0.9, 1.0]	$\rightharpoonup S_3$

We accept the loss of information to increase the understandability and to facilitate a user-friendly adaptation of the learned diagnostic scores.

Including Background Knowledge The presented algorithm can be augmented with background knowledge in order to achieve better learning results. We introduce abnormality information and partition class knowledge as appropriate background knowledge.

If *abnormality* information about attribute values is available, then each value v of a question q is attached with an abnormality label. It explains, whether v is describing a normal or an abnormal state of the question. For example, consider the choice question temperature with the value range: *normal*, *marginal*, *high*. The values *normal* and *marginal* denote normal values of the question, whereas the value *high* describes an abnormal value. We will use these abnormalities, for further shrinking the size of the generated rule base. Let $r = q{:}v \xrightarrow{s} d$ be a scoring rule. If $s \in \Omega_{scr}$ denotes a positive category and v is a normal value of attribute q, then we omit rule r, since findings describing normal behavior usually should not increase the confirmation of a diagnosis. Furthermore, if s denotes a negative category and v is an abnormal value of attribute q, then we likewise omit rule r, because an abnormal finding usually should not decrease the confirmation of a diagnosis, but possibly increases the confirmation of other diagnoses.

As a second type of background knowledge the expert can provide *partition class* knowledge describing how to divide the set of diagnoses and attributes into partially disjunctive subsets, i.e., partitions. These subsets correspond to certain problem areas of the application domain. For example, in the medical domain of sonography, we have subsets corresponding to problem areas like *liver*, *pancreas*, *kidney*, *stomach*, and *intestine*. This knowledge is especially useful when diagnosing multiple faults. Since a case may contain multiple diagnoses, attributes occurring with several diagnoses will be contained in several diagnostic profiles. We reduce noise and irrelevant dependencies by pruning such discovered dependencies $f \rightarrow d$, for which f and d are not in the same partition class.

5 Evaluation

We evaluated the presented methods with cases taken from a medical application, which is currently in routine use. The applied SONOCONSULT case base contains 1340 cases, with a mean of diagnoses $M_d = 4.32 \pm 2.79$ and a mean of relevant findings $M_f = 76.89 \pm 20.59$ per case. SONOCONSULT [11] is a knowledge-based documentation and consultation system for sonography. It is developed and maintained by the domain experts using the shell-kit D3 [12]. The quality of the derived diagnoses usually is very good, i.e., the solutions are correct in nearly all cases.

For the evaluation of our experiments we adopted the F-measure introduced in Section 3, adapting this to the multiple disorder problem occurring in our case base (cf. [10] for details). Furthermore, a stratified 10-fold cross-validation method was applied. We performed two experiments, to determine the impact of including background knowledge into the learning process. For experiment *E0* we applied no background knowledge at all. To demonstrate how the utilization of knowledge improves the results, we used both partition class knowledge and abnormality knowledge for experiment *E1*. We created several sets of scores depending on the parameter $threshold_c$, which describes the correlation threshold used in the learning algorithm Two criteria – accuracy and simplicity – as outlined in Section 3, were used to define the quality of the scores.

The results are presented in the following tables. Column $threshold_c$ specifies the correlation threshold, QM_1 shows the combined quality measure with the default parameter $\gamma = 1$, whereas QM_5 and QM_{10} show the measure with a parameter $\gamma = 5$ and $\gamma = 10$, respectively. *MR* corresponds to the measure MEAN SCORING RULES, attached with standard deviation. *RBS* describes the measure RULE BASE SIZE with total number of rules in addition to the number of weighted rules in parentheses (as described in Section 3). The column *SC* corresponds to the measure MEAN SCORE CATEGORIES. Column *AF* shows the number of applied findings, i.e., the values of the measure APPLIED FINDINGS. Finally, we depict the accuracy of the rule base using the F-measure in column *ACC*.

Experiment E0: no knowledge used								
$threshold_c$	QM_1	QM_5	QM_{10}	ACC	RBS *(w)*	MR	AF	SC
0.2	0.15	0.33	0.46	0.94	2201 (3798)	30.58 ± 16.83	395	3.49
0.3	0.21	0.41	0.54	0.92	1510 (2466)	20.97 ± 10.59	348	3.20
0.4	0.27	0.49	0.61	0.90	1069 (1647)	14.85 ± 7.02	297	2.98
0.5	0.34	0.56	0.68	0.89	770 (1101)	10.70 ± 4.90	247	2.67
0.6	0.40	0.61	0.72	0.83	594 (789)	8.24 ± 3.51	207	2.32
0.7	0.51	0.70	0.77	0.81	369 (413)	5.13 ± 2.13	158	1.44

Experiment E1: using partition class and abnormality knowledge								
$threshold_c$	QM_1	QM_5	QM_{10}	ACC	RBS *(w)*	MR	AF	SC
0.2	0.39	0.59	0.68	0.88	594 (990)	8.25 ± 5.00	180	2.60
0.3	0.45	0.64	0.72	0.86	437 (693)	6.07 ± 3.30	153	2.38
0.4	0.51	0.68	0.75	0.85	328 (495)	4.56 ± 2.15	131	2.12
0.5	0.58	0.72	0.77	0.84	240 (335)	3.34 ± 1.36	113	1.78
0.6	0.62	0.73	0.77	0.78	188 (245)	2.61 ± 1.04	101	1.49
0.7	0.68	0.75	0.77	0.76	131 (149)	1.81 ± 0.70	81	1.10

The high values of the accuracy for low values of $threshold_c$ and the large number of rules per diagnosis indicate overfitting of the learned knowledge. This is of course domain dependent, and therefore the expert needs to tune the threshold carefully. With greater values for $threshold_c$ less rules are generated, since only strong dependencies are taken into account. If $threshold_c$ is too high, i.e., if too many rules are pruned, this obviously degrades the accuracy of the learned scores. In our experiments this occurs for $threshold_c = 0.6$, for which the accuracy decreases significantly in comparison to $threshold_c = 0.5$. Furthermore, the number of rules per diagnosis (MR) is reduced considerably without decreasing the accuracy (ACC) significantly from $threshold_c = 0.2$ to $threshold_c = 0.5$. Analogously, the number of applied findings (AF) is reduced with an increasing value of $threshold_c$ but a decreasing accuracy. Column SC indicates that the number of applied confirmation categories is reduced by an increased $threshold_c$, i.e., simpler scoring rules are learned.

These findings are also reflected in the unifying quality measure. It is obvious, that the balance between the scores' accuracy and simplicity depends on the generalization parameter γ. QM_1 has a stronger increase from $threshold_c = 0.6$ to $threshold_c = 0.7$ than QM_5 since the size of the score, i.e., the number of rules has a higher impact.This depends on the priorities of the user: If $\gamma = 1$, then we have a strong bias favoring minimal scores, one rule per diagnosis in the best case. If γ is set to higher values, then we generalize this, such that the accuracy is more important. This can be seen in experiment *E1* for $\gamma = 10$ considering thresholds $0.5, 0.6,$ and 0.7 where the increased score simplicity is balanced by the decrease in score accuracy. In the case of such a plateau, the user either has to consult the detailed quality measures to trade-off accuracy vs. simplicity, or can tune the quality measure with respect to the parameter α, i.e., the weighting factor which trades-off simplicity vs. accuracy. Then, either a clear cut-off point is found, where the quality measure has a maximum value, or the appropriate cut-off point has to be selected from a limited number of options, in the case of a plateau, i.e., a set of equal values.

6 Conclusion

We presented a semi-automatic learning method for simple diagnostic scoring rules with appropriate quality measures. These focus on the understandability, i.e., simplicity. An unifying measure also takes the accuracy of the learned knowledge into account. This measure allows for a first quick evaluation of the learned patterns. The measure can be fine-tuned guided by the user's expectations and goals. This is especially important in the context of semi-automatic learning methods, which can be refined incrementally taking different amounts of background knowledge into account. As an example of such a semi-automatic approach, we outlined a method for learning simple diagnostic scores, and presented an evaluation of the proposed methods using a case base from a real-world application. This demonstrated the applicability of the presented simplicity measures and the unifying quality measure balancing simplicity and accuracy aspects.

In the future, we are planning to apply the measures on other rule-based patterns, such as subgroups. Additionally, interpretation and evaluation of the learned knowledge together with the proposed quality measures by medical experts should further demonstrate the significance of these measures.

References

1. Ho, T., Saito, A., Kawasaki, S., Nguyen, D., Nguyen, T.: Failure and Success Experience in Mining Stomach Cancer Data. In: International Workshop Data Mining Lessons Learned, International Conf. Machine Learning. (2002) 40–47
2. Gamberger, D., Lavrac, N.: Expert-Guided Subgroup Discovery: Methodology and Application. Journal of Artificial Intelligence Research **17** (2002) 501–527
3. Puppe, F., Ziegler, S., Martin, U., Hupp, J.: Wissensbasierte Diagnosesysteme im Service-Support (Diagnostic Knowledge Systems for the Service-Support). Springer Verlag (2001)
4. Ohmann, C., et al.: Clinical Benefit of a Diagnostic Score for Appendicitis: Results of a Prospective Interventional Study. Archives of Surgery **134** (1999) 993–996
5. R., M., Pople, H.E., Myers, J.: Internist-1, an Experimental Computer-Based Diagnostic Consultant for General Internal Medicine. NEJM **307** (1982) 468–476
6. Neumann, M., Baumeister, J., Liess, M., Schulz, R.: An Expert System to Estimate the Pesticide Contamination of Small Streams using Benthic Macroinvertebrates as Bioindicators, Part 2. Ecological Indicators **2** (2003) 391–401
7. Mitchell, T.: Machine Learning. McGraw-Hill Comp. (1997)
8. Yen, S.J., Chen, A.L.P.: An Efficient Algorithm for Deriving Compact Rules from Databases. In: Ling, Masunaga: Proceedings of the 4th International Conference on Database Systems for Advanced Applications-95. Volume 5 of Advanced Database Research and Development Series., World Scientific (1995) 364–371
9. Freitas, A.A.: On Rule Interestingness Measures. Knowledge-Based Systems **12** (1999) 309–325
10. Atzmueller, M., Baumeister, J., Puppe, F.: Inductive Learning of Simple Diagnostic Scores. In: Medical Data Analysis, Proceedings of the International Symposium of Medical Data Analysis (ISMDA). Number 2868 in LNCS, Springer Verlag (2003) 23–30
11. Huettig, M., Buscher, G., Menzel, T., Scheppach, W., Puppe, F., Buscher, H.P.: A Diagnostic Expert System for Structured Reports, Quality Assessment, and Training of Residents in Sonography. submitted to Medizinische Klinik (2003)
12. Puppe, F.: Knowledge Reuse Among Diagnostic Problem-Solving Methods in the Shell-Kit D3. Int. J. Human-Computer Studies **49** (1998) 627–649