Learning and Inspecting Classification Rules from Longitudinal Epidemiological Data to Identify Predictive Features on Hepatic Steatosis

Uli Niemann[a], Henry Völzke[b], Jens-Peter Kühn[c], Myra Spiliopoulou[a]

[a]*Faculty of Computer Science, Otto-von-Guericke University Magdeburg, Universitätsplatz 2, D-39106 Magdeburg, Germany*
[b]*Institute for Community Medicine, Ernst-Moritz-Arndt University Greifswald, Walther-Rathenau-Straße 48, D-17475 Greifswald, Germany*
[c]*Institute for Diagnostic Radiology and Neuroradiology, Ernst-Moritz-Arndt University Greifswald, Sauerbruchstraße, D-17487 Greifswald, Germany*

Abstract

Personalized Medicine requires the analysis of epidemiological data for the identification of subgroups sharing some risk factors and exhibiting dedicated outcome risks. We investigate the potential of data mining methods for the analysis of subgroups of cohort participants on hepatic steatosis. We propose a workflow for data preparation and mining on epidemiological data and we present InteractiveRuleMiner, an interactive tool for the inspection of rules in each subpopulation, including functionalities for the juxtaposition of labeled individuals and unlabeled ones. We report on our insights on specific subpopulations that have been discovered in a data-driven rather than hypothesis-driven way.

Keywords: medical data mining, classification rules, subpopulation mining, interactive data mining, longitudinal epidemiological studies, hepatic steatosis

1. Introduction

Medical research on epidemiological data aims to identify risk factors for diseases and to contribute thus to prevention and diagnosis [22]. Epidemiological data come from *population-based* studies with a large number of randomly selected participants (the *cohorts*), for whom several variables are recorded; these may include sociodemographics, results of medical tests (e.g. on blood samples) and, recently, also magnetic resonance imaging (MRI) of body parts. Epidemiological research is commonly *hypothesis-driven*: scholars formulate a hypothesis on how a behaviour (e.g. alcohol consumption), a chronic disease (e.g. diabetes), a genetic predisposition or other factor may affect the prevalence of a certain outcome (e.g. cirrhosis, fatty liver); they then perform a careful selection of cohort participants with and without the outcome, on which the role of the candidate determinant is investigated with statistical analysis.

With the proliferation of medical engineering technology, an enormous number of variables, including biomarkers, can be recorded in population-based studies. Formulating hypotheses on each and every variable of those truly *Big epidemiological Data* is impractical, so that *data-driven* analysis of epidemiological data, namely epidemiological data mining, is now gaining momentum. Its purpose is to identify factors that are potentially associated with an outcome, so that hypothesis-driven analysis concentrates on them. This trend is further strengthen by the demands of personalized medicine [12], which requires the detection of previously unspecified subpopulations of patients that share common determinants (i.e. factors associated with an outcome).

In this study, we propose a mining workflow and an interactive tool for the discovery of potential determinants and corresponding value intervals that are associated with the multifactorial disorder hepatic steatosis (aka: fatty liver). Our emphasis is on highlighting fatty liver determinants that are not characteristic of the whole cohort but of cohort subgroups, e.g. female participants aged over 50.

Data mining methods are used widely on *clinical data* for diagnostic and therapeutic purposes, and there are also comparative studies on the performance of different mining algorithms for a specific clinical task, see e.g. [21]. However, clinical data mining analyzes data on patients, while epidemiology cohorts consist of participants with and without the outcome. Medical research on epidemiological data delivers the diagnostic indices that are later used for clinical diagnosis, e.g. the "fatty liver index" proposed by Bedogni et al. in [2]. Hence, although our workflow encompasses mining methods that have been used on clinical data, studies on clinical data mining do not provide evidence on the potential of these same methods for epidemiological data analysis tasks, as we study here.

Mining methods on epidemiological data are rather rarely used. Multiple regression is often the approach of preference [5], but linear models, Cox regression and Poisson regression have also been used - notably on the epidemiological data we analyze here [1, 11, 18, 28]. However, the use of regression in such studies is still mostly hypothesis-driven, e.g. on whether liver ultrasonography

Preprint submitted to Elsevier

Accepted 20 February 2014 DOI: 10.1016/j.eswa.2014.02.040

can predict mortaility rask from elevated serum gamma-glutamyl transpeptidase levels [11]. Our objective in this study is to demonstrate the potential of data-driven analysis for class separation in epidemiological data. Finding *new, previously unsuspected* determinants is not within the scope of our study. Rather, the potential of our approach is reflected on the quality of the learners, and, more importantly, on the data-driven identification of subpopulations that differ with respect to the class distribution, and on the data-driven discovery of associations that through independent, hypothesis-driven studies have earlier shown to exist.

We study our mining workflow on an multifactorial disorder, hepatic steatosis, using data from the first cohort of the "Study of Health in Pomerania" (SHIP); SHIP consists of population-based samples selected from Pomerania in Northeast Germany [27]. SHIP has already been extensively used for hypothesis-driven reseach on hepatic steatosis and lends itself excellently to the purposes of our analysis: we use the results of liver MRI recordings as target variable and a multitude of sociodemographic variables and medical tests for classification.

The contributions of our work are as follows. First, we propose a mining process for the classification of the participants of an epidemiological study with respect to a target outcome; we choose exemplarily the multifactorial disorder hepatic steatosis. Moreover, we propose an interactive tool, which we call InteractiveRuleMiner, with which a medical expert can drill into a derived model and investigate the properties of those subpopulations she considers interesting. Albeit mining workflows are often proposed for clinical data, mining methods for epidemiological data are rare and, in contrast to our method, they are hypothesis-driven. With our approach, a mining expert does not need to formulate hypotheses in advance, but can rather study the insights delivered by the models, identify subpopulations, drill-down on them and acquire further insights interactively.

The paper is organized as follows. In the next section, we discuss related work. In section 3 we describe materials and methods for data preparation, population partitioning and classification. In section 4 we report on the discovered models and important features for the different partitions. In section 5 we present our tool "InteractiveRuleMiner". The last section concludes the paper with a discussion and an outlook towards learning disorder progression.

2. Related Work

Medical decisions concerning the diagnosis of multifactorial diseases are based on clinical and epidemiological studies. The latter accommodate information on participants with and without the disorder and allow for discriminative model learning and, in the longitudinal design, for understanding the progress of a disorder (possibly towards a disease). There are several studies on the identification of factors (like obesity or alcohol consumption) and outcomes (like cardiovascular diseases) associated with hepatic steatosis. Findings on genetic and non-genetic factors include [13, 16, 25]; findings on associated outcomes include [26] and [19]. However, these studies identify risk factors and/or associated outcomes that pertain to the whole population. Our study emanates from the necessity to identify such factors and outcomes for subpopulations and thus to stimulate personalized diagnosis and treatment, as expected in personalized medicine [12, 29].

Classification on subpopulations is studied by Zhanga and Kodell in [32], albeit they analyze clinical data for diagnosis, while we analyze epidemiological data to identify variables associated with the outcome. Zhanga and Kodell point out that the complete population can be very heterogeneous, so that classifier performance on the whole dataset can be low. Therefore, they first train an ensemble of classifiers, then associate with each training instance the predictions made on it by each ensemble member, thus creating a new feature space where the variables are the predictions. They then perform hierarchical clustering on the instances, thus building three subpopulations: one where the prediction accuracy is high, one where it is intermediate and one where it is low [32]. With this approach, Zhanga and Kodell split the original dataset into subpopulations that are easy or difficult to classify [32]. The method seems appealing in general, but does not look promising in our case: we investigate a three-class problem with a very skewed distribution, so we already know that low accuracy is partially caused by the skew. Hence, we study the dataset exploratively *before* classification, to identify subpopulations that exhibit less skew, and exploratively *after* classification, to identify variables inside each subpopulation, which are associated to the outcome with high likelihood.

Pinheiro et al. perform association rule discovery on patients with liver carcinoma [20]. The authors point out that early detection of liver cancer may help reducing the five-year mortality rate (which is currently 86 % [20]), but early detection is difficult, because in the onset of a liver carcinoma, the patient often observes no symptoms [20]. Pinheiro et al. leverage the association rule miner FP-growth [10] to discover high-confidence association rules and high-confidence classification rules with respect to mortality in a liver cancer patients dataset. We also consider association rules promising for the analysis of medical data, because they are easy to compute and deliver results that are understandable by humans. Therefore, we also use association rules as baseline mining method, though for epidemiological data and for classification rather than mortality prediction. To use association rules for classification, we specify that the rule consequent should be the target variable (the rules are then called "classification rules"; we use this term hereafter).

Next to its advantages, association rule discovery (and classification rule discovery) has an inherent disadvantage: namely it generates large or even huge numbers of rules,

among which the expert has to search for the truly interesting ones. Scholars have often proposed visualization as a remedy, and there is substantial research on comprehensible visual representations of large numbers of association rules. For example, Hahsler and Chelluboina group association rules' antecendents by their shared attributes to create a grid, where more important rules are displayed as circles; a circle's size and color reflects the rule's support and lift values, respectively [8]. This tool further allows the user to zoom into interesting areas of the visualized set of rules [8].

In [24], Sekhavat and Hoeber stress that "... in spite of the advantages of previous works in visualizing association rules, the most common problem they encounter is their inability to handle a large collection of rules. In general, this results in occlusion and screen clutter problems due to the need to compress the visual representation into a single view." They propose SARV (Scalable Association Rule Visualisation), an interactive panel containing (i) a table-like grid view where rows represent rule antecedents, columns represent rule consequents, and a cell captures a single rule, colored (in grey-scale) according to the rules support value – this view is interactive, so that potentially interesting rules can be selected; (ii) a graph view for the visual exploration of the relationships between rules selected in the grid view; (iii) a textual view for displaying a rules support and confidence values [24]. The InteractiveRuleMiner of our approach also contains visualization aids and allows the medical expert to select rules for inspection. However, a medical expert is less interested in a rule's support and confidence, and more on how the rule manifests itself inside each class. Hence, our InteractiveRuleMiner shows graphically how the instances supporting each rule are distributed among the classes, and it allows the medical expert to sort and juxtapose rules according to different criteria.

For mining on medical data, there are many supervised learning methods to choose from, next to association rules. Similarly to [17], who study warafin medication among elderly patients, and similarly to our earlier work on the malignance of breast tumors [7], we also use decision trees for classification. The SHIP data we study have been so far analyzed mainly with regression methods, including Cox regression [18, 11], generalized linear and mixed models [1], generalized estimating equations, structural equation models, median and Poisson regression [28]. So, what should be the method of choice? Pombo et al. compare 39 studies, where supervised learning methods had been used for pain assessment in clinical decision support [21]. The learning task common to these studies, had been to predict whether pain treatment would be necessary. The studies had used rule-based algorithms, artificial neural networks, nonstandard set theory, and statistical learning algorithms. Pombo et al. compare them on accuracy and demonstrate that none outperforms all others [21]. This indicates that the choice of the classification algorithms depends on the medical problem and should also be dictated by further requirements of the medical task. In our study, we use decision trees, regression trees and classification rules for supervised learning, because these methods deliver models that can be directly interpreted by a human expert. Our InteractiveRuleMiner tool is a further aid for model inspection and interpretation.

In [7], we have proposed a workflow for data preparation and classification on a patient cohort, paying emphasis on the identification of variables that separate well among the classes. The work reported here is based on a simpler workflow, but uses more methods and an elaborate data partitioning step to deal with severe class imbalance and with different class distributions in the partitions.

3. Materials and Methods

The data used for population partitioning and class separation for hepatic steatosis come from the Study of Health in Pomerania (SHIP). We describe the data in subsection 3.1. In subsection 3.2, we explain what motivated us to partition the data and present our partitioning steps. Then, we discuss the methods we used for class separation on the whole dataset and on the partitions (cf. 3.3).

3.1. The Dataset

The Study of Health in Pomerania (SHIP) encompasses two independent cohorts. Cohort inclusion criteria were age from 20 to 79 years and main residency in the study region. Baseline examinations for the first cohort were performed between 1997 and 2001 (SHIP-0, n= 4308). Followup examinations were done in 2002-2006 (SHIP-1, n= 3300) and 2008-2012 (SHIP-2, n= 2333). Baseline information for a second, independent cohort (SHIP-TREND-0, n= 4420) was collected in 2008-2012.

For our analysis, the target variable is derived from the participant's liver fat concentration computed with magnetic resonance imaging (MRI). Preliminary MRI results are currently available for 578 SHIP-2 participants. These MRI results are preliminary, because the MR technique used to compute the values of the original target variable mrt_liverfat_s2 included a correction of $T2\star$ effects, but other confounders for chemical shift MR fat quantification, such as multi-spectral complexity of fat and T1 effects were ignored. However, as shown in [15], these latter confounders behave linearly with respect to the target. Through conservative choice of the cut-off value (see below) and discretization, this problem was partially amended, so that the mining methods still behave reliably.

We use the data of these participants for classifier learning, while our interactive InteractiveRuleMiner (cf. section 5) also juxtaposes these data to the data of the remaining 1755 participants, for whom the MRI recordings were not made available. We derive the target variable from the result of the MRI report through discretization. In particular, participants with a liver fat concentration of no more than 10 % are mapped to class A ("negative" class,

corresponds to the absence of the disorder); values greater than 10 % and lower than 25 % are mapped to class B (increased liver fat / fatty liver tendency); values greater than 25 % are mapped to class C (high liver fat). We consider classes B and C as "positive". The cut-off of 10 % is higher than the value of 5% suggested in [14] for separation the between healthy subjects and subjects with hepatic steatosis. However, the primary interest from the medical perspective was the identification of important variables for individuals that are likely to be ill. The selection of a high cut-off value made the mining task substantially more challenging, as is explained hereafter.

Out of the 578 participants, 438 belong to class A (≈ 76 %), 108 to B (≈ 19 %) and 32 to C (≈ 6 %).

Next to the target variable, the dataset contains 66 variables extracted from participants' questionnaire answers and medical tests (cf. [27]). They are variables on sociodemographics (gender, age etc), variables on consumption behaviour (e.g. alcohol and nicotin), SNPs, variables extracted from laboratory data (e.g. sera concentrations), and two variables on the results of the liver ultrasound – `stea_s2` and `stea_alt75_s2`. Both variables take symbolic values that reflect the likelihood that the participant has fatty liver; the latter is a combination of the former and the ALAT recording for the participant; details are in [27].

Some values of `stea_s2` and `stea_alt75_s2` are correlated to the target variable, in the sense that they occur often under the classes A or C, but they are not adequate for class separation. Since MRI was performed only on SHIP-2, while liver ultrasound recordings are also available in SHIP-0, we are particularly interested in identifying additional predictive features for subpopulations characterized by specific values of the two liver sonography variables.

Almost all variables we mention hereafter have the suffix `_s2`. This stands for values recorded in the SHIP-2 followup, as opposed to SHIP-0 and SHIP-1. Exceptions are gender, highest school degree and the 10 SNP variables.

3.2. Partitioning the Dataset into Subpopulations

Our decision for partitioning before classification was motivated by the observation that the dataset is imbalanced with respect to gender (314 women, 264 men). Our first step (cf. 3.2.1) is the investigation of the class distributions in the two partitions on gender, whereupon we see that the distributions are very different, most notably with respect to class B. The second step (cf. 3.2.2) is the investigation of the class distributions on age, whereupon we detect that age has an influence on the posteriors for the partition PartitionF but not for the partition PartitionM. The third step of our approach is then the identification of the cut-off point for age: we introduce a heuristic that identifies the age value which minimizes the standard deviation with respect to the target variable. Supervised learning (cf. 3.3) is then performed separately on the partitions of male and of female participants, while an additional learner is built for the subpopulation of older female participants (aged above 52, the cut-off point for age) .

3.2.1. Partitioning the Original Dataset on Gender

In Table 1, we depict the absolute and relative distribution of the target variable: for the whole dataset, for the subset of female participants and the subset of male participants. These subsets are called PartitionF and PartitionM hereafter. The gender-based separation leads to very different target variable distributions: the portion of A participants in PartitionM is much lower than in PartitionF (69 % vs. 81 %). The last entry of Table 1 is discussed in 3.2.2.

		absolute			relative		
Partition	total	A	B	C	A	B	C
All	578	438	108	32	76 %	19 %	6 %
PartitionM	264	183	66	15	69 %	25 %	6 %
PartitionF	314	255	42	17	81 %	13 %	5 %
F:age>52	183	131	40	12	72 %	22 %	7 %

Table 1: Class Distribution on Gender

The disparity of the distributions on gender becomes more clear in Figure 1, where the values for median, first and third quartile are different in the two partitions: (i) the median of PartitionF is lower than the median of PartitionM (3.7 % vs. 5.9 %) and (ii) the difference between the first and the third quartile for PartitionF is smaller than for PartitionM (4.2 % vs. 7.7 %). The maximum length of the whiskers in this boxplot is defined by $1.5 \cdot (q_3 - q_1)$, where q_1 is the value of the first quartile and q_3 is the value of the 3rd quartile for the distribution of mrt_liverfat_s2: a participant with a liver fat concentration outside the whiskers is then termed an "outlier". There are more outliers with very high fat liver concentrations in PartitionF than in PartitionM (47 vs. 15). Additionally, if we observe the number of outliers with a fat liver concentration of more than 23.3 %, i.e. at the approximate position of the upper whisker in PartitionM, we see that (iii) the absolute number of female participants with even higher concentration is larger than the corresponding number of male participants (19 vs. 15). These findings lead us to further investigations of the class distribution in the partition PartitionF.

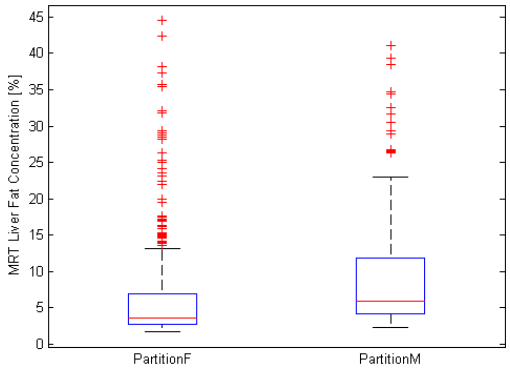

Figure 1: Boxplots for PartitionF and PartitionM

3.2.2. Splitting the Set of Female Participants on Age

We performed an analysis of the class distribution in the partition PartitionF on the variable age_s2. For simplicity of notation, we refer to this variable as "age" hereafter.

To understand how age affects the class distribution, we introduce a heuristic that determines the age value at which PartitionF should be split (into two bins), so that the standard deviations of the liver fat concentrations in each bin are minimized. Let $splitAge$ denote the cutoff value and $X_y = \{x \in \text{PartitionF} | \text{age of } x \leq splitAge\}$, $X_z = \{x \in \text{PartitionF} | \text{age of } x > splitAge\}$ denote the bins. Further, let n be the cardinality of $X_y \cup X_z$ i.e. of PartitionF. Then, we define the Sum of weighted Standard Deviations ($SwSD$) as

$$SwSD\,(X_y, X_z) = \frac{|X_y|}{n}\sigma(X_y) + \frac{|X_z|}{n}\sigma(X_z) \qquad (1)$$

where $|X_i|$ is the cardinality of X_i and $\sigma(X_i)$ the standard deviation of the original liver fat values, for $i = y, z$. Our heuristic selects X_y, X_z in such a way that $SwSD()$ is minimized. For PartitionF, the minimum value was 7.44 at the age of 52, i.e. close to the onset of menopause.

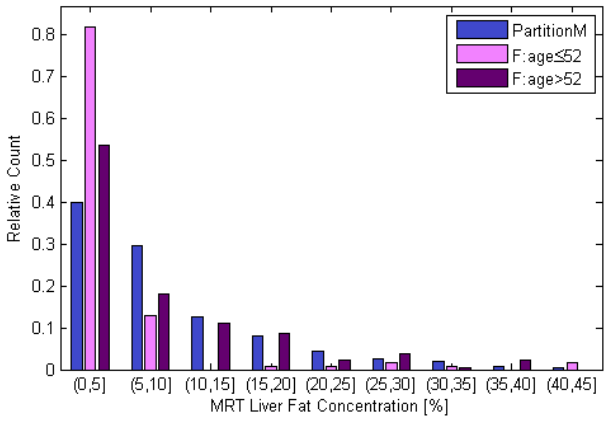

Figure 2: Class Distribution on male participants and on females older than 52 and younger: the horizontal axis shows the liver fat concentration in bins of 5 %, while the vertical axis shows the ratio of participants in each bin among the participants in the whole partition (PartitionM , F:age≤52 and F:age>52 respectively)

The histogram of Figure 2 depicts the differences in the class distributions at the cut-off position age of 52. Next to PartitionM, we show the subpartitions F:age>52 and F:age≤52 of PartitionF. The liver fat concentration values (continuous variable mrt_liverfat_s2) are put in bins with a step of 5 %. We observe that the absolute majority of female participants in F:age≤52 have no more than 5 % liver fat concentration and 90 % of them have no more than 10 %, i.e. belong to the negative class A. In contrast, more than 10 % of the participants in F:age>52 have a liver fat concentration of more than 10 % and at most 15 %, a bit less than 10 % have a liver fat concentration of more than 15 % and at most 20 %, i.e. around 20 % of the participants in F:age>52 belong to the positive class B.

The last entry of Table 1 depicts partition F:age>52, listing the number of participants in each class. It is clear that F:age>52 contains most of the positive female participants. Hence, we analyze this partition separately.

3.3. Classification Methods

For the classification of the cohort participants we concentrate on algorithms that deliver human-understandable models, since we want to identify predictive *"features", i.e. variables and values/ranges* in the models. So, we consider decision trees, classification rules and regression trees.

3.3.1. Decision Trees

We employ the J4.8 decision tree classification algorithm (equivalent to the C4.5 algorithm [23]) of the Waikato Environment for Knowledge Analysis (Weka) library [9]. This algorithm builds a tree gradually, by splitting each node (subset of the dataset) on the variable that maximizes information gain within that node. The original algorithm operates only on variables that take symbolic values and creates one child node per value. However, the implementation in the Weka library also provides an option that forces the algorithm to always create exactly two child nodes: one for the best separating value and one for all other values. We use this option in our experiments, because it delivers trees of better quality. Moreover, the Weka algorithm also supports variables that take numeric values: a node is split into two child nodes by partitioning the valuerange of the variable into two intervals. [1]

To deal with the skewed distribution, we consider following classification variants:

Naive: The problem of imbalanced data is ignored.

InfoGain: We keep only the top-30 of the 66 variables, by sorting the variables on information gain towards the target variable.

Oversampling: We use SMOTE to resample the dataset with minority-oversampling [4]: for class B, 100 % new instances are generated, for class C 300 % new instances are generated, resulting in following distribution A:438, B:216, C:128.

CostMatrix: We prefer to misclassify a negative case rather than not detecting a positive case, so we penalize false negatives (FN) more than false positives (FP). We use the cost matrix depicted in Table 2.

[1] It must be noted that all variables of our data subset of SHIP-2 participants were originally modeled as numbers. However, some of these variables (e.g. gender or `stea_s2`) should be better observed as symbols rather than numbers, as ordering and mathematical operations (like mean and standard deviation) do not make sense on them. We have therefore re-declared such variables as symbolic.

Classified as ⇒	A	B	C
A	0	1	2
B	2	0	1
C	3	2	0

Table 2: Cost Matrix penalizing misclassification under class skew

3.3.2. Classification Rule Discovery

Classification rules are association rules, the consequent of which consists of one of the classes. For example, consider following association rules, where & stands for "AND":
• stea_s2 = 1 & gx_rs11597390 = 1 & age_ship_s2 > 59 → class = B
• som_waist_s2 ≤ 80 → stea_s2 = 1

The features (variables and value ranges) at the left of the arrow constitute the rule's *antecedent*, while the feature at the right is the rule's *consequent*. The first rule in the example is a *classification rule* referring to class B. The second rule is not a classification rule, since the variable in the consequent is not a class.

The reader may notice that variables with real values (e.g. age, waist) are restricted within a specific range. This range is chosen by the mining algorithm in such a way as to maximize the support of the rule within the cohort and the confidence of the rule's consequent given the antecedent. The first rule in the example belongs to the results on PartitionF (see Table 4, described in the next section): out of the 20 participants supporting the antecedent, 17 belong to class B. Note that we use the expressions *"[number] participants support rule [ruledescription] "* and *"[number] participants support the rule's* antecedent*"*.

For classification rule discovery we use the Weka algorithm HotSpot. For each class, this algorithm determines the rules with the best confidence *and* the optimal boundary values for the features in the antecedent. We have wrapped the algorithm into a mechanism that selects for each class only rules supported by at least τ participants. In our experiments, we use $\tau = \frac{1}{3}$, but our interactive tool (section 5) allows the expert to set this threshold freely.

3.3.3. Regression Trees

We learn regression trees on the original continuous target variable mrt_liverfat_s2 using the Weka algorithms REPTree (Error Pruning Tree), M5P and MP5Rules.

4. Experiments and Findings

We learned models on the full dataset and on each partition for each of the classification variants described in 3.3.1 and for HotSpot rules. We also studied tree regression on the complete dataset. However, the predicitive power of the regression trees was very poor: either the regression tree consisted solely of one node with the mean of the complete dataset as predictor, i.e. the regression algorithm could not find appropriate split attributes, or two or more leaf nodes had very similar prediction values, whereupon interpreting the tree was very hard. We therefore focussed on classification trees and classification rules. We report on our findings with these methods hereafter.

4.1. Results of Decision Tree Classifiers

For the evaluation of decision tree classifiers, we consider accuracy, i.e. the ratio of correctly classified participants to all participants in the selected partition (or full dataset), the specificity and sensitivity, and the F1-Score, i.e. the harmonic mean between precision and recall. For specificity, precision and recall, we consider as positive class the two classes B and C together.

The performance of the decision tree classifiers on the whole dataset was poor: *Oversampling* achieved best performance with an accuracy of ca. 80 % but an F1-score of 62 %. The best decision trees were found for partition F:age>52, followed by those for PartitionF, then PartitionM. The large discrepancy between accuracy and F1-score appears also in the models of the partitions, underlying that accuracy scores are unreliable in such a skewed distribution. Therefore, we do not report on accuracy hereafter.

On partition F:age>52, the overall best decision tree is achieved by the *Oversampling* variant. On the larger PartitionF, best performance was achieved by the decision tree produced with the *InfoGain* variant, while the best decision tree on PartitionM was built with the *CostMatrix* variant. The specificity and sensitivity values for these trees are shown in Table 3, while the trees themselves are depicted in Figures 3 – 5 respectively and discussed in subsection 4.3.

Partition	Decision Tree variant	Sensitivity	Specificity	F1-score
F:age>52	*Oversampling*	63.5 %	93.9 %	81.5 %
PartitionF	*InfoGain*	52.4 %	94.9 %	69.7 %
PartitionM	*CostMatrix*	38.3 %	86.3 %	53.0 %

Table 3: Best decision trees for the three partitions: best separation is achieved in F:age>52; PartitionM is the most heterogeneous one, the performance values are lowest

Table 3 indicates that the decision tree variants perform differently on different partitions. It is natural that *Oversampling* is beneficial for F:age>52, because it partially compensates the skew problem. PartitionM is very heterogeneous, all classifiers perform poorly on it. So, we expect most insights from the decision trees on F:age>52 and PartitionF, where better separation is achieved.

4.2. Discovered Classification Rules

The classification rules found by Hotspot on the whole dataset were conclusive for class A but not for the positive classes B, C. These rules are not useful for diagnostic purposes, so we do not report on them.

The classification rules found on the partitions were more informative. However, classification rules with only one feature in the antecedent had low confidence. To ensure high confidence, we restricted the output on rules

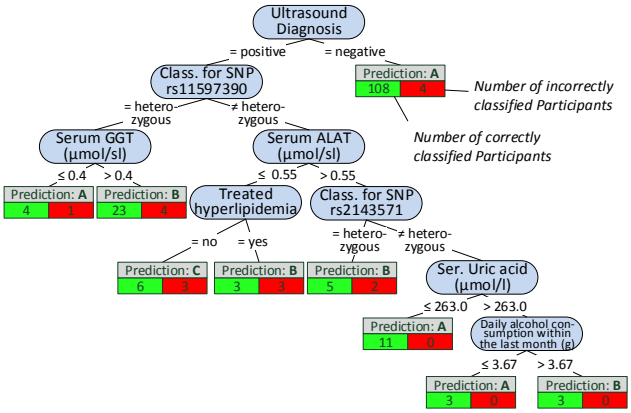

Figure 3: Best Decision Tree for F:age>52, achieved by the variant *Oversampling*

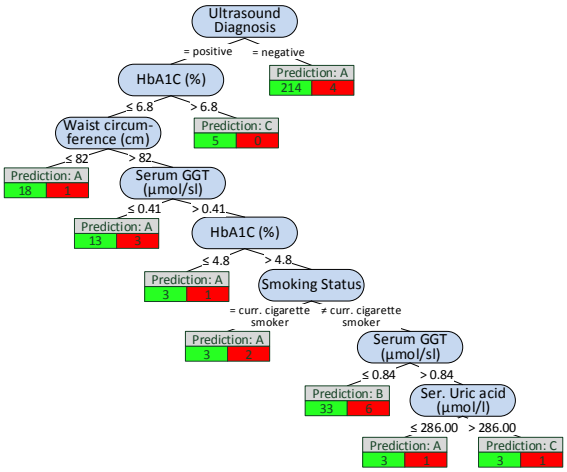

Figure 4: Best Tree for PartitionF, achieved by the variant *InfoGain*

with at least two features in the antecedent. To ensure a still high support, we allowed for at most three features. A selection of high confidence and high support rules for each partition and class are shown in Tables 4 – 6, respectively. We describe the most important features in the antecedent of these rules in the next subsection, together with the most important features of the best decision trees.

4.3. Important Features for Each Subpopulation

The most important features in the decision trees of Figures 3 – 5 are those closer to the root. For better readability, the tree nodes in the figures contain short descriptions instead of the original variable names.

On all three decision trees, the root node is the ultrasound diagnosis variable `stea_s2`. A negative ultrasound diagnosis points to the negative class A, but a positive ultrasound diagnosis does not directly lead to one of the positive classes B, C. The decision trees of the three partitions differ in the nodes placed near the root.

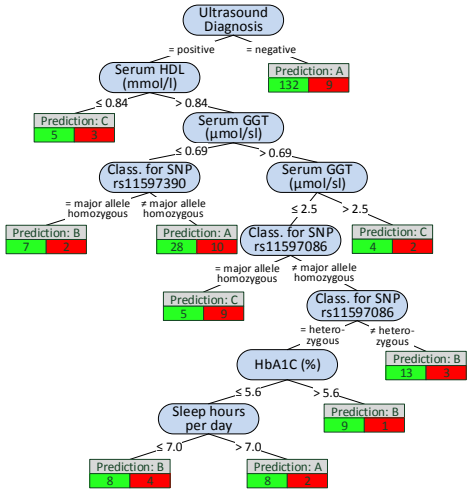

Figure 5: Best Tree for PartitionM, achieved by the variant *Cost-Matrix*

4.3.1. Important Features for PartitionF

In the best decision tree of PartitionF (cf. Figure 4) we observe that if the ultrasound report is positive *and* the HbA1C concentration is more than 6.8 %, the class is C. The classification rules with high support and confidence on Table 4) point to further interesting features: a waist circumference of at most 80 cm, a BMI of no more than 24.82 kg/m^2, a hip circumference of 97.8 cm or less characterize participants of the negative class. All 6 participants having a serum glucose concentration greater than 7 mmol/l *and* a TSH concentration greater than 0.996 mu/l belong to class C. Further, severe obesity (a BMI value of more than 38.42 kg/m^2) points to class C with high confidence - but only in combination with other variables.

4.3.2. Important Features for F:age>52

In contrast to the best tree for PartitionF, the best decision tree for the subpartition F:age>52 (cf. Figure 3) also contains nodes with SNPs, indicating potentially genetic associations to fatty liver for these participants. Classification rules with high support and confidence for class B also contain SNPs, as can be seen on Table 5.

Similarly to PartitionF, high BMI values point to a positive class when combined with other features: on Table 5, we see that all four participants with `stea_alt75_s2` = 3 (i.e. a positive ultrasound diagnosis combined with a critical ALAT value) and a BMI larger than 38.42 kg/m^2 belong to class C. A similar association holds for `stea_alt75_s2` = 3 combined with a high waist circumference (> 124 cm). 19 out of 20 participants in class B having a positive ultrasound diagnosis, a genetic marker `gx_rs11597390` = 1 and a serum HDL concentration of at most 1.53 mmol/l.

4.3.3. Important Features for PartitionM

The role of the ultrasound report in predicting the negative class is the same for PartitionM (cf. Figure 5) as for

Rule antedecedent			Participants supporting antecedent	Target class of the rule	Participants supporting the rule		Rule confidence
Variable 1	Variable 2	Variable 3			absolute number	percentage in class	
som_waist_s2 $\leq$ 80	-	-	132	A	132	52 %	100 %
som_bmi_s2 $\leq$ 24.82	-	-	109	A	109	43 %	100 %
som_huef_s2 $\leq$ 97.8	-	-	118	A	117	46 %	99 %
stea_s2 = 0	-	-	218	A	214	84 %	98 %
stea_alt75_s2 = 0	-	-	202	A	198	78 %	98 %
stea_s2 = 1	gx_rs11597390 = 1	age_ship_s2 > 59	20	B	17	40 %	85 %
stea_alt75_s2 = 1	hrs_s_s2 > 263	age_ship_s2 > 59	20	B	17	40 %	85 %
stea_alt75_s2 = 1	hrs_s_s2 > 263	ldl_s_s2 > 3.22	20	B	17	40 %	85 %
stea_s2 = 1	age_ship_s2 > 66	tg_s_s2 > 1.58	17	B	14	33 %	82 %
stea_s2 = 1	age_ship_s2 > 64	hrs_s_s2 > 263	17	B	14	33 %	82 %
gluc_s_s2 > 7	tsh_s2 > 0.996	-	6	C	6	35 %	100%
som_bmi_s2 > 38.42	age_ship_s2 $\leq$ 66	asat_s_s2 > 0.22	6	C	6	35 %	100%
som_bmi_s2 > 38.42	sleeph_s2 > 6	blt_beg_s2 $\leq$ 38340	6	C	6	35 %	100%
som_bmi_s2 > 38.42	sleeph_s2 > 6	stea_s2 = 1	6	C	6	35 %	100%
hrs_s_s2 > 371	sleepp_s2 = 0	ggt_s_s2 > 0.55	6	C	6	35 %	100%

Table 4: Best HotSpot Classification Rules ($maxLength = 3$) for PartitionF (excerpt)

Rule antedecedent			Participants supporting antecedent	Target class of the rule	Participants supporting the rule		Rule confidence
Variable 1	Variable 2	Variable 3			absolute number	percentage in class	
crea_u_s2 $\leq$ 5.39	stea_s2 = 0	-	75	A	75	57 %	100 %
crea_u_s2 $\leq$ 5.39	stea_alt75_s2 = 0	-	72	A	72	55 %	100 %
som_waist_s2 $\leq$ 80	-	-	54	A	54	41 %	100 %
som_bmi_s2 $\leq$ 24.82	-	-	50	A	50	38 %	100 %
crea_u_s2 $\leq$ 5.39	ggt_s_s2 $\leq$ 0.43	-	50	A	50	38 %	100 %
stea_s2 = 1	ggt_s_s2 > 0.48	ggt_s_s2 $\leq$ 0.63	15	B	15	38 %	100 %
stea_s2 = 1	gx_rs11597390 = 1	hdl_s_s2 $\leq$ 1.53	20	B	19	48 %	95 %
stea_s2 = 1	gx_rs11597390 = 1	fib_cl_s2 > 3.4	15	B	14	35 %	93 %
crea_s_s2 $\leq$ 61	som_waist_s2 > 86	stea_s2 = 1	15	B	14	35 %	93 %
stea_s2 = 1	gx_rs11597390 = 1	hrs_s_s2 > 261	20	B	18	45 %	90 %
som_bmi_s2 > 38.42	age_ship_s2 $\leq$ 66	-	4	C	4	33 %	100%
som_bmi_s2 > 38.42	stea_alt75_s2 = 3	-	4	C	4	33 %	100%
som_huef_s2 > 124	stea_alt75_s2 = 3	-	4	C	4	33 %	100%
som_waist_s2 > 108	gluc_s_s2 > 6.2	-	4	C	4	33 %	100%
stea_alt75_s2 = 3	som_bmi_s2 > 37.32	-	4	C	4	33 %	100%

Table 5: Best HotSpot Classification Rules ($maxLength = 3$) for F:age>52 (excerpt)

PartitionF. As with the best tree for F:age>52, the best tree for PartitionM contains nodes with SNPs and serum GGT value ranges. Such features are also in the antecedent of top Hotspot rules (cf. Table 6): a Serum GGT concentration of more than 1.9 μmol/sl in combination with creatinine concentration of at most 90 mmol/l or a thromboplastin time ratio (quick_s2) of more than 59 % point to class C. Similarly, positive ultrasound diagnosis and a serum HDL concentration not exceeding 0.84 mmol/l point to class C.

4.3.4. Conclusion on important features

The decision trees and classification rules give insights into features that seem diagnostically important. However, the medical expert needs additional information to decide whether a feature is worth further investigation. In particular, decision trees highlight the importance of a feature only in the context of the subtree it is located; a subtree describes a subpopulation that is usually very small. In contrast, classification rules return information on larger subpopulations. However, these subpopulations may overlap; for example, the first four rules on class C for PartitionM (cf. Table 6) may refer to the same 6 participants. Moreover, unless a classification rule has a confidence close to 100 %, there may be participants in the other classes that also support it. Hence, to decide whether the features in the rule's antecedent deserve further investigation, the expert also needs insights on the rule's statistics for the other classes as well. To assist the expert in this task, we propose InteractiveRuleMiner, a tool that discovers classification rules for each class *and* delivers information on the statistics of these rules for all classes.

5. Interactive Rule Miner

Classification rules, as those depicted in Tables 4 - 6 can provide valuable insights on potentially prevalent features (variables and their value ranges) for different sub-

Rule antedecedent			Participants supporting antecedent	Target class of the rule	Participants supporting the rule		Rule confidence
Variable 1	Variable 2	Variable 3			absolute number	percentage in class	
stea_alt75_s2 = 0	-	-	106	A	101	55 %	95 %
stea_s2 = 0	-	-	138	A	131	72 %	95 %
ggt_s_s2 ≤ 0.52	-	-	79	A	73	40 %	92 %
hrs_s_s2 ≤ 310	ggt_s_s2 ≤ 0.77	-	81	A	74	40 %	91 %
som_waist_s2 ≤ 90.8	-	-	79	A	72	39 %	91 %
som_huef_s2 > 108.1	age_ship_s2 > 39	crea_u_s2 > 7.59	28	B	22	33 %	79 %
som_bmi_s2 > 32.29	hdl_s_s2 > 0.94	ATC_C09AA02_s2 = 0	29	B	22	33 %	76 %
som_bmi_s2 > 32.29	hgb_s2 > 8.1	gout_s2 = 0	29	B	22	33 %	76 %
som_waist_s2 > 109	sleeph_s2 ≤ 8	jodid_u_s2 > 9.44	29	B	22	33 %	76 %
som_huef_s2 > 108.1	hdl_s_s2 > 0.97	crea_u_s2 > 5.38	29	B	22	33 %	76 %
ggt_s_s2 > 1.9	crea_s_s2 ≤ 90	quick_s2 > 59	6	C	6	40 %	100%
ggt_s_s2 > 1.9	crea_s_s2 ≤ 90	chol_s_s2 > 4.3	6	C	6	40 %	100%
ggt_s_s2 > 1.9	crea_s_s2 ≤ 90	fib_cl_s2 > 1.9	6	C	6	40 %	100%
ggt_s_s2 > 1.9	crea_s_s2 ≤ 90	crea_u_s2 > 4.74	6	C	6	40 %	100%
ggt_s_s2 > 1.9	tg_s_s2 > 2.01	som_waist_s2 > 93.5	6	C	6	40 %	100%

Table 6: Best HotSpot Classification Rules ($maxLength = 3$) for PartitionM (excerpt)

populations of the cohort under study. However, as can be easily seen in Tables 4 - 6 the number of rules produced is large, the contents of the rules overlap and some features are present under each of the three classes. Hence, the medical expert needs inspection aids to decide which rules are informative and which features should be studied further. Our InteractiveRuleMiner is an interactive mining tool that allows the expert to (a) discover classification rules subject to frequency constraints, inspect the frequency of those rules (b) towards each class and (c) against the *unlabeled part of the cohort*, and (d) study the statistics of each rule for the values of selected variables. We describe these functionalities below, referring to the screenshot on Figure 6.

The user interface of InteractiveRuleMiner has six areas. The fill-in area "Settings" (upper left) allows the medical expert to specify the criteria for rule mining before pressing the button "Build Rules"; below this area appear then the discovered rules. The area "Sorting preference" (next to the area "Settings") allows the expert to specify whether the output should be sorted on confidence of the rule's consequent (selected class) towards the antecedent, on support of the whole rule, or rather alphabetically for better overview of overlapping rules.

The mining criteria concern the dataset (choosing between the whole dataset versus one of the partitions), the class for which rules should be chosen (drop-down list `Class`) and the constraints with respect to this class: Min Value Count , Max Rule Length, Max Branching Factor and Min Improvement used as follows. The Min Value Count is either an absolute number or a percentage over the number of cohort participants in the selected class. The miner constructs the rules by gradually adding features (variables and value intervals) in the antecedent, while preserving the Min Value Count constraint. When the miner adds features gradually, it has several candidates to choose from. For example, consider the rule "stea_s2 = 1 & `hrs_s_s2` > 263 ⟶ B" (see 3^{rd} rule in the output listed on the lower left corner of Figure 6), which has a confidence c: 0.55 (31 out of 56 participants) towards the target class B. The miner considers all features that are supported by Min Value Count among the participants supporting "`stea_s2` = 1", but rejects those that improve the confidence of the target class (consequent) for less than Min Improvement. The remaining candidates are sorted on support and the top ones are chosen - as many as the Max Branching Factor. An example of such an expansion is "`stea_s2` = 1 & `gx_rs11597390` = 1" (see second rule in the output list on Figure 6); its confidence towards class B is c:0.58. The miner extends antecedents with additional features up to the Max Rule Length threshold.

The output list of an execution run (area below the "Settings") is scrollable and interactive. When the expert clicks on a rule, the top middle area "Summary Statistics for selected Rule" is updated. The first row shows the distribution of the cohort participants among the classes for the whole dataset (or partition!), while the second row shows how the participants supporting the rule's antecedent (column "All" in the second row) are distributed among the classes. Hence, the expert can specify the discovery of classification rules for one of the classes and then study how often the antecedent of each rule appears among the participants in the other classes. Clearly, a rule that is supported by most of the participants of the selected class (class B on Figure 6) is interesting, but the rarer it appears among the participants of the other classes the more interesting it is. For example, the rule antecedent "stea_s2 = 1 & hrs_s_s2 > 263" is supported by 31 + 14 positive participants and 11 negative ones, and is most frequent for the positive class B (cf. histogram in Figure 6). In the top middle area, we see that 73.8 % of the class B participants support this rule (31 out of 42), while this percentage drops to 4.3 % for the negative class A.

The areas of InteractiveRuleMiner described thus far de-

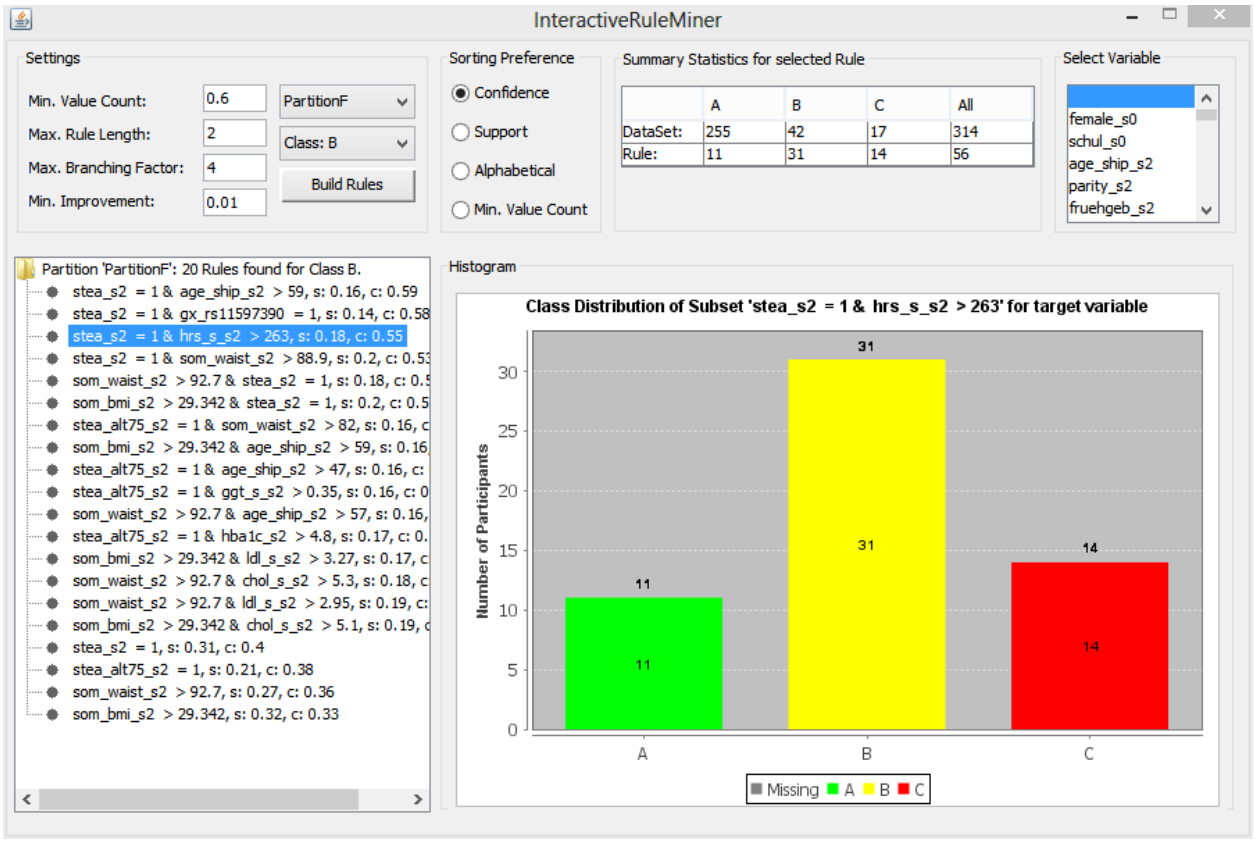

Figure 6: The InteractiveRuleMiner: Classification rules are discovered for PartitionF for class B and shown in the bottom left area. For the selected (marked) rule "stea_s2 = 1 & hrs_s_s2 > 263 → B", we see the distribution of the participants in the rule's antecedent among all three classes in absolute values (top middle area) as histogram on the values of the target variable (bottom right).

liver information on the rules found on labeled data. However, not all participants in the cohort have been subjected to liver MRI. Hence, it is also of interest to know the distribution of the unlabeled participants who support the antecedent of a given rule. The area "Histogram" can be used to this purpose: the expert chooses a further variable from the interactive area "Select variable" in the upper right corner and can then see how the values of these variable are distributed among the study participants - both the labeled ones and the unlabeled ones; the latter are marked as "Missing" in the histogram's legend. For plotting the histograms we make use of the free Java chart library JFreeChart [6].

On Figure 6, the expert has not chosen any variable, hence the target variable is used by default, and only the distribution of labeled participants is visible. On Figure 7, we see the distribution of both labeled and unlabeled participants supporting the antecedent of our example rule "stea_s2 = 1 & hrs_s_s2 > 263" with respect to the second part of the antecedent, i.e. the variable hrs_s2 for values 263 and more. The value distribution among the labeled participants shows that the likelihood of the negative classes drops as the values of hrs_s2 increase.

Figure 8 also shows a histogram for the example rule. However, this time the variable selected in the top right area is not part of the rule. We see a histogram of labeled and unlabeled participants for the variable ldl_s_s2. As on Figure 7, the histogram still shows how the labeled participants supporting the antecedent "stea_s2 = 1 & hrs_s_s2 > 263" are distributed among the three classes.

The distribution of participants on Figure 8 is bimodal with respect to the negative class. A study of further labels in this subpopulation, especially among participants in the second bin may help explaining the bimodality. Hence, the visualization of the participants' statistics for selected rules can deliver indications on subpopulations that should be monitored closer.

6. Discussion and Outlook

To date, analysis of population-based cohort data is mostly *hypothesis-driven*. In this study, we have presented a new mining workflow and interactive mining tool for *data-driven analysis* of population-based cohort data on the example of hepatic steatosis.

Our mining workflow encompasses steps (i) for the identification of subpopulations that exhibit different distributions with respect to the target variable, (ii) for the classification of each subpopulation, taking class skew into account, and (iii) for the identification of variables that

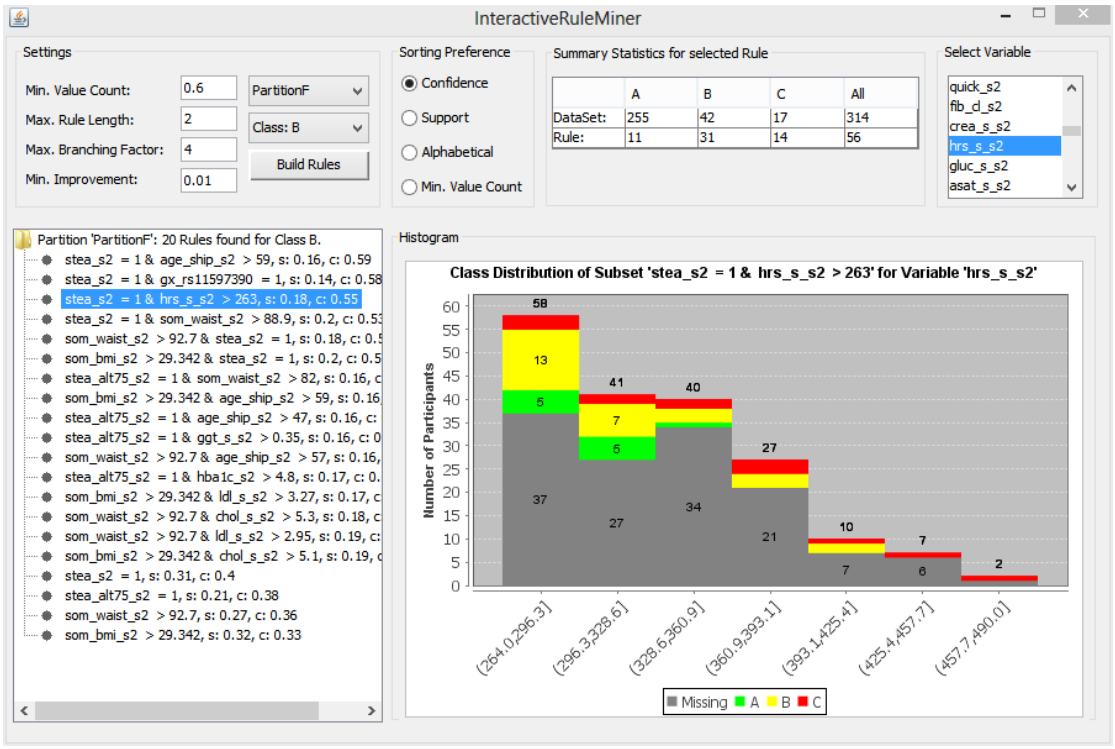

Figure 7: InteractiveRuleMiner: Classification rules as on Figure 6 with additional information on the distribution of the unlabeled participants supporting the rule antecedent "stea_s2 = 1 & hrs_s_s2 > 263" among the participants supporting only the second feature of the antecedent

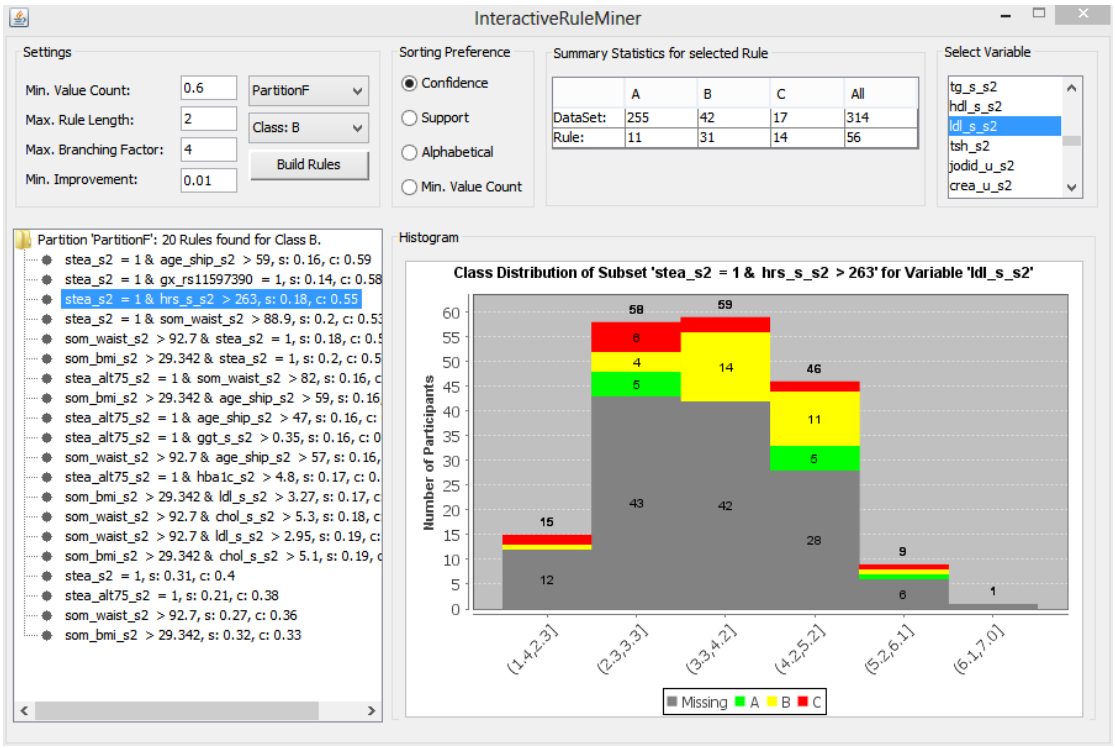

Figure 8: InteractiveRuleMiner: Classification rules as on Figure 6 with additional information on the value distribution for variable `ldl_s_s2` among the labeled and unlabeled participants supporting the rule antecedent "stea_s2 = 1 & hrs_s_s2 > 263"

are associated with the outcome. Our workflow has shown that it is imperative to (a) identify subpopulations before classification, so as to reduce skew, and to (b) drill into the derived models, so as to identify important variables and subpopulations worth studying further.

To assist the human expert in the latter objective (b), we have developed the InteractiveRuleMiner. The InteractiveRuleMiner is an interactive tool that allows the user to study classification rules closer and understand how the cohort participants who support each rule are distributed across the three classes. This inspection step is essential for the identification of not yet known associations between some variables the target outcome. These variables must be then investigated further - in hypothesis-driven studies. Hence, our mining workflow and InteractiveRuleMiner bear the potential of data-driven analysis in delivering insights on a multifactorial disorder and in hypothesis generation for hypothesis-driven studies.

With respect to the concrete multifactorial disorder, our findings verify the potential of our data-driven approach, since most variables in the top-positions of our decision trees and classification rules have been shown earlier in independent studies to be associated with hepatic steatosis[2]. In particular, indices of fat storage in the body (BMI, waist circumference) and the liver enzyme GGT were proposed by Bedogni et al. as a reliable "Fatty Liver Index" [2]. The SNPs rs11597390, rs2143571 and rs11597086 have been shown in [31] to be among the "Independent SNPs Associated with Liver-Enzyme Levels with Genome-wide Significance in Combined GWAS Analysis of Discovery and Replication Data Sets" (Table 3 of [31]. Concerning the impact of alcohol consumption, Baumeister et al. mention that "...the toxic effects of ethanol on the liver are well established ..." but point out that "The literature suggests an even greater role of overweight than heavy drinking in the accumulation of fat in the liver [3]" [1] (1st page); indeed, a variable related to alcohol consumption only appears in our decision tree on F:age>52 (cf. Figure 3) and not among our top classification rules, where we rather see variables associated with a person's weight and adipositas (cf. variables: `som_bmi_s2, som_huef_s2, som_waist_s2` in all figures and tables with findings). The subpopulation F:age>52 itself has been identified without using prior knowledge of the semantics of this subpopulation, but it is remarkable that the age of 52 is close to the onset of menopause – in [30] it has been shown that menopausal status is associated with hepatic steatosis. Our findings also verified a further fact known to the medical experts through independent observation: the sonography outcome (cf. variables: `stea_s2, stea_alt75_s2` in all figures and tables with findings) is associated with the liver fat concentration found by MRI, yet the ultrasonography alone does not predict hepatic steatosis [3, 2].

Did our approach deliver *new* insights? Our algorithms do not simply deliver variables associated to the outcome, but also identify the value intervals that are associated with a specific class, see e.g. the value intervals of the BMI associated with the class B for PartitionM (Table 6) and with the classes A and C for F:age>52 (Table 5). These intervals do not mean that a person with BMI inside the specific interval *does* belong to the corresponding class, but may rather serve as starting point for hypotheses-driven analyses.

A limitation of our approach concerns the interaction with the medical expert. The InteractiveRuleMiner has been designed with the demands of the medical expert in mind, but it has not yet been evaluated by medical experts. As a result, we maximized flexibility through a set of parameters, but it remains to be shown whether the presentation of these parameters is intuitive to the user. Also, the visualization aids (histograms) are rudimentary and must yet be evaluated with respect to the expert's intuition. To this purpose, we intend to set up an appropriate environment in which an expert will interact with the tool and will give us feedback. With respect to the concrete findings for hepatic steatosis, a shortcoming is that some confounders for chemical shift MR fat quantifications have not been corrected in the target variable we used (cf. beginning of section 3.1), hence the exact value intervals of the associated variables should not be taken at face value. This is not a shortcoming of the approach itself, though; our next step is to apply it on the corrected dataset.

Our approach allows for the inspection of subpopulations in two moments. Prior to data mining, we identify subpopulations that exhibit different class distributions. During data mining, our InteractiveRuleMiner highlights the subpopulation supporting each classification rule; these are *overlapping* subpopulations. The overlap among subpopulations is not necessarily a disadvantage, especially for very small subpopulations. However, working with overlapping datasets may be unintuitive to an application expert. Therefore, we investigate the potential of clustering methods for the identification of further subpopulations prior to data mining, so as to perform classification on each cluster independently.

Acknowledgements

Work of the first author was supported by the German Research Foundation project SP 572/11-1 "IMPRINT: Incremental Mining for Perennial Objects".

The Study of Health in Pomerania is part of the Community Medicine Research Network of the University Medicine Greifswald, funded by the State Mecklenburg-West Pomerania.

The data used in this work were made available through the cooperation SHIP/2012/06/D "Predictors of Steatosis Hepatis".

[2]The studies we cite hereafter did not consider exactly the same set of variables as we did, so it is natural that they did not find some of the associations we identified, and that they found associations that we did not find.

[3]In [1], this statement is supported through citation of [3].

We are indepted to Sebastian Baumgärtner for providing extensive literature on known predictors of hepatic steatosis and to Carsten Oliver Schmidt for his valuable comments on processing and destilling this literature.

References

[1] Baumeister, S. E., Völzke, H., Marschall, P., (...), Schmidt, C., Flessa, S., Alte, D., 2008. Impact of fatty liver disease on health care utilization and costs in a general population: A 5-year observation. Gastroenterology 134 (1), 85–94.

[2] Bedogni, G., Bellentani, S., Miglioli, L., Masutti, F., Passalacqua, M., Castiglione, A., Tiribelli, C., 2006. The Fatty Liver Index: a simple and accurate predictor of hepatic steatosis in the general population. BMC Gastroenterology 6 (33), 7 pages.

[3] Bellentani, S., Saccoccio, G., Masutti, F., Passalacqua, M., Croce, L. S., Brandi, G., Sasso, F., Cristanini, G., Tiribelli, C., 2000. Prevalence of and risk factors for hepatic steatosis in northern italy. BMC Gastroenterology 132 (2), 112–117.

[4] Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P., 2002. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16 (1), 321–357.

[5] Friedman, J. H., Meulman, J. J., 2003. Multiple additive regression trees with application in epidemiology. Statistics in Medicine 22 (9), 1365–1381.

[6] Gilbert, D., 2005-2013. JFreeChart (Free Java class library for creating charts). Last accessed December 03, 2013 from `http://www.jfree.org/jfreechart/index.html`.

[7] Glaßer, S., Niemann, U., Preim, B., Spiliopoulou, M., 2013. Can we Distinguish Between Benign and Malignant Breast Tumors in DCE-MRI by Studying a Tumor's Most Suspect Region Only? In: Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on. pp. 77–82.

[8] Hahsler, M., Chelluboina, S., 2011. Visualizing Association Rules in Hierarchical Groups. In: 42nd Symposium on the Interface: Statistical, Machine Learning, and Visualization Algorithms (Interface 2011). The Interface Foundation of North America.

[9] Hall, M., Frank, E., Holmes, J., Pfahringer, B., Reutemann, P., Witten, I. H., 2009. The WEKA Data Mining Software: An Update. SIGKDD Explorations 11 (1), 10–18.

[10] Han, J., Pei, J., Yin, Y., 2000. Mining frequent patterns without candidate generation. ACM SIGMOD Record 29 (2), 1–12.

[11] Haring, R., Wallaschofski, H., Nauck, M., Dörr, M., Baumeister, S. E., Völzke, H., 2009. Ultrasonographic hepatic steatosis increased prediction of mortality risk from elevated serum gammaglutamyl transpeptidase levels. Hepatology 50 (5), 1403–1411.

[12] Hingorani, A. D., van der Windt, D. A., Riley, R. D., (...), Sauerbrei, W., Altman, D. G., Hemingway, H., 2013. Prognosis research strategy (PROGRESS) 4: Stratified medicine research. BMJ: British Medical Journal 346 (e5793), 9 pages.

[13] Ittermann, T., Haring, R., Wallaschofski, H., (...), Meyer zu Schwabedissen, H. E., Rosskopf, D., Völzke, H., 2012. Inverse Association Between Serum Free Thyroxine Levels and Hepatic Steatosis: Results From the Study of Health in Pomerania. Thyroid 22 (6), 568–574.

[14] Kühn, J.-P., Evert, M., Friedrich, N., (...), Mensel, B., Hosten, N., Puls, R., 2011. Noninvasive quantification of hepatic fat content using three-echo dixon magnetic resonance imaging with correction for T2∗ relaxation effects. Investive Radiology 46 (12), 783–789.

[15] Kühn, J.-P., Hernando, D., Mensel, B., (...), Mayerle, J., Hosten, N., Reeder, S. B., 2013. Quantitative chemical shift-encoded MRI is an accurate method to quantify hepatic steatosis. Journal of Magnetic Resonance Imaging 39 (2), 8 pages.

[16] Lau, K., Lorbeer, R., Haring, R., (...), John, U., Baumeister, S. E., Völzke, H., 2010. The association between fatty liver disease and blood pressure in a population-based prospective cohort study. Journal of Hypertension 28 (9), 1829–1835.

[17] Liu, K. E., Lo, C.-L., Hu, Y.-H., 2014. Improvement of adequate use of warfarin for the elderly using decision tree-based approaches. Methods of Information in Medicine 53 (1), 47–53.

[18] Lorenz, M. W., Polak, J. F., Kavousi, M., (...), Ziegelbauer, K., Bots, M. L., Thompson, S. G., 2012. Carotid intima-media thickness progression to predict cardiovascular events in the general population (the PROG-IMT collaborative project): a meta-analysis of individual participant data. The Lancet 379 (9831), 2053–2062.

[19] Markus, M. R. P., Baumeister, S. E., Stritzke, J., (...), Wallaschofski, H., Völzke, H., Lieb, W., 2013. Hepatic Steatosis Is Associated With Aortic Valve Sclerosis in the General Population: The Study of Health in Pomerania (SHIP). Arteriosclerosis, Thrombosis, and Vascular Biology 33 (7), 1690–1695.

[20] Pinheiro, F., Kuo, M.-H., Thomo, A., Barnett, J., 2013. Extracting association rules from liver cancer data using the FP-growth algorithm. In: Computational Advances in Bio and Medical Sciences (ICCABS), 2013 IEEE 3rd International Conference on.

[21] Pombo, N., Arajo, P., Viana, J., 2014. Knowledge discovery in clinical decision support systems for pain management: A systematic review. Artificial Intelligence in Medicine 60 (1), 1–11.

[22] Preim, B., Klemm, P., Hauser, H., Hegenscheid, K., Oeltze, S., Toennies, K., Völzke, H., 2014. Visualization in Medicine and Life Sciences III. Springer, Ch. Visual Analytics of Image-Centric Cohort Studies in Epidemiology.

[23] Quinlan, J. R., 1992. Learning with continuous classes. In: 5th Australian Joint Conference on Artificial Intelligence. Vol. 92. pp. 343–348.

[24] Sekhavat, Y. A., Hoeber, O., 2013. Visualizing Association Rules Using Linked Matrix, Graph, and Detail Views . International Journal of Intelligence Science 3, 34–49.

[25] Stickel, F., Buch, S., Lau, K., (...), Wodarz, N., Völzke, H., Hampe, J., 2011. Genetic variation in the PNPLA3 gene is associated with alcoholic liver injury in caucasians. Hepatology 53 (1), 86–95.

[26] Targher, G., Day, C. P., Bonora, E., 2010. Risk of Cardiovascular Disease in Patients with Nonalcoholic Fatty Liver Disease. New England Journal of Medicine 363 (14), 1341–1350.

[27] Völzke, H., Alte, D., Schmidt, C. O., (...), Biffar, R., John, U., Hoffmann, W., 2011. Cohort Profile: The Study of Health in Pomerania. International Journal of Epidemiology 40 (2), 294–307.

[28] Völzke, H., Craesmeyer, C., Nauck, M., (...), John, U., Baumeister, S. E., Ittermann, T., 2013. Association of Socioeconomic Status with Iodine Supply and Thyroid Disorders in Northeast Germany . Thyroid 23 (3), 346–353.

[29] Völzke, H., Fung, G., Ittermann, T., (...), Rettig, R., Rao, B., Kroemer, H. K., 2013. A new, accurate predictive model for incident hypertension. Arteriosclerosis, Thrombosis, and Vascular Biology 31 (11), 2142–2150.

[30] Völzke, H., Schwarz, S., Baumeister, S. E., Wallaschofski, H., Schwahn, C., Grabe, H. J., Kohlmann, T., John, U., Dören, M., 2007. Menopausal status and hepatic steatosis in a general female population. Gut 56, 594–595.

[31] Yuan, X., Waterworth, D., Perry, J. R., (...), Frayling, T. M., Kooner, J. S., , Mooser, V., 2008. Impact of fatty liver disease on health care utilization and costs in a general population: A 5-year observation. Gastroenterology 134 (1), 85–94.

[32] Zhanga, C., Kodell, R. L., 2013. Subpopulation-specific confidence designation for more informative biomedical classification. Artificial Intelligence in Medicine 58 (3), 155–163.