Knowledge-Based Systems 240 (2022) 108088

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

A combinatorial optimization approach for multi-label associative classification

Yuchun Zou, Chun-An Chou *

Mechanical and Industrial Engineering, Northeastern University, Boston, MA 02115, USA

ARTICLE INFO

Article history:
Received 21 April 2021
Received in revised form 2 October 2021
Accepted 25 December 2021
Available online 31 December 2021

Keywords:
Associative classification
Combinatorial optimization
Interpretable data mining
Machine learning

ABSTRACT

Mining associations between variables corresponding to multiple class labels (or outcomes) is prevalent in various applied domains, such as medical diagnosis, text mining, e-commence, and social behavior analysis. While most associative classification algorithms have been developed to discover association rules of binary variables for single-label classification problem, there are limited methods designed for the problem, called multi-label classification, that accounts for multi-labels. In this study, we consider the multi-label classification problem as a multi-class classification problem and formulate it as a 0–1 integer optimization model. We then leverage combinatorial optimization and association rule techniques to solve this hard problem. More specifically, we propose a ranking metric for selecting and aggregating stronger rules to form an optimal multi-label classifier. The computational results for multiple real applications show that our algorithm is able to identify significant association rules between key variables and multiple labels, and in turn achieves a competitive classification performance compared to state-of-the-art machine learning methods such as logistic regression, decision trees, and random forest. Moreover, we design a user-interface tool interfaced with the developed algorithm and demonstrate a medical diagnosis problem to predict multiple high-risk subgroups during emergency care in practice.

1. Introduction

Classification is one of the fundamental tasks in data mining and machine learning, which aims to predict the class label of unseen samples based on input variables (or features). Multi-label classification (MLC) is a variant of supervised learning problem, which accounts for multiple label outcomes [1], and has received increasing attention in various real-world applications [2–5], such as medical diagnosis, text mining (assigning documents to several topics), e-commerce, social behavior analysis, bioinformatics, among others. Comparing to single label classification that classify samples by mapping variables into one class label, MLC could assign each sample to multiple class labels simultaneously. To tackle this multi-label classification problem, there are two common approaches such as problem transformation and algorithm adaptation. The former transforms a MLC problem into a series of binary classification problems that are settled by single class classifiers, which however cannot handle the dependencies between different labels well, and into a multi-class classification problem. The latter designs new algorithms based on state-of-art methods such as *k*-nearest neighbor rule, decision trees, or neural network to directly solve the MLC problem rather than converting the problem to a simpler one [6].

For data in a binary format, associative classification (AC) is the major concept that aims to solve classification problems using association rule mining to construct significant associations between variables and class label [7–10]. Most AC algorithms mine and filter strong association rules where target class is set to be the consequent or right-hand-side (RHS), including Classification Based on Associations (CBA) [8], Classification based on Multiple Association Rules (CMAR) [11], and Classification based on Predictive Association Rules (CPAR) [12]. Such methods were originally developed for solving single-label classification problems. To solve the MLC problem, some methods that adapt the AC concept have been developed, such as Lazy AC (LAC) [13], Multi-label Multi-class AC (MMAC) [14], Enhanced Multi-label Classifiers based Associative Classification (eMCAC) [15], Associative Rule Mining Technique for Multi-label Classification (ARM-MLC) [16], and Multi-objective optimization to combine Multiple Association rules into an interpretable Classification (MoMAC) [17]. However, these methods cannot promise to select more representative and strong rules that have a maximum coverage of samples and include the fewest features corresponding to the multi-label class outcome of interest to make insightful summary of the data.

In this paper, we study the MLC problem with two class labels in particular, where one is (primary) binary class label and the

* Corresponding author.
E-mail address: ch.chou@northeastern.edu (C.-A. Chou).

https://doi.org/10.1016/j.knosys.2021.108088
0950-7051/

Table 1
Literature summary of multi-label associative classification.

Year	Method	Index
2004	Multi-label multi-class AC (MMAC)	[14][18]
2010	Hierarchical Multi-Label Associative Classification (HMAC)	[19]
2011	Multi-label Lazy Associative Classification (LAC)	[13]
2014	Enhanced Multi-label classifier (eMCAC)	[20][21][15]
2018	Weighted Multi-label Associative Classifiers (WMAC)	[22]
2019	extended Hierarchical Multi-label Associative Classification (eHMAC)	[23]
2020	ARM-MLC	[16]

other is (secondary) multi-class label. We propose to formulate the MLC problem as a 0–1 integer optimization model and solve it by a two-phase algorithm based on association rule mining and combinatorial optimization techniques. First, we employ apriori algorithm to generate association rules corresponding to multi-class labels. Then, we develop a heuristic algorithm to solve the optimization model that integrates a ranking metric based on selection criteria such as support, confidence, lift, and length of rule to prioritize rule selection with high coverage and few features included in rules. Finally, we aggregate all selected rules to form an ensemble classifier for multi-class classification. We validate its effectiveness and generalizability by comparing to state-of-the-art machine learning methods for various datasets. Finally, to demonstrate the practical contribution, we design a prediction tool interfaced with the developed method for a medical diagnosis problem in emergency care to predict high-risk unplanned transfers to intensive care unit (ICU).

The remainder of our paper is organized as follows. In Section 2, we discuss related works. Section 3 describes the overall framework of the proposed method. In Section 4, experimental results and analysis are provided to demonstrate the effectiveness of the proposed method. In Section 5, we conclude current work and mention potential future works.

2. Related works

In this section, we review relevant studies on multi-label associative classification in literature, summarized in Table 1, and discuss other research studies related to it.

Association rule mining was firstly introduced by Agrawal et al. to discover regularities between products in large-scale transaction data for market bask analysis [24]. Liu et al. coined the term associative classification [8]. Associative Classifier (AC) employs association rule mining in supervised learning for classification purpose. Thus, mining association rules is the foundation of associative classification. Various accurate classifiers that use the associative classification as the framework have been developed to deal with the traditional single-label classification problems. For example, Sun et al. presented a comprehensive review of the existing associative classifiers [25]. Later, multi-label associative classification increasingly received much attention and has been pervasive in various application areas due to its generality. Achieving high classification accuracy is then one of major tasks because it determines the success of the applications. Thabtah et al. firstly proposed a label-ranking-based assignment method, called Multi-label Multi-class Associative Classifier (MMAC), to merge multiple labels and performed recursive learning from parts of training datasets [14]. Li et al. further investigated associative classification applied in the multi-label problem and conducted experiments with multi-label data to confirm the performance of MMAC upon multi-label learning [18]. Sangsuriyun et al. developed hierarchical multi-label associative classifier (HMAC) method to primarily study hierarchical data with importing negative class association rules (lift < 1) [19]. Afterwards, Sangsuriyun et al. further came up with an extended HMAC (eHMAC) for protein function prediction base on the previous study, which newly introduced Gene Ontology (GO) terms as background knowledge to discover more hidden valuable rules [23]. Veloso et al. described a lazy strategy based on test instances for mainly searching small disjuncts upon exploring correlation among labels [13]. Liu et al. proposed a weighted MAC (WMAC) to learn an universal weight vector of features by embedding the set of rules into a linear model and weighing the association rules by their confidence [22]. Prathibhamol et al. presented an ensemble method, called ARM-MLC, for the multi-label classification by using a clustering algorithm and two association rule mining algorithms including Aprior algorithm and FP-Growth algorithm [16]. Thanajiranthorn and Songram studied the generation of multi-label rules via AC from single-label datasets, which is often applied to analyze phishing classification [20].

In addition, some of above-mentioned researches have applied the ranking algorithm in dealing with multi-label classification problems. Thabtah et al. filtered top-rank rules and discarded some insignificant ones based on the threshold values of confidence, support, and length of rules in the process of association rule mining [14,26]. In this study, we consider a similar ranking metric for rules under the AC framework, but apply our ranking metric as the weight value to further select rules and our ranking process will not discard any bottom-rank rules.

Decision rule denotes as *if*…*then*… statement consisting of a condition and a prediction. A single decision rule or a combination of several rules can be used to support analysis and prediction. The concept of decision list (sets of decision rules) was recently adopted for interpretable classification in various domains [17,27–29]. Letham et al. designed a Bayesian generative decision list for Stroke prediction [30]. Wang et al. presented a Bayesian framework to discover short rule sets by the simulated study for user behavior analysis in the personalized recommendation system [31]. Bao et al. proposed a framework to construct decision list for predicting paper acceptance by collecting of disordered if-then rules that can both accurately predict class labels and interpretably describe its decision boundaries instead of organizing them in a hierarchy [32]. Ospina-Mateus et al. applied a heuristic approach based on genetic algorithm in conjunction with simulated annealing to yield hierarchical decision rule sets for identifying the severity of motorcycle traffic accidents [33]. Unlike their approaches, our work regards strong association rules obtained from a combinatorial optimization model as the decision lists that are used for multi-label classification.

Combinatorial optimization technique can be generally applied to data mining problems such as classification, clustering, feature selection, and association rule mining. Lim et al. proposed a numerical optimization approach to solve multi-label feature selection problem by designing a score function based on mutual information between features and labels [34]. Bayati et al. firstly used Particle Swarm Optimizer (PSO) and Competitive Swarm Optimizer (CSO) to present a novel filter approach for feature selection for multi-label classification [35]. Chang et al. proposed an optimization model to generate rules and then learned optimal ranking for rules to build a decision list as a classifier for the purpose of binary classification [36]. Bui-Thi et al. formulated

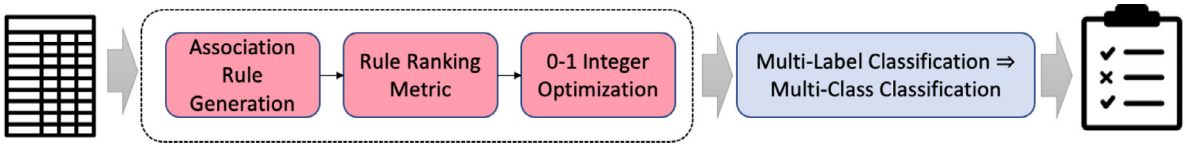


Fig. 1. The proposed methodological framework.

interestingness measure learning as a multi-objective optimization framework that attempts to balance between classifier's size and its performance and developed an interpretable classification model consists of an optimal rule list that is greedily selected based on known interestingness measure [17]. Chou et al. formulated the unplanned ICU transfer prediction problem as a binary classification problem and solved it using a two-phase heuristic strategy [37]. In our work, we propose and solve a multi-label classification problem, and our optimization approach includes multi-stage phases by combining rule ranking to maximize the coverage of rules.

To the best of our knowledge, there is no study that focuses on optimization modeling and solution approach for the multi-label associative classification. In this study, we present a 0–1 integer optimization approach for it to discover key features and their associations corresponding to the multi-label outcome of interest.

3. Proposed method

3.1. Problem definition

Consider a dataset of n samples that are characterized by m binary features $\{a_{ij}\} \in \{0, 1\}$, where $i = 1, \ldots, n$ and $j = 1, \ldots, m$, and categorized by two outcomes (or class labels), where the primary outcome is binary, $c_1 \in \{0, 1\}$, and the secondary outcome is categorical, $c_2 \in \{1, 2, \ldots, p\}$. The goal is to discover patterns or rules (i.e., combinations of binary features) in association with the two class labels. It is formally defined as a multi-label associative classification (MLAC) problem. First, we formulate it as a 0–1 integer optimization problem in terms of rule properties such as maximum coverage, high confidence, and low complexity. Moreover, we design a ranking metric that integrates all mentioned rule properties to prioritize rule selection in the optimization model. Since this is an extremely hard combinatorial problem, we then propose a two-phase solution approach for (1) mining *strong* association rules from binary data corresponding to target class (c_1 = 1) by the apriori algorithm and (2) selecting *superior* association rules or *decision rules* by a designed heuristic algorithm. As a result, we derive a set of *if-then* rules that include discriminating features for classifying samples between different categorical target classes ($c_1 = 1 \cap c_2 = p$) and non-target class ($c_1 = 0$). The overall framework is illustrated in Fig. 1.

3.2. Association rule mining

To discover the knowledge between features and class outcomes, we adopt the concept of association rule mining, originated from market basket analysis [38]. The mining of association rules is a two-step process of searching high frequent combinations of features and then generating strong rules with a high confidence of associations between antecedent and consequent. Given a set of m binary features $I = \{I_1, I_2, \ldots, I_m\}$ and outcome c_{new} in a dataset, an association rule has the form of $\{\bar{I}\} \Longrightarrow \{C\}$, where $\{\bar{I}\} \subset I$. In the following, we briefly describe the important definitions for association rule mining. The criteria of rule selection are often based on these measurements.

Definition 1. *SupportCount*$(\{\bar{I}\} \cap \{C\})$ denotes the number of samples that contain feature sets $\{\bar{I}\}$ and $\{C\}$, which measures the frequency of a rule.

Definition 2. *MinSupp* is defined as a minimum number of samples that contain feature sets $\{\bar{I}\}$ and $\{C\}$, and used for filtering frequent combinations of feature sets $\{\bar{I}\}$ and $\{C\}$.

Definition 3. *Confidence*$(\{\bar{I}\} \rightarrow \{C\})$ is defined as the percentage of samples having feature set $\{C\}$ that contain feature set $\{\bar{I}\}$.

Definition 4. *MinConf* is defined as the minimum percentage of samples having feature set $\{C\}$, which contain feature set $\{\bar{I}\}$, and used for filtering strong association rules.

Definition 5. Lift l is defined as the ratio of the probability of samples containing feature sets $\{\bar{I}\}$ and $\{C\}$ to the probability of those that have feature set $\{C\}$. When $l > 1$, it expresses a positive dependence between variable sets $\{\bar{I}\}$ and $\{C\}$, and an association rule that contains feature sets $\{\bar{I}\}$ and $\{C\}$ is useful.

Definition 6. *MaxRuleLen* is defined as the maximum length of a rule or the number of features used in a rule. For example, *MaxRuleLen* is 3 for the rule $\{I_1, I_3\} \Longrightarrow \{C\}$.

Definition 7. A strong association rule is defined as a rule that must meet *MinSupp*, *MinConf*, *MaxRuleLen*, Lift $l > 1$.

3.3. Ranking association rules

Furthermore, to select better or more representative rules, we design a new ranking metric that integrates the above-mentioned rule criteria such as support, confidence, lift, and rule length to rank association rules. let us consider two rules R_1 and R_2. R_1 is superior to R_2 when the following conditions are met:

1. The confidence of R_1 is greater than the confidence of R_2.
2. When both confidence values of R_1 and R_2 are the same, the support value of R_1 is greater.
3. When the confidence and support values of R_1 and R_2 are the identical, the lift value of R_1 is greater.
4. When the confidence, support, and lift values of R_1 and R_2 are the same, but the rule length of R_1 is less than that of R_2 because R_1 has fewer features.

The *ranking metric* for rule selection is defined as follows:

$$rk(R_k) = \frac{|K| - (k-1)}{|K|} \quad k \in \{1, 2, \ldots, |K|\}, \tag{1}$$

where k is the ordered index and $|K|$ is the total number of rules. The larger $rk(R_k)$ is, the stronger rule R_k is. Here we give an example of three rules R_1, R_2 and R_3. The parameters of rule R_1 are *support* = 0.4, *confidence* = 0.8, *lift* = 3.56, and *length* = 4, the parameters of rule R_2 are *support* = 0.5, *confidence* = 0.7, *lift* = 3.49, and *length* = 3, and the parameters of rule R_3 are *support* = 0.5, *confidence* = 0.7, *lift* = 3.49, and *length* = 4. According to the four above-mentioned conditions, R_1 is superior to R_2 and R_2 is

prior to R_3. The corresponding ranking matrices $rk(R_1) = 1$, $rk(R_2) = 2/3$, and $rk(R_3) = 1/3$. We will integrate this ranking metric in the rule selection optimization model in the next section.

3.4. Rule selection optimization model

In this section, we formally develop a 0–1 integer optimization model for strong rule selection. Suppose we have a set of rules generated from a dataset of n samples categorized in target class (I^+) versus non-target class (I^-), respectively, where $I^+ \cup I^- = I$ and $I^+ \cap I^- = \emptyset$. Note that in this study, we focus on the generation of association rules for target class; that is, any rules that include more samples in target class than non-target class are favorable. Each generated rule is composed of several features and covers a subset of samples in target and non-target classes. The overall objective is to select one or more top-ranked decision rules that can cover as many as possible samples in target class.

Assume that there are m rules to be ordered after applying the ranking metric. We denote b_{jk} as a binary parameter to indicate if feature j is used in the selected rule k or not, where $j = 1, \ldots, m \in J$ and $k = 1, \ldots, m \in K$. We also denote Cr_{ik} to measure how likely sample i is covered by ranked rule k or not, where $i = 1, \ldots, n \in I$. Note that Cr_{ik} is not a binary value because we take into consideration its ranking metric shown in Eq. (1). The measurement is shown as follows:

$$Cr_{ik} = \begin{cases} 1 + rk(R_k), & \text{if sample } i \text{ is covered by rule } k \\ 0, & \text{if sample } i \text{ is not covered by rule } k. \end{cases} \tag{2}$$

Three decision variables are defined as follows:

- $x_i \in \{0, 1\}$ indicates whether sample i can be covered or not, where $i = 1, \ldots, n \in I$.
- $y_j \in \{0, 1\}$ indicates whether feature j is included in the decision model or not, where $j = 1, \ldots, m \in J$.
- $z_k \in \{0, 1\}$ indicates whether ranked rule k is used in the decision model or not, where $k = 1, \ldots, |K| \in K$.

The integer optimization model for rule selection is formulated as follows:

$$\min \quad \alpha \sum_{j=1}^{J} y_j + \beta \sum_{k=1}^{K} z_k + \gamma \sum_{i \in I^-} x_i - \lambda \sum_{i \in I^+} x_i, \tag{3}$$

$$\text{s.t.} \quad \sum_{k \in K} Cr_{ik}^{+} z_k \geq x_i \qquad \forall i \in I^+, \tag{4}$$

$$\sum_{k \in K} Cr_{ik}^{-} z_k \leq M_1 x_i \qquad \forall i \in I^-, \tag{5}$$

$$\sum_{k \in K} B_{jk} z_k \leq M_2 y_j \qquad \forall j \in J, \tag{6}$$

$$x_i, y_j, z_k \in \{0, 1\} \tag{7}$$

The objective function in Eq. (3) is to minimize the number of features and the number of rules included in the decision model while ensuring that the selected rules can minimize non-target class coverage and maximize target class coverage. The constraint set in Eq. (4) ensures that if target sample i is covered and it is covered by at least one rule with a larger ranking metric. The constraint set in Eq. (5) adds the penalty function to the whole model if non-target sample i is covered by selected ranked rules. M_1 and M_2 are big number and here set to $|K| + 1$. The constraint set in Eq. (6) indicates whether feature j is included in any selected rules. α, β, γ, and λ are weight parameters, which are determined by end users depending on the emphasis of the model. In practice, λ is set to a relatively big number to ensure that the model can cover as many target samples as possible. In our study, the setting of these parameters (M_1, M_2, α, β, γ, and λ) in the model follows the previous study in [37].

3.5. Heuristic solution approach

The proposed optimization model is a NP-hard problem which is difficult to be solved as the number of possible association rules increases exponentially with the feature size. Here we introduce a simulated annealing-based heuristic approach (SARules) for solving the rule selection model. Because the decision variables are all binary, we can consider the optimization model as a multi-variable 0–1 knapsack problem.

We define $R_{Gen} = \{z_1, z_2, z_3, \ldots, z_k\}$, as the set of superior association rule selection, where $z_k \in \{0, 1\}$ and $k \in \{1, 2, \ldots, K\}$. R_{Gen} presents the details about the coverage of samples I in target class and selected feature sets J, which are easily read and counted for the generated rule set. In this way, the above-mentioned optimization model can be transformed into the model based on R_{Gen} as follows:

$$\min \; F(z) = \beta \sum_{k=1}^{K} z_k + \Theta_J(z, y_j) + \Theta_I(z, x_{i^-}) - \Theta_I(z, x_{i^+}) \tag{8}$$

$$W(R_{Gen}, \; \forall I, \; J): \begin{cases} W_1(z, \; \forall I^+) = x_{i^+} - \sum_{k \in K} Cr_{i^+k} z_k, \\ W_2(z, \; \forall I^-) = \sum_{k \in K} Cr_{i^-k} z_k - M_1 x_{i^-}, \\ W_3(z, \; \forall J) = \sum_{k \in K} B_{jk} z_k - M_2 y_j, \end{cases} \leq 0 \tag{9}$$

The objective function $F(z)$ in Eq. (8) is equivalent to one in Eq. (3), where Θ_J and Θ_I are functions of z_k and y_j, and z_k and x_i, respectively. Three independent constraints in Eqs. (4)–(6) are integrated into Eq. (9).

For the SARules heuristics, we define several input parameters: initial temperature (t_0), final temperature (t_f), the length of Markov chain L (i.e., the number of iterations at each temperature level), and cooling rate (α). R_{Gen} and R^*_{Gen} are defined as the current solution and new solution, respectively. Here we define the initial solution $R^K_{Gen} = \{0, 0, \ldots, 0\}$, which means none of the rules are selected at the beginning. The whole cooling procedure mainly works by iteratively exploring new feasible neighbor solutions and storing the best one found in the process. We analyze a neighbor solution at each iteration by adding a new neighbor rule to the current solution and possibly removing some rules from it to assure the feasibility. The random choice is performed (line 9 in Algorithm 1) to randomly add a rule if it is not included in the current solution (lines 10–23 in Algorithm 1). Otherwise, it will be removed from the solution (lines 25–32 in Algorithm 1). In terms of accepting new solutions, we adopt the metropolis criterion invented by Gelman et al. (1996) to determine whether or not to accept better and worse solutions with a probability defined as follows:

$$p = \begin{cases} exp(\frac{F(R^*_{Gen}) - F(R_{Gen})}{T}), & \text{if } F(R^*_{Gen}) \geq F(R_{Gen}) \\ 1, & \text{if } F(R^*_{Gen}) < F(R_{Gen}), \end{cases} \tag{10}$$

where the parameter T controls the annealing process. If the objective function value of new solution $F(R^*_{Gen})$ is better than that of old solution $F(R_{Gen})$, new solution R^*_{Gen} will be adopted as the replacement of current solution R_{Gen}, otherwise new solution R^*_{Gen} is accepted with the probability of $exp(\frac{F(R^*_{Gen}) - F(R_{Gen})}{T})$.The time complexity of the algorithm is $O(K \times L \times \frac{t_f - t_0}{\alpha})$. The pseudocode of SARules heuristics is shown in Algorithm 1.

Algorithm 1 SARules.

Procedure $SARules(t_0, t_f, L, W, F, \alpha)$
Initialize $R_{Gen}, R^*_{Gen} \leftarrow \emptyset$
$BestR_{Gen} \leftarrow R_{Gen}$
Random select a rule $m \in \{1, ..., K\}$ in ranked rule list
$t \leftarrow t_0$
while t $\geq t_f$ **do**
  for $it \leftarrow 0, L$ **do**
    $R^*_{Gen} \leftarrow R_{Gen}$
    Random select a rule $m \in \{1, ..., K\}$
    if rule m $\notin R_{Gen}$ **then**
      $R^*_{Gen} \leftarrow R^*_{Gen} \cup \{m\}$
      if $W(R^*_{Gen}, \forall I, \forall J) \leq 0$ **then**
        $R_{Gen} = R_{Gen} \cup \{m\}$
        $F = F(R_{Gen}), W = W(R^*_{Gen}, \forall I, \forall J)$
      else
        Randomly select a rule n satisfies rule n $\in R^*_{Gen}$
        $R^*_{Gen} \leftarrow \{R^*_{Gen}$ - rule $n\}$
        $\Delta = F(R^*_{Gen}) - F(R_{Gen})$
        $Rand \leftarrow$ a random value between 0 and 1
        if $W(R^*_{Gen}, \forall I, \forall J) \leq 0$ and ($\Delta \leq 0$ or $Exp[\Delta/t] > Rand$) **then**
          $R_{Gen} \leftarrow R^*_{Gen}$
        end if
      end if
    else
      Random select a rule n satisfies rule n $\notin R^*_{Gen}$
      $R^*_{Gen} \leftarrow \{R^*_{Gen} - rule\ m + rule\ n\}$
      $\Delta = F(R^*_{Gen}) - F(R_{Gen})$
      $Rand \leftarrow$a random value between 0 and 1
      if $W(R^*_{Gen}, \forall I, \forall J) \leq 0$ and ($\Delta \leq 0$ or $Exp[\Delta/t] > Rand$) **then**
        $R_{Gen} \leftarrow R^*_{Gen}$
      end if
    end if
    if $F(R_{Gen}) < F(BestR_{Gen})$ **then**
      $BestR_{Gen} \leftarrow R_{Gen}$
    end if
  end for
  $t \leftarrow \alpha * t$
end while
return $BestR_{Gen}$

3.6. Transformation of multi-label classification into multi-class classification

As a key step, we transform the multi-label classification problem into a multi-class classification problem by applying a logical operation to combine multiple class labels into one single multi-class label. In Table 2, we show an exemplar. The primary labels c_1 = 1 and $c_1 = 0$ represent the target class and non-target class, respectively. The secondary label c_2 = 1, 2, 3, or 4 is combined with the primary label to form a new class label, expressed as $c_{new} = c_1 \cap c_2 \in \{0, 1\}$. In a practical example, medical practitioners would like to know if patients with cardiovascular diseases (secondary label) is at high risk (primary label).

3.7. Multi-class classification

In the section, we build a multi-class classification model by grouping the set of obtained decision rules, which include distinguishing features, to classify samples between different categorical target classes ($c_{new} \neq 0$) and non-target class ($c_{new} = 0$). Assume we obtain $|K|$ decision rules in R_{Gen} from our SARules, we know the resulting coverage x_{ik} to indicate that sample i is covered by rule k and resulting selection $r_{kc_{new}} \in \{0, 1\}$ to indicate rule k in the decision rule set R_{Gen} is applied to class $c_{new} = \{1, 2, \ldots, p\}$. Note that each selected rule k in R_{Gen} has its $l(r_k)$ and $rk(r_k)$ obtained from Sections 3.2–3.3. We also denote $S_{c_{new}}$ as the set of rules that are classified to class c_{new}, and $L_{c_{new}}$ as the sample size in class c_{new}. Besides, $P(x_{ic_{new}})$ is denoted as the probability of sample i classified to class c_{new}. The multi-class classification, shown in Algorithm 2, includes two steps: ruleset selection in each categorical target class and sample prediction based on the coverage. We first count samples in every class c_{new} covered by each rule k, and then group rule k into the class c_{new} with the largest coverage (lines 3–10). In the second step, sample i is categorized into the class with the highest probability by comparing its rule coverage in each class c_{new} (lines 12–19).

Table 2
The transformation of primary label and secondary into a combined label.

Primary label (c_1)	Secondary label (c_2)	Combined label (c_{new})
1	1	1
1	2	2
1	3	3
1	4	4
0	1, 2, 3 or 4	0

Algorithm 2 Multi-class Classification based on Rule Generation.

1: **Input:** R_{Gen}, *samples* I, c_{new}, $L_{c_{new}}$, rk, l
2: **Output:** $P(x_{ic_{new}})$
3: **for** $r_1, r_2, ..., r_k$ in R_{Gen} **do**
4: Initiation c_{new}, $S_{c_{new}} \leftarrow \emptyset$
5: **for** $k = 1 : |K|$, $c_{new} = 1 : p$ **do**
6: $P(r_k|c_{new}) = \frac{\sum_{k=1}^{K} \sum_{c_{new}=1}^{p} r_{kc_{new}} l(r_k)}{L_{c_{new}}}$
7: $P(r_k) \leftarrow \frac{max\{P(r_k|c_{new})\}}{\sum_{c_{new}=1}^{p} P(r_k|c_{new})}$
8: $r_k \leftarrow r_k$ applied to $argmax\{P(r_k|c_{new})\}$
9: $S_{c_{new}} \leftarrow S_{c_{new}} \bigcup r_k$
10: **end for**
11: **end for**
12: **for** $i = 1 : I$, $k = 1 : |K|$ **do**
13: Initiation $P(x_{ic_{new}}) \leftarrow \emptyset$, $Pr(x_{ic_{new}}) \leftarrow \emptyset$
14: **if** $x_{ik} = 1$ **then**
15: $\{P(r_k), c_{new}\} \leftarrow when\ r_k$ applies to $S_{c_{new}}$
16: $Pr(x_{ic_{new}}) = \sum_{k \in \{rules\ within\ same\ class\ g\}} P(r_k) rk(r_k))$
17: **end if**
18: $P(x_{ic_{new}}) \leftarrow max\{\frac{Pr(x_{ic_{new}})}{\sum_{c_{new}=1}^{p} Pr(x_{ic_{new}})}\}$
19: **end for**
20: Return $P(x_{ic_{new}})$

4. Experimental results

In this section, we demonstrate the performance of our proposed approach (SARules) using four datasets including ICU [37], Mushroom [40], Breast Cancer [41], German Credit [42], and Stroke [43]. Each dataset (excluding Stroke dataset) has one primary class label and one secondary class label. The sample size ranges from a few hundred to tens of thousands and there are various categorical features. The area under ROC curve (AUC) is employed as the major metric to evaluate the classification performance of our method. All computational experiments are coded in python 3.8.0. and Gurobi 9.0.0.

In this study, we first demonstrate the ranked association rule analysis in the overall group and different subgroups for the above five different datasets in our experiments.

4.1. Real applications and data description

Table 3 displays a descriptive summary of all datasets including sizes of sample and feature.

- ICU dataset includes 32 diagnostic features for 1049 patients, where 313 patients are labeled as high risk (unplanned ICU transfer (UIT)) in four subgroups such as Infection, Gastrointestinal disease, Cardiovascular/respiratory disease, and Neurological/other diseases based on the reasons of their emergency department visits. The feature set

Table 3
Description of datasets used in this paper.

Dataset	Sample size	Target class	Subgroup description	Subgroup sample #/ Target sample #	Feature #	Feature description
ICU	1049	High-risk to Unplanned ICU Transfer (UIT)	Based on ED Visits	Infection/High-risk to UIT: 353/138 Gastrointestinal disease/High-risk to UIT: 241/62 Cardiovascular and respiratory diseases/High-risk to UIT: 146/64 Neurological and other diseases/High-risk to UIT: 309/49	32	1. Triage 1, 2. Triage 2, 3. > 65 years, 4. Diabetes, 5. Hypertension, 6. Coronary artery disease, 7. Cerebral vascular disease, 8. Cerebral performance category, 9. Respiratory failure history, 10. Congestive heart failure history, 11. Liver cirrhosis history, 12. End-stage renal disease, 13. Cancer, 14. Immune compromise, 15. BT>38C, 16. Heart rate>130, 17. Abnormal white blood cell counts, 18. SIRS, 19. Hypertension, 20. Respiratory compromise, 21. Renal dysfunction, 22. Liver dysfunction, 23. Hematological dysfunction, 24. Metabolic dysfunction, 25. Respiratory arrest, 26. Respiratory distress, 27. Heart rate <40, 28. Oliguria, 29. Altered mental Status, 30. Seizure, 31. Arrhythmia, 32. Chest pain
Mushroom	8124	Poisonous	Gill Spacing	Close/Poisonous: 6812/3804 Crowded/Poisonous: 1312/112	20	1. cap-shape, 2. cap-surface, 3. cap-color,4.bruises, 5. odor, 6. gill-size, 7. population, 8. veil-color, 9. gill-color, 10. gill-attachment, 11. ring-type, 12. ring-number, 13. stalk-surface-above-ring, 14. stalk-color-above-ring, 15. stalk-shape, 16. talk-root, 17. sstalk-surface-below-ring, 18. stalk-color-below-ring, 19. spore-print-color, 20. habitat
German Credit	1000	Good Risk	Housing	Free/Good Risk: 108/64 Own/Good Risk: 713/527 Rent/Good Risk: 179/109	8	1. Age, 2. Sex, 3. Job, 4. Saving account, 5. Checking account, 6. Credit amount, 7. Duration, 8. Purpose
Breast Cancer	286	Recurrence	Breast Quadrant	Left/Recurrence: 97/26 Center/Recurrence: 156/46 Right/Recurrence: 33/13	8	1. Age, 2. menopause, 3. irradiat, 4. tumor-size, 5. inv-nodes, 6. breast, 7. node-caps, 8. deg-malig
Stroke	29072	Stroke	stroke	Overall/Stroke: 29072/548	10	1. Gender, 2. Residence, 3. Age, 4. bmi, 5. Hypertension, 6. Glucose level, 7. Heart Disease, 8. Smoking Status, 9. Marriage, 10. Work type

consists of demographics (Indice 3), comorbid conditions (Indices 4–8), chronic organ insufficiency (Indices 9–14), physiological responses (Indices 15–18), organ dysfunctions (Indices 19–24), and other symptoms/signs (Indices 25–32), as well as triage information (Indices 1–2). Triage is the sorting allocation of treatment to patients (as in an emergency room) according to the urgency of their need for care, which is a subjective diagnosis of patients' urgency level from experienced nurses and physicians. In our experiments, triage is divided into two levels: Triage 1 is a very high priority of medical attention and Triage 2 is a common level of medical attention. The other diagnostic features comprehensively consider various medical indicators and diagnostic criteria.

- Mushroom dataset contains 8120 records represented by 20 features and grouped into two classes (edible and poisonous) based on the edibility of mushrooms. There are only 112 poisonous mushrooms out of 1,312 mushrooms with crowded gill-spacing, accounting for 9%. The main goal is to identify poisonous mushroom feature combinations associating with their close and crowded gill-spacing.
- German Credit dataset is used to discover key factors associating with good credit risk, which contains 8 risk factors and 1000 records grouped into three classes based on the types of Germans' housing.
- Breast Cancer dataset contains 8 diagnostic features and 286 patient records, and we want to discover associative patterns among risk factors corresponding to breast cancer recurrence under different conditions. These patient records can be also classified into three subgroups: left, center, and right based on the quadrant of the recurrence of their breast cancer.
- Stroke dataset contains 10 features, where only 783 stroke cases out of 29,072 patient records. There are 30% missing data in smoking status and 3% missing data in BMI. In our experiments, samples with missing values are removed. We keep the completeness of applied sample data and target to figure out the diagnostic key factors leading to the stroke outcome.

The parameter settings in our SARules heuristics for solving the rule selection optimization model are shown in Table 4. Such suggested parameter values are set generally based on the size of the different datasets. As the dataset grows, the parameter values including the temperature value t and iteration number L increase accordingly. Besides, the cooling rate α we suggest in the following table for each dataset ($\alpha \rightarrow 1$ to ensure a large search space) is obtained through our many experiments and adjustments taking into account the computation time and complexity and has a relatively good performance.

4.2. Experimental results for unplanned ICU transfer prediction

In this section, we in particular demonstrate the analysis detail for ICU dataset. Chou et al. (2020) proposed a rule-based decision model (ARSOM) based on the same dataset to identify key diagnostic features in association with high-risk patients who undergo unplanned ICU transfer (UIT) in emergency department. We are motivated to consider this UIT prediction as a multi-label associative classification problem and then develop the SARules heuristics. We herein present a comparison for experimental results among SARules, ARSOM, and decision trees (DT).

Table 5 displays a summary of rule selection. There are similar numbers of association rules generated by SARules and ARSOM

Table 4
The settings in SA algorithm for solving rule generation optimization model.

Dataset	Initial temperature t_0	Final temperature t_f	cooling rate α	Iterations # L at each temperature
ICU	100	0.5	0.9	10000
Mushroom	100	0.5	0.85	80000
German Credit	100	0.5	0.9	10000
Breast Cancer	100	0.5	0.95	5000
Stroke	97	3	0.85	100000

Table 5
Summary of selected rules for ICU dataset of SARules, ARSOM, and DT.

Method	Selected rule #	Selected rule # in 4 subgroups	Covered feature #	Frequent features
SARules	29	Infection:17 Gastrointestinal disease:4 Cardiovascular and respiratory diseases:6 Neurological and other diseases:2	25	"SIRS", "Abnormal white blood cell counts", "Hypertension" and ">65years"
ARSOM	25	Infection:13 Gastrointestinal disease:13 Cardiovascular and respiratory diseases:11 Neurological and other diseases:4	23	"SIRS","Hypertension", ">65years"
DT	14	Infection:4 Gastrointestinal disease:6 Cardiovascular and respiratory diseases:6 Neurological and other diseases:4	20	"SIRS", "Renal dysfunction", "Liver cirrhosis history", "Hypotension"

and both methods cover the same features except for the feature 'Triage' which is newly introduced by SARules. However, it is observed that over 50% selected rules are slightly or completely different between SARules and ARSOM, which may result from that the import of a new feature such as 'Triage' updates associative patterns among risk factors. It seems obvious that the number of rules generated by DT, whatever for the overall group or four subgroups, is significantly fewer than the numbers of rules generated by the last two methods, even if the numbers of covered features among three methods are comparable. For the infection subgroup and the cardiovascular and respiratory diseases subgroup, the numbers of rules they covered are relatively more than those in other subgroups among the three methods; that is, UIT is associated with relevant risk factors in infection disease and cardiovascular and respiratory diseases.

Besides, all three models show that feature 'SIRS' appears frequently, and features 'Hypertension' and '$>$ 65 years' also have relative high frequencies in the resulting rules of SARules and ARSOM. We summarized that the three diagnostic features are most significant for the identification of unplanned ICU transfer patients.

In Table 6, we display 29 top-ranked association rules generated by SARules as all patient records are considered in our proposed model. Most rules cover all patients in three subgroups of infection, gastrointestinal disease, and cardiovascular/respiratory diseases while neurological/other diseases subgroup is covered by fewer rules. For the infection subgroup, most diagnostic features in comorbid conditions, chronic organ insufficiency, and physiological responses presenting in selected rules, are shown to be associated with UIT. It is observed that the diagnostic feature 'Hematological dysfunction' appearing the most times is associated with a high-risk outcome for patients with gastrointestinal disease. Some relevant features in cardiovascular and respiratory diseases such as 'BT$>$38C', 'Congestive heart failure history' are associated with UIT. For the subgroup neurological and other diseases, UIT could be identified by only two risk factors: 'Altered mental status' and 'Metabolic dysfunction'.

The sets of selected rules obtained by DT for different patient groups are presented in Table 7. For the group in which all patients are considered as a whole, most of features in organ dysfunctions are covered in rules from DT and they are associated with UIT. We observe that one of demographic features 'Triage1' appearing in every rule is very critical for UIT identification for patients with infection disease. For the subgroup gastrointestinal disease, unlike SARules, it seems that the feature 'Hematological dysfunction' is unable to discover UIT, because it is never covered by any rules from DT. For the cardiovascular and respiratory diseases subgroup, UIT is discovered by several diagnostic features from other symptoms/signs including 'Altered mental status', 'Arrythmia', 'Chest pain' in most cases. DT yields more rules including more features that are able to identify UIT than those generated from SARules for the subgroup neurological and other diseases. Among all features included in the top-ranked rules generated by both methods, SIRS appears frequently and may be perceived as an important UIT indicator. However, in contrast with our proposed method (SARules), rules obtained from DT seem to be redundant and less interpretable. Table 7 clearly shows that many feature combinations appear repeatedly in multiple rules. For instance, '{No SIRS} $\cap$ {No Renal dysfunction}' is included in 8 out of 14 rules and '{SIRS} $\cap$ {Hyotension}' appears 4 times in the overall group. However, decision rules containing such recurring feature combinations are unconcise and unrealistic for medical diagnosis in the real world. In addition, as a rule-based classifier, DT may not be fully optimized if the root node of the tree happens to be selected badly in the beginning of the construction procedure, which would result in a lack of general interpretability of rules generated by DT.

Furthermore, we design an online tool that is interfaced with SARules for UIT prediction, as shown in Fig. 2. End users such as medical practitioners or patients can easily check those diagnostic features in the table according to patient conditions. It then predicts the possible diagnostic subgroup along with associated probabilities based on association rules which patients belong to.

4.3. Experimental results for other applications

The rule analysis of ICU dataset primarily verifies the feasibility of the proposed method (SARules) and show what generated

Table 6
The top-ranked rules selected by SARules for unplanned ICU transfer prediction.

Rule	Support	Confidence	Lift	Subgroup Probability	Ranked Value	Subgroup
{Cerebral performance category}∩{Renal dysfunction}	0.017	0.900	3.016	0.392	1.988	Infection
{SIRS}∩{Hypotension}	0.038	0.800	2.681	0.514	1.905	Infection
{SIRS}∩{End-stage renal disease}	0.015	0.800	2.681	0.432	1.881	Infection
{Cerebral performance category}∩{Congestive heart failure history}∩{Hypertension}	0.011	0.800	2.681	0.446	1.869	Infection
{Heart rate > 130}∩{Respiratory distress}	0.016	0.773	2.590	0.369	1.750	Infection
{TRIAGE1}∩{Respiratory distress}	0.025	0.765	2.563	0.633	1.738	Infection
{Abnormal white blood cell counts}∩{Hypotension}	0.031	0.762	2.553	0.537	1.702	Infection
{TRIAGE1}∩{Hypotension}	0.018	0.760	2.547	0.629	1.679	Infection
{Cerebral performance category}∩{Respiratory distress}	0.027	0.757	2.536	0.560	1.655	Infection
{TRIAGE1}∩{Renal dysfunction}	0.023	0.750	2.514	0.546	1.619	Infection
{SIRS}∩{Cancer}∩{Diabetes}	0.013	0.737	2.469	0.625	1.512	Infection
{Diabetes}∩{Hypotension}	0.021	0.733	2.458	0.492	1.488	Infection
{Cerebral performance category}∩{Cerebral vascular disease}∩{Hypertension}∩{Coronary artery disease}	0.010	0.733	2.458	0.413	1.417	Infection
{SIRS}∩{Respiratory compromise}∩{Hypertension}∩{> 65years}	0.027	0.718	2.406	0.447	1.321	Infection
{Respiratory compromise}∩{Abnormal white blood cell counts}	0.029	0.714	2.394	0.465	1.298	Infection
{Cancer}∩{Diabetes}∩{Abnormal white blood cell counts}	0.014	0.714	2.394	0.863	1.250	Infection
{SIRS}∩{Hypertension}∩{Abnormal white blood cell counts}∩{Coronary artery disease}	0.013	0.700	2.346	0.450	1.012	Infection
{SIRS}∩{Liver cirrhosis history}	0.025	0.897	3.005	0.760	1.976	Gastrointestinal disease
{Renal dysfunction}∩{Hematological dysfunction}	0.015	0.889	2.979	0.413	1.940	Gastrointestinal disease
{SIRS}∩{Hematological dysfunction}	0.029	0.833	2.793	0.547	1.929	Gastrointestinal disease
{Liver dysfunction}∩{Hematological dysfunction}	0.010	0.733	2.458	0.560	1.476	Gastrointestinal disease
{SIRS}∩{Congestive heart failure history}∩{Coronary artery disease}	0.023	0.800	2.681	0.518	1.893	Cardiovascular and respiratory diseases
{Respiratory compromise}∩{Hypertension}∩{Renal dysfunction}	0.010	0.733	2.458	0.477	1.440	Cardiovascular and respiratory diseases
{BT > 38 C}∩{Congestive heart failure history}∩{>65years}∩{Coronary artery disease}	0.016	0.708	2.374	0.431	1.155	Cardiovascular and respiratory diseases
{Respiratory distress}∩{Coronary artery disease}	0.011	0.706	2.366	0.589	1.119	Cardiovascular and respiratory diseases
{TRIAGE2}∩{Hypertension}∩{Respiratory failure history}	0.011	0.706	2.366	0.589	1.095	Cardiovascular and respiratory diseases
{BT > 38 C}∩{SIRS}∩{Hypertension}∩{Respiratory distress}	0.011	0.706	2.366	0.328	1.060	Cardiovascular and respiratory diseases
{Altered mental status}	0.023	0.889	2.979	0.711	1.964	Neurological and other diseases
{Metabolic dysfunction}	0.024	0.758	2.539	0.323	1.667	Neurological and other diseases

Table 7
Rules generated by decision trees for ICU dataset.

Rule	Subgroup
{No SIRS}∩{No Renal dysfunction}∩{No Altered mental status}∩{No Cerebral performance category}∩{Chest pain}	Overall
{No SIRS}∩{No Renal dysfunction}∩{No Altered mental status}∩{ Cerebral performance category}	Overall
{No SIRS}∩{No Renal dysfunction}∩{Altered mental status}	Overall
{No SIRS}∩{Renal dysfunction}∩{TRIAGE 1}∩{No Cerebral vascular disease}	Overall
{No SIRS}∩{Renal dysfunction}∩{TRIAGE 1}∩{Cerebral vascular disease}∩{Hypotension}	Overall
{No SIRS}∩{Renal dysfunction}∩{TRIAGE 2}∩{No Cerebral performance category}∩{Hematological dysfunction}	Overall
{SIRS}∩{No Hypotension}∩{No Liver cirrhosis history}∩{No Respiratory compromise}∩{Respiratory distress}	Overall
{SIRS}∩{No Hypotension}∩{No Liver cirrhosis history}∩{Respiratory compromise}	Overall
{SIRS}∩{No Hypotension}∩{Liver cirrhosis history}∩{No Renal dysfunction}	Overall
{SIRS}∩{No Hypotension}∩{Liver cirrhosis history}∩{Renal dysfunction}∩{Hematological dysfunction}	Overall
{SIRS}∩{Hypotension}	Overall
{SIRS}∩{Hypotension}∩{No Abnormal white blood cell counts}	Overall
{SIRS}∩{Hypotension}∩{Abnormal white blood cell counts}∩{No Coronary artery disease}	Overall
{SIRS}∩{Hypotension}∩{Abnormal white blood cell counts}∩{Coronary artery disease}∩{Cerebral vascular disease}	Overall
{TRIAGE 1}∩{No Renal dysfunction}∩{No Respiratory distress}∩{Hypotension}	Infection
{TRIAGE 1}∩{No Renal dysfunction}∩{Respiratory distress}	Infection
{TRIAGE 1}∩{Renal dysfunction}∩{No Cerebral vascular disease}	Infection
{TRIAGE 1}∩{Renal dysfunction}∩{Cerebral vascular disease}∩{Coronary artery disease}	Infection
{No Liver cirrhosis history}∩{No Hypotension}∩{Heart rate > 130}∩{Abnormal white blood cell counts}	Gastrointestinal disease
{No Liver cirrhosis history}∩{Hypotension}∩{ Heart rate ≤ 130}	Gastrointestinal disease
{Liver cirrhosis history}∩{No SIRS}∩{No Cerebral performance category}∩{Heart rate > 130}	Gastrointestinal disease
{Liver cirrhosis history}∩{No SIRS}∩{Cerebral performance category}	Gastrointestinal disease
{Liver cirrhosis history}∩{SIRS}∩{No Renal dysfunction}	Gastrointestinal disease
{Liver cirrhosis history}∩{SIRS}∩{Renal dysfunction}∩{BT > 38C}	Gastrointestinal disease
{No SIRS}∩{No Altered mental status}∩{No Chest pain}∩{Abnormal white blood cell counts}	Cardiovascular and respiratory diseases
{No SIRS}∩{No Altered mental status}∩{Chest pain}∩{No Arrythmia}	Cardiovascular and respiratory diseases
{No SIRS}∩{Altered mental status}	Cardiovascular and respiratory diseases
{SIRS}∩{No Chest pain}∩{No Cerebral vascular disease}∩{Respiratory failure history}	Cardiovascular and respiratory diseases
{SIRS}∩{No Chest pain}∩{Cerebral vascular disease}	Cardiovascular and respiratory diseases
{SIRS}∩{Chest pain}	Cardiovascular and respiratory diseases
{No Altered mental status}∩{No Cerebral performance category}∩{Respiratory failure history}	Neurological and other diseases
{No Altered mental status}∩{Cerebral performance category}∩{Renal dysfunction}	Neurological and other diseases
{Altered mental status}∩{No Hypertension}	Neurological and other diseases
{Altered mental status}∩{Hypertension}∩{No Renal dysfunction}∩{SIRS}	Neurological and other diseases

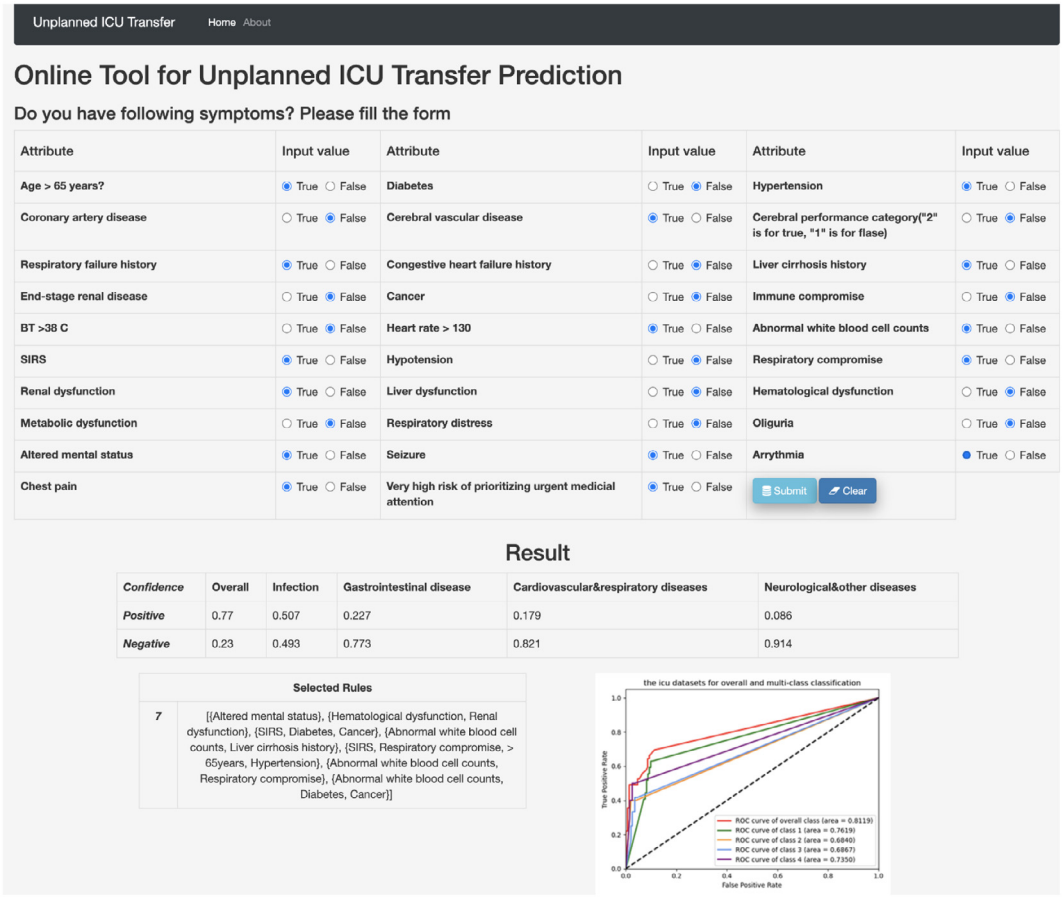


Fig. 2. An illustrative online tool interfaced with our proposed method for Unplanned ICU transfer prediction.

rules are like. We now use benchmark dataset studies to demonstrate the generalization of SARules. We consider a collection of datasets that are frequently used in other similar classification studies: Mushroom, German Credit, Breast Cancer and Stroke.

In Tables 8–11, all decision rules for the four benchmark datasets generated by SARules are listed, together with their corresponding *support*, *confidence*, *lift*, *ranked value*, subgroup categories. The features in each rule are no more than 5. In Table 8, 13 generated association rules were displayed for all patients with breast cancer as a whole. Most of the selected associative patterns are presented in the center and right breast quadrants which may be regarded as the main locations of breast cancer recurrence. We observe that features 'node-caps' and 'medium-size tumor' appearing many times are two important risk diagnostic factors leading to breast cancer recurrence. German Credit study shown in Table 9 displays 12 rules used for identifying Germans with good credit risk. These rules are separately listed into three types of Germans' housing. For the type Germans having their own housing, good credit risk can be identified by the factor 'skilled job', for a skilled job usually brings people a stable income, supporting them to pay for their own house. Besides, we observe that 'no checking account' is also another indicator of good credit risk. Germans with rental housing and good credit risk are shown to be young people who have not opened or used a savings account and residents with rich savings. For the people living in free housing, it seems obvious that good credit risk is associated with their sex (mainly male), medium age, and consuming purposes. For poisonous mushroom prediction as shown in Table 10, there are just 4 selected rules which are all composed of a single feature, whereas they have high confidence. We could see that poisonous mushrooms appear in most cases when the gills of the mushrooms are close. Four features 'foul odor', 'silky stalk-surface-above-ring', 'several populations', and 'narrow gill-size' are shown to be the most significant factors to discover poisonous mushrooms. Table 11 lists 8 generated rules that are all made of one single feature as all patients with stroke are considered as a whole. It is observed that cerebral stroke is strongly associated with the three features 'age', 'marriage' and a high BMI (overweight) among all features covered in the top-ranked selected rules. Besides, we also compared a number of obtained rules using our SARules with DT for the above experimental datasets displayed in Table 12. There is no large gap between the numbers of association rules generated by both methods in the subgroups of each dataset separately, however, the included features for the first three datasets in our SARules are fewer than those in DT. In terms of Stroke diagnosis, our SARules generates two more rules and covers more diagnostic features than those obtained from DT, over the whole imbalanced dataset.

4.4. Performance comparison with state-of-the-art machine learning methods

We analyze and evaluate the classification performance of the selected association rules by our proposed method (SARules)

Table 8
Selected association rules for breast cancer dataset.

Rule	Support	Confidence	Lift	Subgroup probability	Ranked value	Subgroup
{node-caps}∩{Medium-Age}∩{inv-nodes-large}	0.028	0.667	2.243	0.413	1.602	breast-quad-left
{tumor-size-medium}∩{inv-nodes-large}∩{breast-right}	0.010	0.500	1.682	0.500	1.032	breast-quad-left
{Young}∩{node-caps}∩{breast-left}	0.017	1.000	3.365	1.000	1.971	breast-quad-center
{menopause}∩{irradiat}∩ {Olds}∩{breast-left}	0.021	0.750	2.524	0.739	1.742	breast-quad-center
{irradiat}∩{inv-nodes-small}∩{tumor-size-medium}∩{breast-left}	0.042	0.667	2.243	0.861	1.613	breast-quad-center
{irradiat}∩{tumor-size-medium}∩{node-caps}∩{Medium-Age}	0.024	0.538	1.812	0.486	1.165	breast-quad-center
{Young}∩{inv-nodes-small}∩{tumor-size-medium}∩{breast-left}	0.024	0.538	1.812	0.586	1.161	breast-quad-center
{tumor-size-medium}∩{node-caps}∩{Medium-Age}∩{breast-left}	0.031	0.643	2.163	0.569	1.484	breast-quad-center
{irradiat}∩{inv-nodes-small}∩{Medium-Age}∩{tumor-size-large}	0.010	0.500	1.682	1.00	1.025	breast-quad-center
{irradiat}∩{inv-nodes-small}∩{Olds}∩{node-caps}	0.010	1.000	3.365	0.561	1.914	breast-quad-right
{menopause}∩{breast-right}∩{node-caps}∩{deg-malig-High}	0.010	0.600	2.019	0.561	1.341	breast-quad-right
{inv-nodes-small}∩{tumor-size-medium}∩{deg-malig-High}	0.101	0.537	1.807	0.370	1.158	breast-quad-right
{menopause}∩{Olds}∩{tumor-size-large}∩{breast-right}	0.014	0.500	1.682	0.484	1.047	breast-quad-right

Table 9
Selected association rules for German credit dataset.

Rule	Support	Confidence	Lift	Subgroup probability	Ranked value	Subgroup
{Checking account_None}	0.348	0.883	1.262	0.370	1.725	Own
{Job_skilled}	0.444	0.705	1.007	0.357	1.029	Own
{Credit amount_little}	0.590	0.727	1.038	0.374	1.178	Rent
{Job_unskilled and resident}	0.144	0.720	1.029	0.462	1.113	Rent
{Saving accounts_quite rich}	0.052	0.825	1.179	0.405	1.557	Rent
{Saving accounts_None}∩{young}	0.093	0.802	1.145	0.357	1.505	Free
{Saving accounts_None}∩{Purpose_car}	0.057	0.770	1.100	0.480	1.395	Free
{Job_highly skilled}∩{young}∩{male}	0.053	0.746	1.066	0.593	1.291	Free
{Medium-age}	0.183	0.738	1.054	0.440	1.249	Free
{Credit amount_little}∩{Job_highly skilled}	0.061	0.735	1.050	0.469	1.230	Free
{Male}	0.499	0.723	1.033	0.416	1.142	Free
{Medium-age}∩{Purpose_car}	0.064	0.703	1.005	0.482	1.019	Free

Table 10
Selected association rules for mushroom dataset.

Rule	Support	Confidence	Lift	Subgroup probability	Ranked value	Subgroup
(Odor=foul)	0.266	1.000	2.075	1.000	1.988	Gill-spacing close
(Stalk-surface-above-ring=silky)	0.274	0.939	1.949	1.000	1.176	Gill-spacing close
(Population=several)	0.351	0.705	1.462	0.632	1.004	Gill-spacing close
(Gill-size=narrow)	0.274	0.885	1.837	0.643	1.093	Gill-spacing crowded

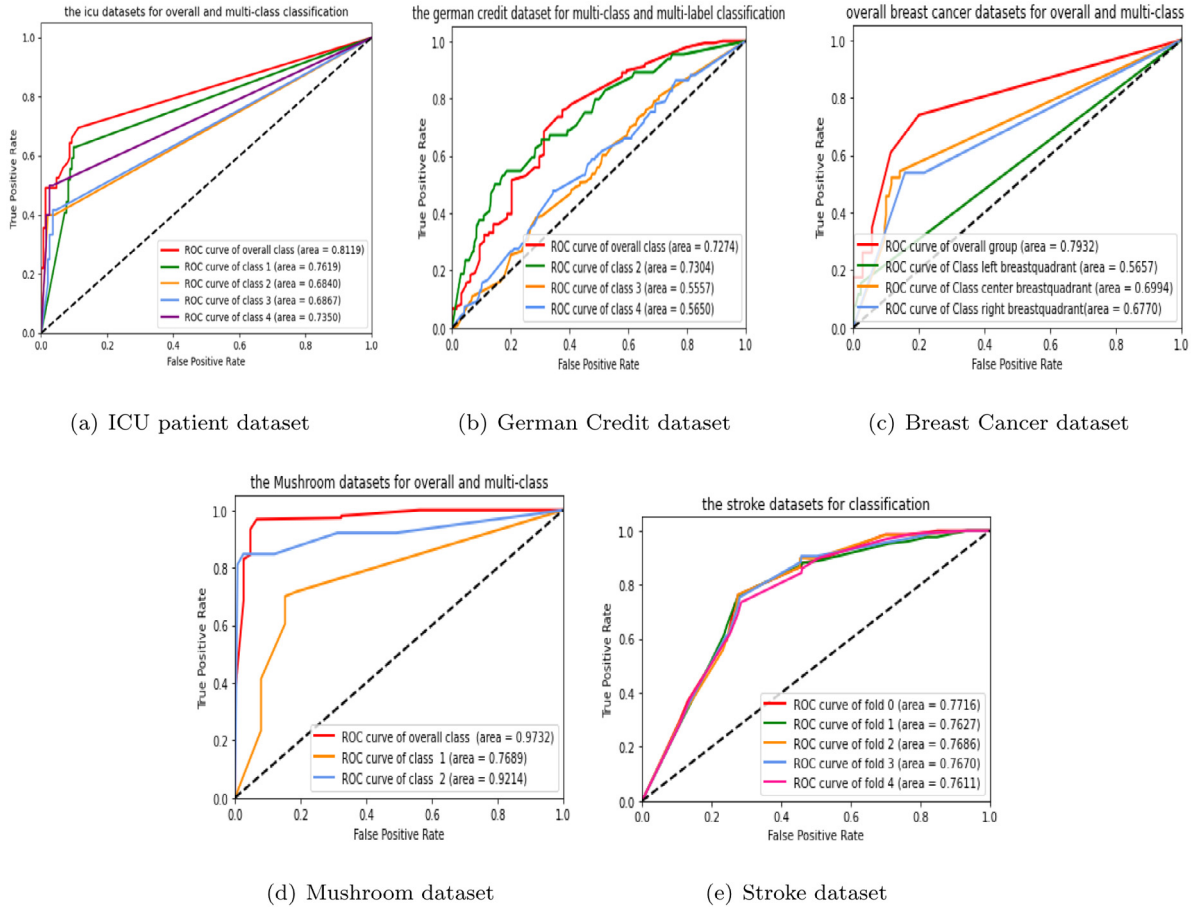
Table 11
Selected association rules for stroke dataset.

Rule	Support	Confidence	Lift	Ranked value	Subgroup
{elderly}∩{Male}	0.006	0.051	2.731	2.000	Stroke
{elderly}∩{never smoked}	0.007	0.050	2.636	1.965	Stroke
{No heart Disease}∩{normal glucose level}∩{elderly}	0.006	0.040	2.118	1.754	Stroke
{No heart Disease}∩{Overweight}∩{elderly}	0.008	0.039	2.092	1.737	Stroke
{Smoked}∩{No Hypertension}∩{married}	0.007	0.022	1.146	1.632	Stroke
{Female}∩{Overweight}∩{married}	0.007	0.021	1.122	1.596	Stroke
{No heart Disease}∩{Urban}∩{married}	0.007	0.019	1.007	1.509	Stroke
{Overweight}∩{No Hypertension}∩{No heart Disease}∩{married}	0.007	0.015	0.822	1.193	Stroke

as compared to ARSOM proposed by Chou et al. (2020) and other three well-developed machine learning methods including logistic regression (LR) [44], DT [45], and random forests (RF) [46]. The computational results for five datasets are shown in Fig. 3, demonstrating the proposed method (SARules) is able to achieve good classification performance. Furthermore, we implement an out-of-sample validation using 10 times 5-fold cross-validation. Table 13 presents the comparison results of the five datasets for overall samples and subgroups separately. Our method outperforms slightly ARSOM, LR, and DT by 2%–10% and is comparable to RF on the overall classification. Besides, our model is not very sensitive to the dataset which contains a large number of imbalanced samples. In terms of subgroup classification, our method is shown to be non-inferior compared to other methods for almost all datasets while ARSOM, LR, DT, and RF slightly outperform our approach SARules around 5%–7%. However, our SARules is simpler than RF which employs hundreds of estimators and all features

Table 12
Rule comparison between our proposed method and decision trees for four datasets.

Dataset		Selected rule #		Covered feature #	
		SARules	DT	SARules	DT
Breast Cancer	Overall	13	7	13	14
	Breastquadrant-Left	2	2		
	Breastquadrant-Center	7	5		
	Breastquadrant-Right	4	2		
German credit	Overall	12	3	11	12
	Housing-Free	7	3		
	Housing-Own	2	3		
	Housing-Rent	3	3		
Mushroom	Overall	4	4	4	10
	Gill-spacing close	3	4		
	Gill-spacing crowded	1	1		
Stroke	Overall	8	6	11	7



(a) ICU patient dataset (b) German Credit dataset (c) Breast Cancer dataset

(d) Mushroom dataset (e) Stroke dataset

Fig. 3. AUC Performance for five datasets.

are included in RF whereas there are limited features covered by SARules. Besides, generative ranked rules in SARules present more interpretable information whereas DT generates discriminative rules for identifying the target class. Comparing to LR and RF, both methods can only present individual feature importance and use its one-size-fit-all result to make classification.

5. Conclusions and future works

In this paper, we proposed a new combinatorial approach for the multi-label associative classification problem. Specifically, we developed a 0–1 integer optimization model along with a rule ranking metric and then a heuristic algorithm (SARules) for selecting strong association rules (or decision rules) to form a multi-label classifier. We validated our proposed method for five datasets in different applications and demonstrated the easy-to-interpret results that help decision-making. Our method achieved a competitive classification performance compared to other state-of-the-art machine learning methods. We also designed a decision tool interfaced with the developed SARules for unplanned ICU transfer prediction. For future work, we will consider to integrate domain-knowledge in the constraints of the rule selection model in order to improve practical interpretability. We will also

Table 13
Classification performance (AUC) comparison for five datasets based on 10 times 5-fold cross validation.

Dataset		SARules	ARSOM	LR	DT	RF
ICU	Overall	**0.81(0.04)**/0.79(0.03)	0.78(0.00)	0.73(0.01)	0.72(0.01)	0.77(0.03)
	Infection	0.76(0.06)/0.73(0.06)	0.76(0.01)	**0.79(0.01)**	0.72(0.01)	0.78(0.04)
	Gastrointestinal disease	0.70(0.18)/0.69(0.07)	0.75(0.03)	0.72(0.02)	0.74(0.03)	**0.80(0.05)**
	Cardiovascular and respiratory diseases	**0.66(0.18)**/0.65(0.12)	0.62(0.03)	0.60(0.01)	0.52(0.04)	0.56(0.09)
	Neurological and other diseases	0.75(0.03)/0.72(0.04)	0.73(0.02)	**0.80(0.05)**	0.56(0.09)	0.79(0.08)
Breast Cancer	Overall	**0.76(0.06)**/0.74(0.09)	0.75(0.06)	0.73(0.04)	0.61(0.09)	0.66(0.05)
	Breastquadrant-Left	0.56(0.07)/0.55(0.06)	**0.61(0.09)**	0.51(0.04)	0.51(0.08)	0.51(0.10)
	Breastquadrant-Center	0.70(0.07)/0.69(0.07)	0.70(0.07)	**0.78(0.08)**	0.63(0.08)	0.70(0.08)
	Breastquadrant-Right	0.70(0.15)/0.68(0.16)	0.73(0.04)	0.67(0.18)	0.61(0.16)	**0.73(0.16)**
German credit	Overall	0.70(0.03)/0.67(0.05)	0.68(0.03)	0.67(0.03)	0.64(0.04)	**0.74(0.03)**
	Housing-Free	0.69(0.10)/0.69(0.03)	**0.73(0.08)**	0.63(0.11)	0.59(0.11)	0.68(0.10)
	Housing-Own	0.58(0.06)/0.55(0.01)	0.59(0.04)	0.65(0.05)	0.67(0.04)	**0.73(0.04)**
	Housing-Rent	0.59(0.06)/0.58(0.02)	0.65(0.08)	0.66(0.08)	0.56(0.09)	**0.70(0.08)**
Mushroom	Overall	0.96(0.02)/0.95(0.01)	0.97(0.00)	0.92(0.14)	0.96(0.10)	**0.99(0.01)**
	Gill-spacing close	0.80(0.03)/0.77(0.01)	0.85(0.04)	0.98(0.01)	**0.98(0.02)**	0.97(0.03)
	Gill-spacing crowded	0.96(0.02)/0.95(0.01)	0.96(0.03)	0.99(0.00)	**0.99(0.01)**	**0.99(0.01)**
Stroke	Overall	0.78(0.03)/0.76(0.02)	0.74(0.03)	**0.81(0.02)**	0.62(0.03)	**0.81(0.02)**

extend the current work to generic data format and take into account data uncertainty using robust optimization techniques.

CRediT authorship contribution statement

Yuchun Zou: Conceptualization, Methodology, Validation, Writing – original draft. **Chun-An Chou:** Investigation, Supervision, Conceptualization, Methodology, Method, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is support by Northeastern Faculty Fund. We acknowledge that Dr. Che-Hung Tsai from Taichung Veterans General Hospital in Taiwan provided the ICU dataset for practical validation. We would like to express our sincere thanks to reviewers for their time during the COVID-19 pandemic to provide the constructive feedback to ensure the presentation quality of the work.

References

[1] G. Tsoumakas, I. Katakis, Multi-label classification: An overview, Int. J.f Data Warehous. Min. (IJDWM) 3 (3) (2007) 1–13.
[2] M.-L. Zhang, Z.-H. Zhou, A review on multi-label learning algorithms, IEEE Trans. Knowl. Data Eng. 26 (8) (2013) 1819–1837.
[3] C. Vens, J. Struyf, L. Schietgat, S. Džeroski, H. Blockeel, Decision trees for hierarchical multi-label classification, Mach. Learn. 73 (2) (2008) 185.
[4] M.-L. Zhang, Z.-H. Zhou, Ml-KNN: A lazy learning approach to multi-label learning, Pattern Recognit. 40 (7) (2007) 2038–2048.
[5] M. Borhani, Multi-label log-loss function using L-BFGS for document categorization, Eng. Appl. Artif. Intell. 91 (2020) 103623.
[6] H. Haripriya, C. Prathibhamol, Y.R. Pai, M.S. Sandeep, A.M. Sankar, S.N. Veerla, P. Nedungadi, Multi label prediction using association rule generation and simple k-means, in: 2016 International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT), IEEE, 2016, pp. 159–163.
[7] K. Ali, S. Manganaris, R. Srikant, Partial classification using association rules., in: KDD, vol. 97, 1997, pp. 115–118.
[8] B. Liu, W. Hsu, Y. Ma, et al., Integrating classification and association rule mining., in: KDD, vol. 98, 1998, pp. 80–86.
[9] F. Thabtah, W. Hadi, N. Abdelhamid, A. Issa, Prediction phase in associative classification mining, Int. J. Softw. Eng. Knowl. Eng. 21 (06) (2011) 855–876.
[10] X. Wang, K. Yue, W. Niu, Z. Shi, An approach for adaptive associative classification, Expert Syst. Appl. 38 (9) (2011) 11873–11883.
[11] W. Li, J. Han, J. Pei, CMAR: Accurate and efficient classification based on multiple class-association rules, in: Proceedings 2001 IEEE International Conference on Data Mining, IEEE, 2001, pp. 369–376.
[12] X. Yin, J. Han, CPAR: Classification based on predictive association rules, in: Proceedings of the 2003 SIAM International Conference on Data Mining, SIAM, 2003, pp. 331–335.
[13] A. Veloso, W. Meira, M. Gonçalves, M. Zaki, Multi-label lazy associative classification, in: European Conference on Principles of Data Mining and Knowledge Discovery, Springer, 2007, pp. 605–612.
[14] F.A. Thabtah, P. Cowling, Y. Peng, MMAC: A new multi-class, multi-label associative classification approach, in: Fourth IEEE International Conference on Data Mining, ICDM'04, IEEE, 2004, pp. 217–224.
[15] N. Abdelhamid, Multi-label rules for phishing classification, Appl. Comput. Infor. 11 (1) (2015) 29–46.
[16] C. Prathibhamol, K. Ananthakrishnan, N. Nandan, A. Venugopal, N. Ravindran, A novel approach based on associative rule mining technique for multi-label classification (ARM-MLC), in: Progress in Advanced Computing and Intelligent Engineering, Springer, 2020, pp. 195–203.
[17] D. Bui-Thi, P. Meysman, K. Laukens, MoMAC: Multi-objective optimization to combine multiple association rules into an interpretable classification, Appl. Intell. (2021) 1–13.
[18] B. Li, H. Li, M. Wu, P. Li, Multi-label classification based on association rules with application to scene classification, in: 2008 the 9th International Conference for Young Computer Scientists, IEEE, 2008, pp. 36–41.
[19] S. Sangsuriyun, S. Marukatat, K. Waiyamai, Hierarchical multi-label associative classification (HMAC) using negative rules, in: 9th IEEE International Conference on Cognitive Informatics, ICCI'10), IEEE, 2010, pp. 919–924.
[20] C. Thanajiranthorn, P. Songram, Generation of efficient rules for associative classification, in: International Conference on Multi-Disciplinary Trends in Artificial Intelligence, Springer, 2019, pp. 109–120.
[21] C. Thanajiranthorn, P. Songram, Efficient rule generation for associative classification, Algorithms 13 (11) (2020) 299.
[22] C. Liu, L. Chen, I. Tsang, H. Yin, Towards the learning of weighted multi-label associative classifiers, in: 2018 International Joint Conference on Neural Networks, IJCNN, IEEE, 2018, pp. 1–7.
[23] S. Sangsuriyun, T. Rakthanmanon, K. Waiyamai, Hierarchical multi-label associative classification for protein function prediction using gene ontology, Chiang Mai J. Sci. 46 (1) (2019) 165–179.
[24] R. Agrawal, T. Imieliński, A. Swami, Mining association rules between sets of items in large databases, in: Proceedings of the 1993 ACM SIGMOD international conference on Management of data, 1993, pp. 207–216.
[25] Y. Sun, A.K. Wong, Y. Wang, An overview of associative classifiers., in: DMIN, Citeseer, 2006, pp. 138–143.
[26] F.A. Thabtah, P. Cowling, Y. Peng, Multiple labels associative classification, Knowl. Inf. Syst. 9 (1) (2006) 109–129.
[27] C. Lawless, O. Gunluk, Fair decision rules for binary classification, 2021, arXiv preprint arXiv:2107.01325.
[28] I. Grau, D. Sengupta, A. Nowe, Interpretable semisupervised classifier for predicting cancer stages, in: Machine Learning, Big Data and IoT for Medical Informatics, Elsevier, 2021, pp. 241–259.
[29] F. Valente, S. Paredes, J. Henriques, Personalized and reliable decision sets: Enhancing interpretability in clinical decision support systems, 2021, arXiv preprint arXiv:2107.07483.

[30] B. Letham, C. Rudin, T.H. McCormick, D. Madigan, et al., Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model, Ann. Appl. Stat. 9 (3) (2015) 1350–1371.
[31] T. Wang, C. Rudin, F. Doshi-Velez, Y. Liu, E. Klampfl, P. MacNeille, A bayesian framework for learning rule sets for interpretable classification, J. Mach. Learn. Res. 18 (1) (2017) 2357–2393.
[32] P. Bao, W. Hong, X. Li, Predicting paper acceptance via interpretable decision sets, in: Companion Proceedings of the Web Conference 2021, 2021, pp. 461–467.
[33] H. Ospina-Mateus, L.A.Q. Jiménez, F.J. López-Valdés, S.B. Garcia, L.H. Barrero, S.S. Sana, Extraction of decision rules using genetic algorithms and simulated annealing for prediction of severity of traffic accidents by motorcyclists, J. Ambient Intell. Humaniz. Comput. (2021) 1–22.
[34] H. Lim, J. Lee, D.-W. Kim, Optimization approach for feature selection in multi-label classification, Pattern Recognit. Lett. 89 (2017) 25–30.
[35] H. Bayati, M.B. Dowlatshahi, M. Paniri, MLPSO: a filter multi-label feature selection based on particle swarm optimization, in: 2020 25th International Computer Conference, Computer Society of Iran, CSICC, IEEE, 2020, pp. 1–6.
[36] A. Chang, D. Bertsimas, C. Rudin, An integer optimization approach to associative classification, in: Advances in Neural Information Processing Systems, 2012, pp. 269–277.
[37] C.-A. Chou, Q. Cao, S.-J. Weng, C.-H. Tsai, Mixed-integer optimization approach to learning association rules for unplanned ICU transfer, Artif. Intell. Med. 103 (2020) 101806.
[38] H. Xiong, Association Analysis: Basic Concepts and Algorithms, URL http://www.columbia.edu/~jwp2128/TeachingW, 4721.
[39] A. Gelman, G.O. Roberts, W.R. Gilks, et al., Efficient Metropolis jumping rules, Bayesian Stat. 5 (599–608) (1996) 42.
[40] G.H.L.P. Jeff Schlimmer, Mushroom dataset, https://archive.ics.uci.edu/ml/datasets/mushroom.
[41] Y. Matjaz Zwitter, Milan Soklic, Ljubljana, Breast Cancer dataset, https://archive.ics.uci.edu/ml/datasets/breast+cancer.
[42] P.D.H. Hofmann, German Credit dataset, https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data).
[43] Y. vnter Ljubljana, Stroke dataset, https://www.kaggle.com/asaumya/healthcare-problem-prediction-stroke-patients.
[44] D.W. Hosner, S. Lemeshow, Applied Logistic Regression, Jhon Wiley & Son, New York, 1989.
[45] J. Quinlan, Program for Machine Learning, Morgan Kaufmann Pub, 1993, C4. 5.
[46] T.K. Ho, The random subspace method for constructing decision forests, IEEE Trans. Pattern Anal. Mach. Intell. 20 (8) (1998) 832–844.