Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 2384130, 12 pages
https://doi.org/10.1155/2020/2384130

Research Article

Risk Factors Discovery for Cancer Survivability Analysis Using Graph-Rule Mining

Chaoyu Yang,[1] **Jie Yang**,[2] **and Zhenyu Yang**[3]

[1]*School of Economics and Management, Anhui University of Science and Technology, Huainan 232001, China*
[2]*School of Computing and Information Technology, Faculty of Engineering and Information Sciences, University of Wollongong, Wollongong, NSW 2522, Australia*
[3]*School of Computer Science and Engineering, The University of New South Wales, Sydney, NSW 2052, Australia*

Correspondence should be addressed to Chaoyu Yang; yangchy@aust.edu.cn

Received 20 May 2020; Accepted 13 July 2020; Published 31 July 2020

Academic Editor: Jia-Bao Liu

Mining and understanding patients' disease-development pattern is a major healthcare need. A huge number of research studies have focused on medical resource allocation, survivability prediction, risk management of diagnosis, etc. In this article, we are specifically interested in discovering risk factors for patients with high probability of developing cancers. We propose a systematic and data-driven algorithm and build around the idea of association rule mining. More precisely, the rule-mining method is firstly applied on the target dataset to unpack the underlying relationship of cancer-risk factors, via generating a set of candidate rules. Later, this set is represented as a rule graph, where informative rules are identified and selected with the aim of enhancing the result interpretability. Compared to hundreds of rules generated from the standard rule-mining approach, the proposed algorithm benefits from a concise rule subset, without losing the information from the original rule set. The proposed algorithm is then evaluated using one of the largest cancer data resources. We found that our method outperforms existing approaches in terms of identifying informative rules and requires affordable computational time. Additionally, relevant information from the selected rules can also be used to inform health providers and authorities for cancer-risk management.

1. Introduction

Recent years have witnessed a significantly increasing amount of electronic health records (EHR), in addition to other data collected for the diagnosis and management purpose. The Surveillance, Epidemiology, and End Results (SEER) resource is one of the typical examples. As one comprehensive and authoritative resource in relation to cancer statistics, SEER is a publicly available dataset originating from the United State. This data repository aims to provide high-quality and comprehensive cancer information, in order to help institutions and laboratories worldwide performing their own research. As such, this SEER dataset has been used for a diverse range of research applications, which results in more than 1500 copies for the public use annually. In addition, this SEER repository has also been evolving and updated from time to time, by the increment of new patient samples, the inclusion of more medical features/variables, the involvement of new types of cancers, etc.

Not surprisingly, numerous machine learning methods have been applied to the SEER dataset for monitoring patient status and facilitating a better understanding of cancer treatment and survivability. Prior research efforts include Expert Systems [1], Fuzzy Systems [2, 3], Evolutionary Computation [4, 5], Support Vector Machines [6], and Neural Networks and/or Deep Learning [7]. Yet, there are still open research questions remaining. Expert/Fuzzy Systems, for instance, are typically reliant on human knowledge to determine (semi)static decision strategies. Intuitively, the a priori knowledge may vary from experts to experts, thereby resulting in significantly different outcomes. In addition, knowledge/expertise acquisition could be very time-consuming and labor-expensive, particularly when the scale/dimension of the given problem is large. On the other hand,

Support Vector Machines and Neural Network based approaches are limited in their interpretability, and relevant results are always questionable for end users.

Alternatively, we consider adopting the association rule-mining (ARM) algorithm in this study. As one of the most popular data-mining algorithms, ARM is characterized by its capability of being data-driven (less dependency on the external knowledge, compared to Expert Systems) and interpretability (high transparency compared to Neural Networks). As such, ARM has attracted much research attention with its wide application in many areas, such as the analysis for smart-phone app usage [8], opinion leadership identification [9], and monitoring patients' disease-development behavior [10]. In addition, ARM-based applications in the medical domain can be found in the preliminary research [11, 12].

Yet, one major problem with ARM is the huge number of generated rules; that is, a typical result from ARM could be hundreds and thousands rules associated with different lengths. On the other hand, many rules are overlapping and/or repeating each other with minor changes, which leads to the issue of the rule redundancy. Obviously, a large number of rules is difficult to exam or interpret manually, not to mention its computational overhead, while applying a small set of rules may not be sufficient to capture the underlying pattern, due to the possibility of lacking of information. Consequently, how to control/manipulate the number of generated rules to accurately describe the given dataset becomes a critical process for any ARM-based applications.

Traditionally, there are two strategies in terms of optimizing generated rules: (i) the application of a priori domain knowledge and (ii) rule summarization technique. The former one usually works with predetermined conditions to filter rules, which relies on external resources, such as expert experience or domain knowledge. In this context, only certain items are permitted to be included in generated rules, while others will be cast as unnecessary items to remove. Intuitively, this strategy has two major drawbacks: firstly, identifying important items is time-consuming, particularly with the large number of available items; secondly, experts could impose their own bias via determining item importance, thereby resulting in questionable rules.

On the other hand, rule summarization is a data-driven and automate method, in which less domain knowledge is required. The basic concept of the summarization technique is to identify important rules automatically, from the entire rule set, without compromising the information loss. Some existing work has been reviewed in Section 2. Inspired by the general applicability of rule summarization, this paper explores the task of discovering patients' pattern using the association rule summarization method. To enhance the summarization capability, we further introduce a cluster-based strategy to identify important rules. More specifically, the proposed algorithm consists of three parts. To begin with, we establish a rule graph based on their similarity, in which rules are grouped into different clusters using the community-detection method. Eventually, significant rules are determined and selected across individual rules, which are cast as the output of the proposed summarization. To the best of our knowledge, this is the first study of proposing a cluster-based rule summarization algorithm to reveal the relationship among cancer-related risk factors.

The remainder of the paper is organized as follows. Section 2 provides the brief review about the ground work, such as data-mining-based medical applications; we also discuss traditional techniques for the rule summarization and community-detection clustering approaches. Section 3 presents the proposed cluster-based summarization algorithm, where three major phases are discussed in terms of similarity graph construction, community-detection-based cluster, and applied summarization strategy within each individual cluster. Our proposed framework is then evaluated in Section 4 using the SEER dataset to explore patient risk factors, followed by concluding remarks in Section 5.

2. Related Work

This section offers a brief discussion of the state-of-the-art research work related to the analysis of patients' pattern. At first, we investigate the application of data-mining algorithms in the medical domain. We then discuss the basic concept behind association rule-mining and summarization methods. Finally, we focus on the clustering approach for community detection.

2.1. Data-Mining-Based Medical Application. Recent years have witnessed a vast amount of applications of data-mining techniques in the medical domain [1, 3, 6, 7, 13]. In [1], an expert system was proposed by integrating geographic information and Online Analytical Processing (OLAP) technologies to facilitate environmental health decision support. More precisely, this expert system aimed to investigate potential relations between health problems and environmental risk factors, such as neighborhood, industrial pollutants, and drinking water quality. Another research [6] was proposed to apply a number of supervised learning techniques to discover lung cancer patients in terms of their survivability. Experimental results suggested that the Gradient Boosting Machine led to the best prediction performance, while Support Vector Machine was the only model that generated a distinctive output. In addition, the work [7] investigated the combination of Neural Networks with adversarial domain adaptation. A couple of scenarios were considered in the experiment for the evaluation purpose, including the standard supervised classification, unsupervised domain adaptation, and supervised domain adaptation. Resultant outcome indicated that the hybrid model of Neural Networks and adversarial domain based adaptation achieved satisfactory performance to measure pathology reports. More recently, a Bidirectional Long Short-Term Memory Neural Network (BLSTM-NN) was employed to build an interaction monitoring system in [13]. In their study, ten volunteers were involved and their activities were recorded using a set of Kinect sensors. Then 3D skeletons from participants were detected and tracked using BLSTM-NN, which revealed the underlying activity pattern and interaction among patients. A more general survey was

represented [3] to discuss the various methodologies, such as Fuzzy Logic, Neural Networks, and Genetic Algorithm, and their various applications in medicine.

The majority of existing systems, however, generally are characterized as expert-defined or black-box style. For instance, Expert and Fuzzy Systems require the domain knowledge to setup prediction strategies, which could be very labor-expensive. Neural Network based approaches, on the other hand, are usually limited by their interpretability, which remain questionable for end users. To sum up, despite the general interest of applying data-mining techniques in the medial domain, discovering patient risk factors is still a difficult task. As an alternative, this paper explores the potential of applying association rule-mining-based methods. In particular, rule-based approaches benefit from their transparency, interpretability, and efficient computation, which have potential to overcome the aforementioned limitations from other approaches.

2.2. Rules Mining and Summarization. Association rules mining (ARM) is one of the most-common data-mining algorithms for the relationship analysis. Its goal is to extract rules of the form "IF-Then," such that if a set of variable values is found, then another set of variables will generally have a specific value. A typical example from patients' rule can be "AGE_DX(1), MAR_STAT(1) $\longrightarrow$ SEQ_NUM(0), SRV (> 60)," which indicates if this patient is diagnosed less than 53 years old (i.e., AGE_DX(1)) and is single (i.e., MAR_STAT(1)), then to some extent she/he will have one primary only in the lifetime (i.e., SEQ_NUM(0)) and survival months is more than 60 months (i.e., SRV (> 60)).

As such, the technique is very useful in terms of associating an immediate subsequence (i.e., consequent) given the previous condition (i.e., antecedent) and discovering patterns of interaction among different factors. On the other hand, the importance of a rule is usually estimated through critical indicators, such as "support" and "confidence." Mathematically, given a rule of $(\mathscr{A} \longrightarrow \mathscr{C})$, its support is the proportion of records which contain all items from $\mathscr{A}$, which can be computed as follows:

$$\text{supp}\,(\mathscr{A} \longrightarrow \mathscr{C}) = \text{supp}\,(\mathscr{A}) = \frac{|\mathscr{A}|}{N}, \tag{1}$$

where $|\mathscr{A}|$ is the number of records containing $\mathscr{A}$ and N is the total number of rules. The confidence of the rule $\mathscr{A} \longrightarrow \mathscr{C}$ is accordingly computed as

$$\text{conf}\,(\mathscr{A} \longrightarrow \mathscr{C}) = \frac{\text{supp}\,(\mathscr{A}) \cup \text{supp}\,(\mathscr{C})}{\text{supp}\,(\mathscr{A})}. \tag{2}$$

Consequently, the "support" indicator is used to measure the extent to which the antecedent and consequent occurs simultaneously, while the "confidence" indicator estimates how often the consequent occurs given the antecedent.

Due to its high interpretability and efficiency, plenty of ARM-based applications have been applied for the analysis of smart-phone app usage [8], opinion leadership identification [9, 14], and monitoring patients' disease-development behavior [10], to name a few. In their pilot work of [8], the authors aimed to investigate how students use their smart-phone apps to support online learning. App data from 148 schools were collected accordingly, and the K-means algorithm was employed to separate students into five groups based on their app usage. By mining pattern rules from each cluster, results suggested that students' online patterns showed a shifting ratio between educational and noneducational apps. In addition, generated rules also revealed unique emphases on different types of apps that could impact on student learning performance. The work in [14], on the other hand, investigated a niche subset of user-generated popular culture content on Douban, a well-known Chinese-language online social network. Built on a dataset comprised of 714,946 comments and 228,806 distinct users, a parallel rule-mining algorithm was proposed. The experimental results demonstrated the flexibility and applicability of the rule-based method for extracting useful relationship from complex social media data. In addition, another work to explore patient's behavior in terms of disease complications and recurrences was reported in [10]. For this particular research, a database about colorectal cancer, with 1516 patients and 126 attributes, was considered. At its core, four heuristic operators and a complete methodology were proposed to implement the rule-mining process. From the experiments, the rule-based approach showed advantages over standard approaches, such as the associative classification methods to identify risk factors.

The major problem, however, with the traditional ARM is the huge number of generated rules, which is manually inefficient to exam them one-by-one. The large number of rules, on the other hand, also reduces the interpretability as a whole. To overcome this problem, one established approach is the rule summarization, i.e., to summarize rules based on their significance without degrading the relationship information from the expression of the entire rule set. There are a few implementations for summarizing important rules, including APRX-COLLECTION [15] and RPGlobal [16]. To begin with, APRX-COLLECTION introduces a concept of the false positive rate (α), which is used to control the level of wrong coverage. More precisely, APRX-COLLECTION firstly forms an aggregated rule set $\mathscr{R}^*$ by enumerating all possible combinations of original rules $\mathscr{R}$. As such, additional rules, even though they might not exist in $\mathscr{R}$, could be created. Next, rules from $\mathscr{R}^*$ are selected according to two criteria: (i) those rules cover most items from $\mathscr{R}$and (ii) the number of additional items should be less than the level of α. On the other hand, RPGlobal adopts similar selection criteria by identifying rules that cover most items from $\mathscr{R}$. The major difference is that RPGlobal chooses rules from $\mathscr{R}$ directly, instead of generating $\mathscr{R}^*$. In addition, to limit the increment of additional rules, RPGlobal further introduces a user-defined parameter β to control the number of selected rules each time.

Overall, rule-mining summarization techniques have been proposed to overcome the problem associated with the huge number of generated rules. The summarized rules make the subsequent interpretation process more efficient and easier, by filtering least-important rules and significantly reducing the scale of rules. Inspired by this insight, an

enhanced rule summarization method is proposed in this study, which is in light of clustering, in particular, the community-detection technique.

2.3. Community Detection. As a graph clustering technique, community detection has attracted a lot of attention in the past decade due to the increasing scale of social network. With the rapid growth of Internet infrastructure, more and more people utilize online resources in their daily life, such as Twitter and Facebook. The result is the generation of a huge social network, in which individual users play a role of nodes/vertexes and their connections (e.g., friendship) become the edge in the network. As such, either the industry stakeholders or academia are interested in analyzing this giant social network to formulate better marketing and/or development strategy. In particular, identifying communities within complex networks is of great importance for many real scenarios. A typical example could be forming an online community in relation to a group of people who share the same interest.

A number of different methods have been proposed to implement the community detection. A pioneer work in [17] was proposed based on the concept of edge betweenness centrality (EBC). For each individual edge within the network, its EBC was measured by the total number of shortest paths (for any two vertices) passing this particular edge. As a result, an edge with higher EBC became a good indicator to separate among communities, while an edge with lower EBC was more likely to exist within a small community. By removing edges with high EBC, the entire network eventually could be split into small groups/communities. Another work was reported in [18] with a similar measurement, that is, edge clustering coefficient (ECC). This measurement was to count the number of triangles for a given edge, compared to the total number of such possible triangles. Compared to EBC, edges with low ECC were considered as the connections among communities. As such, disjoint subnetworks can be formed by eliminating those low-ECC edges.

In addition to edge-based measurements, the Walktrap algorithm from [19] considered the topological similarity between vertices. The main idea was to divide the network based on a distance between vertices such that distance in the same community was small but became larger in different groups. This vertex distance was formally defined by (i) the walking probability from one vertex to another and (ii) the vertex degree. Another vertex-based algorithm was proposed in [20], termed as Label Propagation. To begin with, every vertex was randomly initialized with a unique label. Later, during the iteration, each vertex adjusted its label based on neighbors; that is, new label will be made the same as its majority labels nearby. Finally, communities were formed by grouping vertices with the same labels. The Infomap algorithm, on the other hand, was proposed using the concept of random walks and information diffusion [21]. It started by performing a random walk within the network and calculated the information flows using the trajectory of the previous random walk. An information map was accordingly established, which differentiated communities with a diverse range of map importance. One advantage with this Infomap algorithm is its nearly linear-computational time, thereby leading to a very efficient process.

2.4. Summary. In this section, we briefly discuss the existing work of applying data-mining techniques to address medical problems. We also review the basic concept of rule-mining, rules summarization, and the community-detection approaches. Although preliminary work has been conducted to identify rules in relation to patients risk management, traditional rule-mining algorithms suffer from a major problem associated with the huge number of generated rules. To cope with this issue, rules summarization techniques offer advantage to select important rules and minimize the information loss. Taking all these aspects into account, we propose an enhanced summarization algorithm using the community-detection approaches, which is detailed in the following section.

3. The Proposed Framework

In this section, we discuss a systematic and data-driven approach to discover risk-relevant factors. The main contribution of this study is the proposal of a novel cluster-based summarization algorithm. As illustrated in Figure 1, the proposed approach consists of three phases. To begin with, we apply the traditional rule-mining algorithm on the entire dataset to generate a comprehensive set of potential rules. Due to the large scale of this rule set, we then represent it as a rule-similarity graph; see Section 3.1. Secondly, the community-detection algorithm is employed to identify clusters from this rule graph; see Section 3.2. Finally, informative rules across clusters are summarized, as introduced in Section 3.3.

For convenience, Table 1 summarizes notations used throughout the paper.

3.1. Similarity Graph. The main purpose of this first phase is to generate a completed rule set that represents the entire transaction records and then to construct a rule-similarity graph. Towards this end, there are several steps we need to consider, including data discretization, rule-mining, similarity measurement, and graph construction.

3.1.1. Data Discretization. To begin with, the rule-mining algorithm works well with discrete data, rather than the continuous ones. However, in the real-world scenario, the majority of medical data is continuous and not operable by the rule-mining approaches. To quantify extracted features, a preprocess of data discretization is necessary. For simplicity, this study aims to split a continuous input into L groups (where L is a user-defined threshold). As such, samples belonging to the same group will be assigned with the same label, to convert the continuous data into discrete one. Note that a domain knowledge is required to decide the number of groups (i.e., L), while different business or operational requirements could result in a variety of discretization ranges.

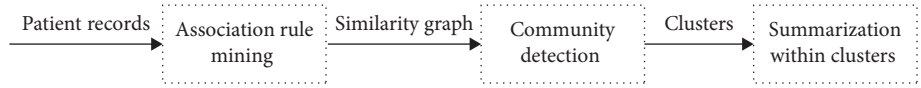

Figure 1: The pipeline of the proposed rule-summarization algorithm, including three phases: (i) applying association rule-mining algorithm to obtain the completed rule set, which is turned into the rule-similarity graph; (ii) employing the community-detection algorithm to cluster the rule graph; (iii) summarizing significant rules from individual clusters.

Table 1: List of adopted notations used in our study.

Notation	Description
$\mathcal{I}$	The complete itemset from the original data
N	Number of transaction records
M	Number of rules generated from these N samples
K	Number of clusters to be found within the rule graph
N_s	Number of rules to be summarized/selected from one cluster

However, the advantage of the discretization is twofold: (i) continuous data is represented using discrete labels to facilitate the subsequent application of the rule-mining algorithm; (ii) the raw continuous dataset is represented by a smaller-sized but meaningful format, which is easy to be interpreted and also saves the computational cost.

3.1.2. Rule Mining. There exists a diverse range of implementations for rule mining, such as Apriori and FP-Growth. In particular, Apriori employs a "bottom-up" strategy to produce frequent-item sets, in which repeated scanning of the entire dataset is required. This typically leads to an expensive computational cost. Therefore, in this study, the FP-Growth algorithm is implemented, which adopts a "top-down" strategy to produce frequent-item sets. The main advantage is that it requires less scanning time to generate possible combinations of frequent sets.

3.1.3. Similarity Measurement. Before we construct the rule-similarity graph, it is essential to define the similarity measurement for any given rules. Consider the typical form of two rules r_1: $(\mathcal{A}_1 \longrightarrow \mathcal{C}_1)$ and r_2: $(\mathcal{A}_2 \longrightarrow \mathcal{C}_2)$. The similarity function between two rules is accordingly defined in terms of the relative item coverage (RIC):

$$\mathrm{RIC}(r_1, r_2) = \frac{\left(\left\|\mathcal{A}_1 \cap \mathcal{A}_2\right\|\right) \cup \left(\left\|\mathcal{C}_1 \cap \mathcal{C}_2\right\|\right)}{\left\|\mathcal{A}_1 \cup \mathcal{C}_1 \cup \mathcal{A}_2 \cup \mathcal{C}_2\right\|}. \tag{3}$$

As observed, the similarity is measured as the portion of the common items from both antecedents and consequences versus the portion of all items occurring within two rules. We then introduce the process of rule-graph construction based on the similarity measurement in equation (3).

3.1.4. Graph Construction. Graph is a very important data structure in computer science, while a large number of existing works have been proposed to demonstrate the solid application and success of graph-based techniques [2, 4]. Inspired by this insight, we also consider representing rules as a graph format. As such, the rule graph is represented as $G = (V, E)$ in our study, where each vertex v_i $(v_i \in V)$ denotes a rule r_i, and the edge e_{ij} is the connection between the i-th and j-th vertex. Furthermore, e_{ij} is associated with the similarity between the rule of r_i and r_j, i.e., $\mathrm{RIC}(r_i, r_j)$. Note that there are a diverse range of options to manipulate e_{ij}. For instance, we can introduce a user-defined threshold ε to filter $\mathrm{RIC}(r_i, r_j)$ if $\mathrm{RIC}(r_i, r_j) < \varepsilon$. That is, two vertices are only connected if their similarity $\mathrm{RIC}(r_i, r_j)$ is larger than the value of ε. Alternatively, we can also employ the concept of k-nearest neighbors, from which only the most similar $(k-1)$ vertices to one specific vertex are connected. Without loss of generality, we consider the full-connect strategy; that is, all vertices will be connected to each other, while the edge e_{ij} equals their similarity $\mathrm{RIC}(r_i, r_j)$.

Table 2: Comparison of different community-detection implementations in terms of their computational complexity.

Algorithm	Cost	Reference
Edge betweenness centrality	$\mathcal{O}(m^2 n)$	[17]
Edge clustering coefficient	$\mathcal{O}(m^4/n^2)$	[18]
Walktrap	$\mathcal{O}(n^2 \log(n))$	[19]
Label propagation	$\mathcal{O}(m+n)$	[20]
Infomap	$\mathcal{O}(m)$	[21]

Note that n and m represent the number of vertexes and edges, respectively.

3.2. Rule Clustering. This second phase intends to apply the community-detection algorithm to identify clusters from the rule-similarity graph. The general idea is to cluster vertices in a way that samples belonging to the same group are similar, while samples from different groups are dissimilar to each other. As mentioned in Section 3.3, there has been a great interest in developing implementations for detecting communities within the graph. Table 2 summarizes the computational complexity of the existing implementations for the community detection.

As observed, a variety of implementations may lead to different costs, as they focus on different optimization strategies on splitting vertices and/or edges. Considering our case of rule-based graph, there will be over 10^4 rules, indicating the final graph is with 10^4 vertices and approximately 10^8 edges. As such, we implement the work [21] in this study as our community-detection executor, due to its efficiency and affordable computation.

3.3. Cluster-Based Summarization. The final phase is used to perform rule summarization by determining or selecting important rules within each individual cluster. More precisely, after forming clusters within the rule graph, rules will be ranked according to their importance. As such, top-N rules from each single cluster will be selected according to

Input: K clusters (C_1, C_2, ..., C_K), the number of selected rules N_s, and the penalty terms of λ_L, λ_A, and λ_R;
Initialization:
  set the rule set S to empty: $S = \varnothing$
for $k = 1$ **to** K **do**
  While ($\|C_k\| \neq 0$ or $\|S\| \leq N_s$) **do**
    Calculate the score of r_i ($\forall r_i \in C_k$) according to equation (7)
    Select one rule with the highest score and label it as r^*
    $S \longleftarrow r^*$ and remove r^* from C_k
  end
end
Output: Return the optimal rules from S.

Algorithm 1: The proposed cluster-based algorithm for rule summarization.

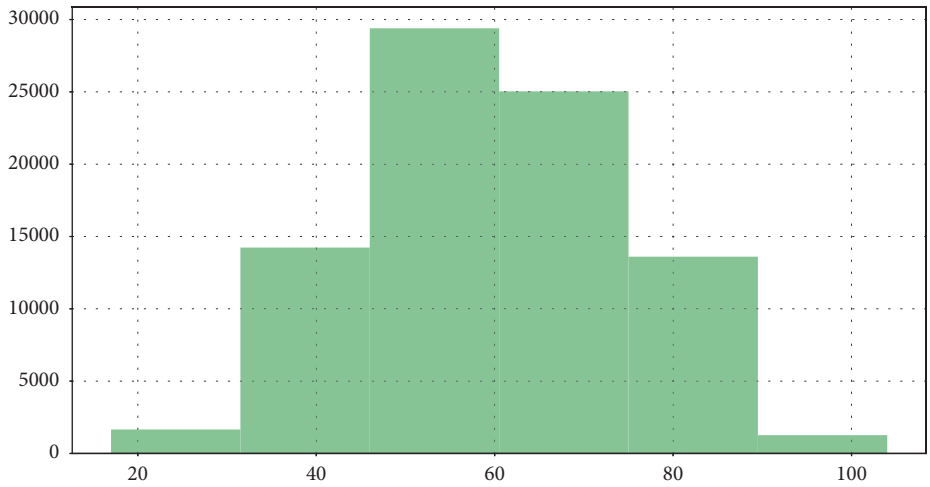

Figure 2: Data distribution of AGE_DX.

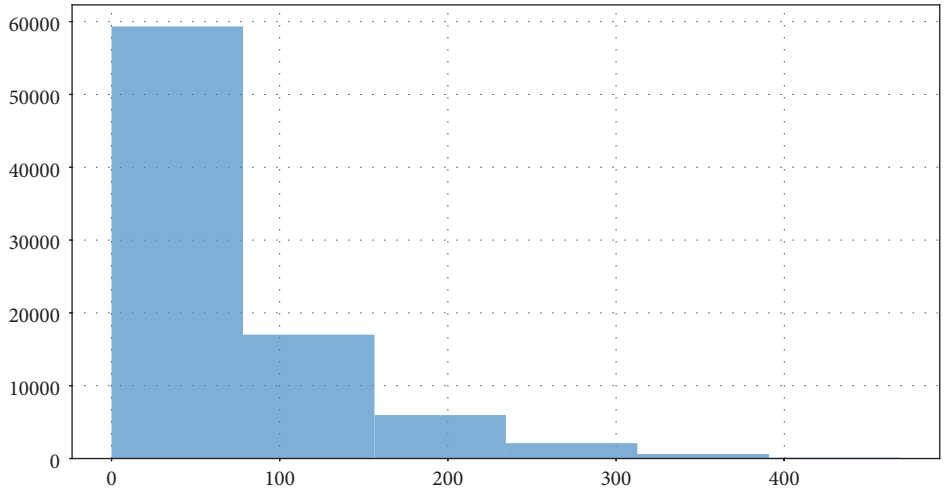

Figure 3: Data distribution of SRV_TIME.

their importance, where N is a user-determined parameter, to produce the final summary. To begin with, we propose three statistical features before aggregating them to measure the rule importance:

(i) Rule length: the first importance measurement is the rule length. This idea is inspired by the work of APRX-COLLECTION in [15], where a long rule has a better item coverage compared to shorter ones. Therefore, we use the rule length as a feature to

estimate the importance. Given a rule of (r_i: $\mathscr{A} \longrightarrow \mathscr{C}$), the feature of length (Length(r_i)) is computed as follows:

$$\text{Length}(r_i) = \frac{1}{\|\mathscr{I}\|}(\|\mathscr{A}\| + \|\mathscr{C}\|), \tag{4}$$

where $\|\mathscr{I}\|$ represents the total number of distinct items from all rules.

(ii) Aggregated similarity: we want to select informative rules that would represent the majority rules within a cluster. As such, rules should be selected if they are with a similar or close form to the rest. Therefore, the second measurement to the rule importance is the aggregated similarity. More specifically, the larger the aggregated similarity of a rule, the higher the rank. Given a cluster C, the aggregate similarity (Aggregate(r_i)) for the i-th rule ($r_i \in C$) is then given as follows:

$$\text{Aggregate}(r_i) = \frac{\sum_{j=1, i\neq j}^{\|C\|} \text{RIC}(r_i, r_j)}{\|C\|}, \tag{5}$$

where RIC(r_i, r_j) is the similarity measurement from (3) and $\|C\|$ represents the total number of rules from the C cluster.

(iii) Redundancy: another critical aspect is to consider the redundancy impact once a rule is selected. Note that each cluster is composed of similar rules. Therefore, if two rules contain similar items, then these two are more likely to convey the similar meaning. The redundancy feature is then employed to eliminate those semantically similar rules, without repeating the same rules all the time. In our study, the following estimation is proposed to measure the redundancy feature (Redundancy(r_i)):

$$\text{Redundancy}(r_i) = \frac{\sum_{j=1}^{\|\mathcal{S}\|} \text{RIC}(r_i, r_j)}{\|\mathcal{S}\|}, \quad \text{if } \|\mathcal{S}\| \neq 0, \tag{6}$$

where $\mathcal{S}$ is the set composed of selected rules and $\|\mathcal{S}\|$ represents the number of rules from S. Note that, at the beginning of the summarization, $\mathcal{S} = \varnothing$ and $\|\mathcal{S}\| = 0$. In this case, Redundancy$(r_i) = 0$ ($\forall r_i \in \mathcal{S}$).

Apart from aforementioned measurements, we further leverage the support degree as an indicator for evaluating the rule importance. Consequently, the final score to one specific rule (i.e., Score(r_i)) is to involve four statistical features, including support degree, rule length, aggregated similarity, and redundancy, and the following equation is proposed to formulate this calculation:

Table 3: List of selected variables from the breast cancer data.

Variables	Description
MAR_STAT	Marital status
SEX	Gender
SEQ_NUM	Seq. of malignant
LATERAL	Laterality
NUMPRIMS	Number of primaries
SRV_TIME	Survival months
ORIGIN	Origin
AGE_DX	Age at diagnosis
SITEO2V	Primary site
RADIATN	Method of radiation therapy
HISTREC	Histology
ADJAJCCSTG	AJCC 6th stage

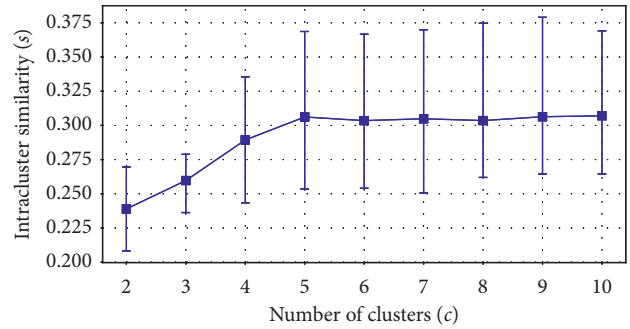

Figure 4: Intercluster similarity $\mathcal{S}_{\text{inter}}$ based on varying values of the number of clusters (c), while the upper and lower bound represent the maximal and minimal values of $\mathcal{S}_{\text{inter}}$. The related computational time is 426.07 s, 564.11 s, 710.78 s, 870.91 s, 1027.49 s, 1188.15 s, 1345.25 s, 1501.67 s, and 1658.97 s, respectively.

Table 4: The completed rule sets generated using 85,189 patient samples, with different settings of support degrees.

Support degrees	Number of rules	Support degrees	Number of rules
(0.0, 0.1]	9096	(0.1, 0.2]	2043
(0.2, 0.3]	287	(0.3, 0.4]	196
(0.4, 0.5]	125	(0.5, 0.6]	93
(0.6, 0.7]	20	(0.7, 0.8]	16
(0.8, 0.9]	10	(0.9, 1.0]	1

Table 5: Statistical information from the result of community-detection clustering when $K = 5$.

Cluster index	Number of rules	Average RIC within one cluster
0	1477	0.2535
1	2640	0.3135
2	2534	0.2949
3	3056	0.2903
4	2180	0.3687

Table 6: List of selected rules from the proposed algorithm, showing the top 15 rules from five clusters.

Index	Rule	Support	Confidence
0	(SEQ_NUM(0), HISTREC(9), RADIATN(0)) ⟶ (SEX(2), NUMPRIMS(1), ORIGIN(0))	0.428	0.942
0	(NUMPRIMS(1), HISTREC(9), RADIATN(0)) ⟶ (SEX(2), SEQ_NUM(0), ORIGIN(0))	0.428	0.938
0	(NUMPRIMS(1), RADIATN(0)) ⟶ (SEX(2), ORIGIN(0))	0.509	0.944
1	(NUMPRIMS(1), HISTREC(9), ORIGIN(0)) ⟶ (SEX(2), SEQ_NUM(0))	0.695	0.987
1	(SEX(2), SEQ_NUM(0), HISTREC(9)) ⟶ (NUMPRIMS(1), ORIGIN(0))	0.695	0.947
1	(SEX(2), NUMPRIMS(1), HISTREC(9)) ⟶ (ORIGIN(0))	0.699	0.948
2	(NUMPRIMS(1), HISTREC(9)) ⟶ (SEX(2), SEQ_NUM(0))	0.733	0.987
2	(NUMPRIMS(1), HISTREC(9)) ⟶ (SEX(2))	0.737	0.993
2	(NUMPRIMS(1)) ⟶ (SEX(2), SEQ_NUM(0))	0.860	0.987
3	(SEQ_NUM(0), ORIGIN(0)) ⟶ (SEX(2), NUMPRIMS(1))	0.816	0.991
3	(NUMPRIMS(1), ORIGIN(0)) ⟶ (SEX(2), SEQ_NUM(0))	0.816	0.987
3	(SRV(< 60)) ⟶ (ORIGIN(0))	0.574	0.948
4	(SRV(< 60)) ⟶ (SEX(2), ORIGIN(0))	0.569	0.940
4	(HISTREC(9), ORIGIN(0)) ⟶ (SEX(2))	0.807	0.993
4	(RADIATN(0)) ⟶ (SEX(2))	0.616	0.993

Table 7: List of selected rules based on their support/confidence degree.

Rule	Support	Confidence
(ORIGIN(0)) ⟶ (SEX(2))	0.943	0.992
(SEQ_NUM(0)) ⟶ (NUMPRIMS(1))	0.866	0.999
(NUMPRIMS(1)) ⟶ (SEQ_NUM(0))	0.866	0.995
(NUMPRIMS(1)) ⟶ (SEX(2))	0.864	0.992
(SEQ_NUM(0)) ⟶ (SEX(2))	0.860	0.992
(NUMPRIMS(1)) ⟶ (SEX(2), SEQ_NUM(0))	0.860	0.987
(SEQ_NUM(0)) ⟶ (SEX(2), NUMPRIMS(1))	0.860	0.991
(SEX(2), NUMPRIMS(1)) ⟶ (SEQ_NUM(0))	0.860	0.995
(SEX(2), SEQ_NUM(0)) ⟶ (NUMPRIMS(1))	0.860	0.999
(NUMPRIMS(1), SEQ_NUM(0)) ⟶ (SEX(2))	0.860	0.992
(HISTREC(9)) ⟶ (SEX(2))	0.851	0.993
(ORIGIN(0), SEQ_NUM(0)) ⟶ (NUMPRIMS(1))	0.822	0.999
(ORIGIN(0), NUMPRIMS(1)) ⟶ (SEQ_NUM(0))	0.822	0.995
(ORIGIN(0), NUMPRIMS(1)) ⟶ (SEX(2))	0.820	0.992
(ORIGIN(0), SEQ_NUM(0)) ⟶ (SEX(2))	0.817	0.992

$$\text{Score}(r_i) = \text{Supp}(r_i) + \lambda_L \text{Length}(r_i) + \lambda_A \text{Aggregate}(r_i) - \lambda_R \text{Redundancy}(r_i), \tag{7}$$

where λ_L, λ_A, and λ_R are the penalty term for balancing four statistical features, respectively. Eventually, the cluster-based summarization algorithm is proposed for determining informative rules, which is further shown in Algorithm 1.

3.4. Summary. The main contribution of this work is to formulate the problem of rule summarization as a graph clustering process. Next, we discuss the computational complexity of the proposed method. Given a dataset with N samples, the FP-Growth algorithm is employed to mining potential rules with the cost of $\mathcal{O}(N)$. Next the rule-similarity graph is constructed, which requires a complexity of $\mathcal{O}(M^2)$ (where M is the total number of generated rules; note that it usually leads to $N \ll M$). Note that with the established graph, there will be M vertex and $M(M-1)/2$ edges as we consider the full-connect strategy. As such, applying the community-detection algorithm to cluster this rule graph costs $\mathcal{O}(M^2)$. Finally, the cluster-based summarization algorithm needs to go through all K clusters to select top N_s rules. As such, for the k-th cluster, the time complexity could be $\mathcal{O}(\|C_k\| \cdot \min(N_s, \|C_k\|) \approx \mathcal{O}(\|C_k\| \cdot N_s))$, where $\|C_k\|$ is the number of rules within the k-th cluster. In the worst case, we have $\|C_k\| = M$, thereby leading to the worst complexity of $\mathcal{O}(M \cdot N_s)$. Overall, the complexity order of the proposed algorithm is $\mathcal{O}(N) + \mathcal{O}(M^2) + \mathcal{O}(M^2) + \mathcal{O}(K \cdot M \cdot N_s) \approx \max(\mathcal{O}(M^2), \mathcal{O}(K \cdot M \cdot N_s))$.

Notice that the overall complexity for the proposed algorithm depends on either the total number of generated rules (i.e., M) or the number of available clusters and rules to be selected. In the worst case, the complexity could be $\mathcal{O}(M^2)$ if we select all rules; by contrast, it will be $\mathcal{O}(K \cdot M)$ as only one rule to be chosen from individual cluster.

4. Experimental Results and Analysis

This section discusses the experimental results by performing the application of our proposed algorithm to the SEER dataset. The details about the employed dataset are presented in the following section. The aim of the

Table 8: Results from traditional rule summarization techniques, using the APRX-COLLECTION method.

Rule from APRX-COLLECTION	Support	Confidence
(SEX(2), RADIATN(0), LATERAL(1), MAR_STAT(2), SEQ_NUM(0)) ⟶ (ORIGIN(0), NUMPRIMS(1), HISTREC(9))	0.108	0.828
(RADIATN(1), LATERAL(2), SEQ_NUM(0), HISTREC(9), NUMPRIMS(1)) ⟶ (ORIGIN(0))	0.119	0.945
(SEX(2), MAR_STAT(5), SRV(< 60)) ⟶ (NUMPRIMS(1), SEQ_NUM(0))	0.122	0.918
(SRV(< 60), C509, RADIATN(0)) ⟶ (ORIGIN(0))	0.109	0.943
(AGE_DX(2), MAR_STAT(2)) ⟶ (SEX(2))	0.106	0.992
(C508, LATERAL(2)) ⟶ (SEX(2))	0.101	0.998
(ADJACCSTG(70.0), ORIGIN(0)) ⟶ (SEX(2))	0.101	0.990
(NUMPRIMS(2), SEQ_NUM(1)) ⟶ (HISTREC(9))	0.101	0.883
(C504, LATERAL(1)) ⟶ (ORIGIN(0))	0.138	0.953
(SRV(> 60), AGE_DX(3)) ⟶ (ORIGIN(0))	0.103	0.964
(MAR_STAT(1), ORIGIN(0)) ⟶ (NUMPRIMS(1))	0.102	0.876
(AGE_DX(1), SRV(> 60)) ⟶ (ORIGIN(0))	0.137	0.944
(NUMPRIMS(2)) ⟶ (SEQ_NUM(1))	0.115	0.993
(MAR_STAT(5)) ⟶ (ORIGIN(0))	0.189	0.958
(ADJACCSTG(70.0)) ⟶ (NUMPRIMS(1))	0.102	0.940

experiments is to (i) evaluate the influence of key parameters to the clustering performance and (ii) compare the proposed algorithm with existing work for the rule summarization.

4.1. Experiments Design. As mentioned before, the SEER dataset consists of samples from a great number of cancer types. In this study, the breast cancer is explicitly employed as the main resource. Relevant samples and variables from SEER that are associated with patient survivability and tumor status are selected based on a set of inclusions (the selection criteria for variables can be found in our preliminary work [22]). As such, 85,189 patient samples (with 12 variables) are identified as the main experimental data, and detailed description for chosen variables can be found in Table 3.

Note that among these 12 variables, two of them are with the continuous type, i.e., AGE_DX and SRV_TIME, and Figures 2 and 3 illustrate their distribution, respectively. In relation to the continuous data, the data discretization process is considered to split them into L groups, while each group will be assigned to one unique label. In this study, we set $L = 3$ and 2 for the variables of AGE_DX and SRV_TIME, respectively. As a result, the discrete results for AGE_DX will be [17, 53), [53, 67), and [67, 104] due to an equal-size separation. Meanwhile, as for the variable of survival months, we are taking SRV_TIME = 60 as the splitting threshold, as the majority of studies has categorized patients' survivability using a threshold of five years [10, 22].

4.2. Results from Cluster-Based Summarization. In this section, we focus on the results from the proposed algorithm, with respect to the rule summarization for patients' behavior. Towards this end, we start by examining the completed rule sets and then performing rules cluster, and more importantly we investigate the summarization results across different clusters. To begin with, we utilize the FP-Growth algorithm to generate the completed rule set, which leads to a total of 11,887 rules. Table 4 shows the generated rules as a function of the support degrees, while a higher support is normally with a smaller number of rules. Note that the threshold for the confidence degree is fixed as 0.5 in all cases.

We then perform the community detection to cluster the generated rule set. A particular problem with the application of community-detection clustering is to determine the number of clusters (K). Yet, the selection of an appropriate value for the number of clusters has crucial impact on the clustering performance. Having said that, a too big value for K would make it difficult to interpret the result, not to mention the computational cost; while a smaller value of K could fail to group similar samples and result in poor clustering performance. To identify an optimal value for the number of clusters, the measurement of intercluster similarity ($\mathcal{S}_{\text{inter}}$) is introduced that is estimated as follows:

$$\mathcal{S}_{\text{inter}} = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{N_k^2}\sum_{i=1}^{N_k}\sum_{j=1,j\neq i}^{N_k}\text{RIC}\left(r_i, r_j\right)\right), \tag{8}$$

where RIC represents the relative item coverage (defined in equation (3)), N_k represents the number of rules in the k-th cluster ($k = [1, 2, \ldots, K]$), and K is the number of clusters.

Using this measurement, our aim is to select the value of K that leads to the biggest value of $\mathcal{S}_{inter}$, so that similar rules are grouped together to maximize the intercluster similarity. Towards this end, Figure 4 plots the results of intercluster similarity for varying values of K, where $K \in [2, 10]$. At first, we confirm that the computational time associated with different sizes of clusters is increasing with respect to K. For instance, the minimal and maximal time has been found from $K = 2$ (with 426.07 seconds) and $K = 10$ (with 1658.97 seconds), respectively. On the other hand, we observe that the intercluster similarity ($\mathcal{S}_{\text{inter}}$) performs stably after $K \geq 5$.

Given the expensive computation with a bigger value of K, we decide to take the optimal value of $K = 5$ in the following study. As such, we perform the community-detection algorithm to cluster rules into five groups, and the statistical information about the particular clustering result is summarized in Table 5.

Table 9: Results from traditional rule summarization techniques, using the RPGlobal method.

Rule from RPGlobal	Support	Confidence
(RADIATN(0), LATERAL(2), SEQ_NUM(0), SRV(< 60), ORIGIN(0), HISTREC(9), NUMPRIMS(1)) ⟶ (SEX(2))	0.135	0.993
(SEX(2), RADIATN(0), LATERAL(2), SRV(< 60), ORIGIN(0), HISTREC(9), NUMPRIMS(1)) ⟶ (SEQ_NUM(0))	0.135	0.998
(RADIATN(1), MAR_STAT(2), SEQ_NUM(0), ORIGIN(0), HISTREC(9), NUMPRIMS(1)) ⟶ (SEX(2))	0.128	0.991
(SEX(2), LATERAL(1), MAR_STAT(2), SRV(< 60), ORIGIN(0), HISTREC(9)) ⟶ (SEQ_NUM(0))	0.104	0.923
(SEX(2), AGE_DX(2), RADIATN(0), SEQ_NUM(0), SRV(< 60)) ⟶ (NUMPRIMS(1))	0.101	0.999
(SEQ_NUM(0), AGE_DX(3), HISTREC(9), MAR_STAT(5), NUMPRIMS(1)) ⟶ (SEX(2))	0.101	0.996
(SEX(2), SEQ_NUM(0), AGE_DX(3), HISTREC(9), MAR_STAT(5)) ⟶ (NUMPRIMS(1))	0.101	0.998
(SEX(2), C504, ORIGIN(0), MAR_STAT(2)) ⟶ (HISTREC(9))	0.135	0.905
(SRV(> 60), NUMPRIMS(1), HISTREC(9)) ⟶ (ORIGIN(0))	0.127	0.962
(AGE_DX(1), NUMPRIMS(1), SEQ_NUM(0), RADIATN(1)) ⟶ (SEX(2))	0.107	0.995
(SEX(2), SRV(> 60), RADIATN(0), LATERAL(2)) ⟶ (HISTREC(9))	0.112	0.893
(SEX(2), C509, ORIGIN(0), LATERAL(1)) ⟶ (NUMPRIMS(1))	0.126	0.881
(NUMPRIMS(1), MAR_STAT(1), ORIGIN(0), SEQ_NUM(0)) ⟶ (SEX(2))	0.123	0.991
(MAR_STAT(1)) ⟶ (SEX(2), NUMPRIMS(1), SEQ_NUM(0))	0.106	0.864
(SEX(2), C508, RADIATN(0)) ⟶ (ORIGIN(0))	0.116	0.953

Table 10: Selected item sets and their descriptions (ordered alphabetically) from our study.

Item	Description
AGE_DX(1)	Age at diagnosis less than 53
AGE_DX(2)	Age at diagnosis more than or equal to 53 and less than 67
AGE_DX(3)	Age at diagnosis more than or equal to 67
ADJACCSTG(70.0)	Breast adjusted AJCC 6th stage (1988+)–IV
C50(4,8,9)	Breast
HISTREC(9)	Histology recode—broad groupings (8500–8549)
LATERAL(1)	Not a paired site
LATERAL(2)	Right: origin of primary
MAR_STAT(1)	Single (never married)
MAR_STAT(2)	Married (including common law)
MAR_STAT(5)	Widowed
NUMPRIMS(1)	One primary
NUMPRIMS(2)	Two primaries
ORIGIN(0)	Non-Spanish/Non-Hispanic
RADIATN(0)	None or diagnosed at autopsy
RADIATN(1)	Beam radiation
SEQ_NUM(0)	One primary only in the patient's lifetime
SEQ_NUM(1)	First of two or more primaries
SEX(2)	Female
SRV (< 60)	Survival months less than 60 months or equal
SRV (> 60)	Survival months more than 60 months

Next, the proposed summarization algorithm is performed on these five clusters to identify informative rules. In this study, we are setting key parameters for summarization as follows: the number of selected rules $N_s = 3$, and the penalty terms of $\lambda_L = \lambda_A = \lambda_R = 0.35$. As a result, summarized rules are listed in Table 6, which shows a diverse coverage of support degree and number of items. For instance, all selected rules, together, cover nearly 98% of patient samples (high support degree) while seven distinct items occur from the results that are identified as key items, including SEQ_NUM(0), HISTREC(9), RADIATN(0), SEX(2), NUMPRIMS(1), ORIGIN(0), and SRV (< 60). We will then compare our proposed approach with others.

4.3. Comparison Results. In this section, we compare the proposed algorithm with the existing approaches in terms of mining risk factors associated with patients' disease development. To begin with, we first extract top 15 rules from the completed rule set (without any summarization techniques), by simply ranking them based on their support degrees. As such, these rules are cast as the baseline results, and Table 7 illustrates this rule set with high support degree.

As observed, the majority of rules from the baseline results is overlapping each other. For instance, there are only five items observed from both the antecedent and consequent, including "SEX(2)," "NUMPRIMS(1)," "ORIGIN(0)," "SEQ_NUM(0)," and "HISTREC(9)," respectively. That is, approximately 88.1% of the items have been repeated

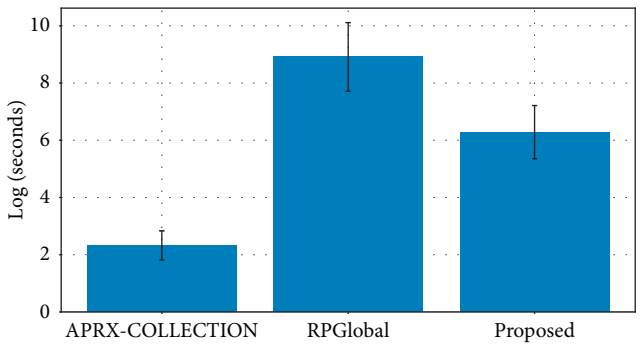

Figure 5: Comparison results from different summarization approaches in terms of their computational cost.

from this baseline result. On the other hand, rules from Table 7 are with relatively simple format (with no more than 3 items), which also indicates we could miss some complex or advanced rules. More importantly, although these top 15 rules are selected based on their support degree, their relevant data coverage is relatively low compared to our method (with 94.6% of the entire data). By contrast, the proposed method performs better than the baseline rules, by identifying more key items (seven) and covering a larger number of patients (98.3%).

Next, traditional rule summarization techniques are considered, including APRX-COLLECTION [15] and RPGlobal [16] method, while their results are shown in Tables 8 and 9, respectively. Again, those methods are applied to summarize rules by selecting the top 15 ones.

The results from the APRX-COLLECTION algorithm clearly indicate its preference of selecting long rules, regardless of their support. As mentioned before, the principle of APRX-COLLECTION is to choose rules with a large item coverage. As a result, the top two rules from APRX-COLLECTION, for instance, are with 8 and 6 items, respectively. However, the major problem with APRX-COLLECTION is the data coverage, that is, the support degree from rules. All selected rules are with support under the value of 0.2, which is associated with a very small population of patients. In other words, results from APRX-COLLECTION are insufficient and cover the majority patient's data, while they could lead to a misleading summarization result.

The similar problem is with the RPGlobal algorithm. Again, the long rules are still preferred and rules with more items are selected. However, RPGlobal also considers removing redundancy by encouraging rules that cover other population. As a result, we can see that the data coverage from RPGlobal is slightly better than that of APRX-COLLECTION, with the average support degrees of 0.1171 (RPGlobal) and 0.1154 (APRX-COLLECTION), respectively. Consequently, the major problem with traditional summarization technologies is that the selected rules are associated with a small data coverage, thereby reducing their generalization capability.

On the other hand, as mentioned before, our summarized rule set (illustrated in Table 6) shows a nice balance between the support degree (data coverage) and the number of covered items. For instance, our algorithm leads to a total number of 7 distinct items, which are more than those of baseline (i.e., five). Therefore, more details or complex rules are allowed to be selected using our method. In addition, compared to traditional summarization methods, the proposed approach leads to a summarized rule set with approximately 98% support degree, which outperforms its peers that are with less than 25% support degree. In other words, our method is able to cover the majority of patient cases. Overall, the comparison results clearly show the summarization applicability of our method to represent an overlarge rule set by identifying important rules (with high support degree in terms of data coverage and less redundancy in terms of item overlapping).

At last, we investigate the computational cost from different approaches, and the comparison results are shown in Figure 5. From the experimental results, we notice that the proposed algorithm requires an affordable time for the rule summarization. For example, compared to the RPGlobal method, the proposed algorithm needs 535.8 seconds, which is much better than that of RPGlobal (with 7422.18 seconds). Although we notice that the APRX-COLLECTION approach comes with the least time of 10.23 seconds, its summarization result is the worst among three cases. As such, the satisfactory performance from our proposed algorithm compensate for its computational cost. More importantly, the computational time is accumulated with five clusters in our approach. Note that we can perform the summarization within individual clusters in a parallel way, which could reduce the total cost further. We will leave this work for our future study.

5. Conclusion

In this paper, we propose a novel rule summarization algorithm for identifying informative rules from a cancer-relevant data repository. Three phases are introduced that are capable of generating a comprehensive rule set and relevant rule-similarity graph, performing the community detection to cluster rules, and then selecting important rules to produce a fluent rule summary.

The proposed method is evaluated using the breast cancer dataset from the Surveillance, Epidemiology, and End Results (SEER) resource, which include 85,189 patient samples and 12 variables. The data leads to a completed rule set with over 11,887 rules. By applying the proposed method, we manage to identify the informative rules with high support degree in terms of data coverage and less redundancy in terms of item overlapping. Experimental results also demonstrate that the proposed method leads to competitive performance compared to existing approaches, in terms of the satisfactory summarization results and affordable computational cost. Overall, the proposed method offers a flexible capability and efficient applicability for processing a large amount of medical data that in turn can be utilized to facilitate patients' risk management.

Table 10 summaries the item sets and related medical information from rules within Tables 6–9, respectively.

Data Availability

The Surveillance, Epidemiology, and End Results (SEER) data used in our manuscript to support the findings have been deposited in the publicly available, open-source data repository, which is accessible from https://seer.cancer.gov/.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (Grant no. 61873004) and the Humanities and Social Sciences Foundation of Anhui Department of Education, China (Grant no. SK2017A0098).

References

[1] Y. Bédard, P. Gosselin, S. Rivest et al., "Integrating gis components with knowledge discovery technology for environmental health decision support," *International Journal of Medical Informatics*, vol. 70, no. 1, pp. 79–94, 2003.

[2] J. B. Liu, Z. Y. Shi, Y. H. Pan, J. Cao, M. Abdel-Aty, and U. Al-Juboori, "Computing the laplacian spectrum of linear octagonal-quadrilateral networks and its applications," *Polycyclic Aromatic Compounds*, pp. 1–12, 2020.

[3] A. Yardimci, "Soft computing in medicine," *Applied Soft Computing*, vol. 9, no. 3, pp. 1029–1043, 2009.

[4] J.-B. Liu, J. Zhao, and Z.-Q. Cai, "On the generalized adjacency, laplacian and signless laplacian spectra of the weighted edge corona networks," *Physica A: Statistical Mechanics and Its Applications*, vol. 540, Article ID 123073, 2020.

[5] C. M. Lynch, B. Abdollahi, J. D. Fuqua et al., "Prediction of lung cancer patient survival via supervised machine learning classification techniques," *International Journal of Medical Informatics*, vol. 108, pp. 1–8, 2017.

[6] H. Lu, H. Wang, and S. W. Yoon, "A dynamic gradient boosting machine using genetic optimizer for practical breast cancer prognosis," *Expert Systems with Applications*, vol. 116, pp. 340–350, 2019.

[7] A. Rios, E. B. Durbin, I. Hands et al., "Cross-registry neural domain adaptation to extract mutational test results from pathology reports," *Journal of Biomedical Informatics*, vol. 97, Article ID 103267, 2019.

[8] J. Yang, J. Ma, and S. K. Howard, "Usage profiling from mobile applications: a case study of online activity for australian primary schools," *Knowledge-Based Systems*, vol. 191, Article ID 105214, 2019.

[9] J. Yang, B. Yecies, and P. Y. Zhong, "Characteristics of Chinese online movie reviews and opinion leadership identification," *International Journal of Human-Computer Interaction*, vol. 36, no. 3, pp. 211–226, 2020.

[10] J. A. Delgado-Osuna, C. García-Martínez, J. Gómez-Barbadillo, and S. Ventura, "Heuristics for interesting class association rule mining a colorectal cancer database," *Information Processing & Management*, vol. 57, no. 3, Article ID 102207, 2020.

[11] M. R. Nalluri, K. Kannan, X.-Z. Gao, and D. S. Roy, "Multiobjective hybrid monarch butterfly optimization for imbalanced disease classification problem," *International Journal of Machine Learning and Cybernetics*, vol. 11, no. 7, pp. 1423–1451, 2019.

[12] D. Nguyen, W. Luo, D. Phung, and S. Venkatesh, "Ltarm: a novel temporal association rule mining method to understand toxicities in a routine cancer treatment," *Knowledge-Based Systems*, vol. 161, pp. 313–328, 2018.

[13] R. Saini, P. Kumar, B. Kaur, P. P. Roy, D. P. Dogra, and K. C. Santosh, "Kinect sensor-based interaction monitoring system using the blstm neural network in healthcare," *International Journal of Machine Learning and Cybernetics*, vol. 10, no. 9, pp. 2529–2540, 2019.

[14] J. Yang and B. Yecies, "Mining Chinese social media UGC: a big-data framework for analyzing Douban movie reviews," *Journal of Big Data*, vol. 3, no. 1, p. 3, 2016.

[15] F. Afrati, A. Gionis, and H. Mannila, "Approximating a collection of frequent sets," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Seattle, WA, USA, pp. 12–19, August 2004.

[16] D. Xin, J. Han, X. Yan, and H. Cheng, "Mining compressed frequent-pattern sets," in *Proceedings of the 31st International Conference on Very Large Data Bases*, VLDB Endowment, Trondheim, Norway, pp. 709–720, August 2005.

[17] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[18] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, "Defining and identifying communities in networks," *Proceedings of the National Academy of Sciences*, vol. 101, no. 9, pp. 2658–2663, 2004.

[19] P. Pons and M. Latapy, "Computing communities in large networks using random walks," in *Computer and Information Sciences. Lecture Notes in Computer Science*, p. 3733, Springer, Berlin, Germany, 2005.

[20] U. N. Raghavan, R. Albert, and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," *Physical Review E, Statistical, Nonlinear, and Soft Matter Physics*, vol. 76, no. 3, Article ID 036106, 2007.

[21] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proceedings of the National Academy of Sciences*, vol. 105, no. 4, pp. 1118–1123, 2008.

[22] N. Shukla, M. Hagenbuchner, K. T. Win, and J. Yang, "Breast cancer data analysis for survivability studies and prediction," *Computer Methods and Programs in Biomedicine*, vol. 155, pp. 199–208, 2018.