EASR

Engineering and Applied Science Research

https://www.tci-thaijo.org/index.php/easr/index

Published by the Faculty of Engineering, Khon Kaen University, Thailand

Exploiting a knowledge base for intelligent decision tree construction to enhance classification power

Sirichanya Chanmee and Kraisak Kesorn*

Department of Computer Science and Information Technology, Faculty of Science, Naresuan University, Phitsanulok 65000, Thailand

Received 10 November 2021
Revised 28 February 2022
Accepted 7 March 2022

Abstract

Decision Trees are a common approach used for classifying unseen data into defined classes. The Information Gain is usually applied as splitting criteria in the node selection process for constructing the decision tree. However, bias in selecting the multi-variation attributes is a major limitation of using this splitting condition, leading to unsatisfactory classification performance. To deal with this problem, a new decision tree algorithm called "*Knowledge-Based Decision Tree (KDT)*" is proposed which exploits the knowledge in an ontology to assist the decision tree construction. The novelty of the study is that an ontology is applied to determine the attribute importance values using the PageRank algorithm. These values are used to modify the Information Gain to obtain appropriate attributes to be nodes in the decision tree. Four different datasets, Soybean, Heart disease, Dengue fever, and COVID-19 dataset, were employed to evaluate the proposed approach. The experimental results show that the proposed method is superior to the other decision tree algorithms, such as the traditional ID3 and the Mutual Information Decision tree (MIDT), and also performs better than a non-decision tree algorithm, e.g., the k-Nearest Neighbors.

Keywords: Classification, ID3, Information gain, Knowledge-base, Semantic

1. Introduction

Data scientists have tried to find an approach to extracting hidden knowledge from a large dataset, an approach latterly called "data mining" [1] which is a constantly evolving concept and practice. However, data mining has its limitation where the semantics and the relationships between data are disregarded in the data analysis process. Recently, new technology has extended this approach to be "semantic data mining" [2, 3], which exploits the data relationships in the analytical process which is now being proposed to the machine learning research community. Semantic data mining is a data mining approach that utilizes domain knowledge to enhance the analysis performance. The domain knowledge can assist in constraining the search space, disclosing valuable data patterns, and identifying data relationships [4]. Ontologies [5] describe concepts and relationships represented in a hierarchy that is structured on relationships which are now increasingly applied to support semantic data mining tasks. The domain knowledge in an ontology is used to identify the inherent semantic information which supports the data analytic processes, such as data preparation [6-8] and modeling tasks [9, 10].

Among the classification techniques, a decision tree [11] is a commonly used classification algorithm because it presents the classification model using a hierarchical structure, allowing data miners to understand and interpret results more easily. The advantage of the decision tree is that it can handle both categorical and numerical data and performs better than some complicated techniques, such as neural networks, in several applications. A decision tree algorithm is superior to the standard deep learning technique on tabular datasets, where each attribute is meaningful but deficient in multiscale temporal or spatial structures [12].

The decision tree algorithm uses the splitting criteria to partition the dataset of interest into more homogenous subsets and then identifies the best attributes that satisfy those criteria for constructing the decision tree model.

Information Gain is a splitting condition that usually applies to determine the decision tree's appropriate nodes used in several tree-based algorithms. However, a drawback of using Information Gain as the splitting criteria is the multi-value bias problem which means that the algorithm favors choosing attributes with a large value range as a decision tree node while ignoring attributes with smaller distinct values [13]. Thus, if the chosen attribute is not appropriate, the decision pattern may be hard to interpret, and the performance for classifying the unseen data may decrease. To solve this bias problem, many approaches have been proposed, such as using new splitting criteria [14-17] and applying the weighted attribute concept [18, 19] to improve the selection process for decision tree nodes. For the weighted attribute approach, the importance value of each attribute is used to modify the splitting measure to obtain the appropriate attribute for the decision tree. Therefore, there is a chance that a significant attribute with a few distinct values would be selected to be a node in the decision tree, and the unimportant attribute with multiple distinct values would be disregarded. As a result, the multi-values bias problem will be reduced, and the classification performance will be improved.

*Corresponding author. Tel.: +669 3635 3926
Email address: kraisakk@nu.ac.th

doi: 10.14456/easr.2022.53

In this research, we proposed an approach that integrates ontology into the decision tree construction process. The knowledge in an ontology is used as the attribute importance values to modify the Information Gain for selecting the appropriate attributes to use as the decision tree nodes. The collaboration of the domain ontology with the modeling phase of the decision tree, as proposed in this research, is a significant contribution to research in the field to enhance the performance of the traditional decision tree.

2. Literature review

This section surveys some existing research related to decision tree improvement. Firstly, the theory of decision tree construction is presented. Then, the related works on modifying the decision tree's splitting criteria are discussed. Finally, we offer the research on identifying the concept importance values of ontology.

2.1 Decision tree

A decision tree [20] is a widely used classification algorithm presented as a tree structure to indicate the decision and the results. The input values of this approach can be numerical and categorical data. A decision tree consists of the test point called nodes and branches, representing the attribute value used to separate the dataset into small subsets. An internal node refers to a test attribute, and a leaf node refers to the target class/classification result. For decision tree construction, the algorithms repeatedly partition the data into subsets based on the most informative attributes that satisfy the splitting criteria. Splitting terminates if all instances in the subset belong to the same class or the set of candidate attributes used to split the data is empty.

Several decision tree algorithms are commonly used to classify the data, including the Iterative Dichotomiser 3 (ID3), C4.5, and Classification and Regression Tree (CART) [11]. Each algorithm uses different splitting criteria to identify the best attribute for constructing the decision tree. In our study, the ID3 is applied to examine our proposed technique because this algorithm is simple and quickly classifies the data.

Information Gain applied the entropy principle, which is a measurement in information theory applied to determine the impurity of each attribute. The attribute with the maximum Information Gain value is selected to be a node in the decision tree. The entropy and the Information Gain can be computed as shown in (1)-(3) [21].

$$Info(D) = -\sum_{i=1}^{m} p_i log_2(p_i) \tag{1}$$

where *info*(*D*) is the information entropy of the dataset *D*, which contains *m* distinct classes. p_i refers to the probability of the instances belonging to class *i* and is estimated by $\frac{|C_{i,D}|}{|D|}$. The term $|C_{i,D}|$ refers to the number of instances belonging to class *i* and $|D|$ is the total number of instances of the dataset *D*.

$$Info_A(D) = \sum_{j=1}^{v} \frac{D_j}{D} \times Info(D_j) \tag{2}$$

where $info_A(D)$ refers to the required information to classify the instances from the dataset *D* based on the partitioning by attribute *A*, which has V distinct values. The term $|D_j|$ refers to the number of instances for which attribute *A* has the value *j*.

$$Gain(A) = Info(D) - Info_A(D) \tag{3}$$

where *Gain*(*A*) refers to the Information Gain values of the attribute *A*.

When the Information Gain is applied as a splitting condition for constructing the decision tree, the problem of node selection occurs. The attribute with multiple distinct values may be selected as a node for the decision tree, while the attribute with a smaller range of distinct values may be disregarded [13]. This situation results in obtaining a complicated model and unsatisfactory classification performance. Various techniques have been applied to the decision tree construction process to deal with the multi-value bias problem. The next section will present the related works on improving the decision tree algorithm.

2.2 Modifying splitting criteria of decision tree

The splitting criteria are the significant factor that affects the performance and structure of the decision tree. To improve the decision tree performance, the approaches that modify the splitting measure have been proposed, and these approaches are illustrated in this section.

The first approach is to replace the traditional splitting criteria with the new splitting measurement. For example, Z. Wang et al. [15] proposed an algorithm to select the attributes used in the decision tree by focusing on the consistency of the attribute. The attribute with the greatest consistency will be selected as the node of the decision tree. The result achieved in applying their algorithm indicates that the performance of their approach outperforms the traditional ID3 and also avoids the multi-value bias problem. Y. Wang et al. [16] also applied rough set theory to improve the ID3 performance. The Information Gain computation is simplified by using Taylor's formula for reducing computation time. Then the coordination degree in the rough set theory is integrated for overcoming the multi-bias problem. The experimental result indicates that the proposed approach is superior to the traditional ID3 with less running time for building the decision tree and the tree structure. Fang et al. [14] presented new splitting criteria based on the mutual information concept for improving decision tree performance. The correlation between the attribute and the defined class is used to identify the best attribute used as a node in the decision tree, which can improve classification accuracy. The importance values of attributes are also used as the new splitting criteria of the decision tree. For example, Zhou et al. [17] proposed a decision tree algorithm based on feature weight (FWDT) to improve the performance of the traditional decision tree. The feature weights are determined using the ReliefF algorithm, and these values are used as the new splitting condition for the decision tree construction process. Their results indicate that the FWDT outperforms the traditional approach on classification accuracy.

The second approach is applying attribute importance values to modify the classical splitting measurement for dealing with the multi-value bias problem. To illustrate, Iqbal et al. [22] present an algorithm termed Importance Aided Decision Tree (IADT) that uses the feature importance value to improve decision tree performance. The feature importance values are determined based on the expert's opinion or calculated from the dataset. Their experimental results show that the IADT is superior to the traditional decision tree algorithm. Soni & Parwar [18] applied the attribute importance value to modify the Information Gain for improving the decision tree performance. The attribute importance values are determined using the correlation function method. Their results indicate that applying the importance value of an attribute helps to obtain better performance than the traditional decision tree on the classification accuracy and the error rate. Es-Sabery & Hair [19] present an approach that uses the attribute importance value to improve the decision tree algorithm. In their approach, the importance value of each attribute is identified based on the correlation function between the decision attribute and other condition attributes, and then the attribute with the highest values of the new splitting measure will be chosen to be a node in the decision tree. Their experimental results show that the proposed approach helps decrease the number of leaves and obtains better classification accuracy.

Although the importance value of the attribute can improve the performance of the decision tree, the limitation behind the use of the correlation function as the attribute importance value is that these values are determined based on the observed values in the dataset. The correlation may be inaccurate if the dataset is incomplete. As a result, the decision tree performance may be reduced. Moreover, providing the attribute importance values by specialists may be inconsistent based on the varied experience of those experts. To avoid the uncertainty problem, determining the attribute importance values based on the knowledge in the ontology is an alternative approach that can be applied to modify the splitting condition of the decision tree. The related research of identifying the importance value of the concepts in the ontology will be illustrated in the next section.

2.3 Identify the concept of the importance value in an ontology

With the increasing size and complexity of ontologies, an approach called “Ontology Summarization” [23] has been introduced. The ontology summarization is applied to better understand knowledge in the interest domain by distilling the critical information from the ontology and then generating the overview version of each ontology. To identify the key concepts in the ontology, the relationships between the concepts and the structure of the ontology are exploited to determine the importance value of each concept. These values will be ranked, and then an abridged version of the ontology is generated. PageRank [24], a well-known algorithm to evaluate the importance of a web page, is applied for determining the importance values of the concepts in the ontology. Several works have applied the ontology summarization technique to assist the data mining process. For example, Kralj et al. [25] applied the PageRank algorithm to obtain critical knowledge in an ontology. Then, Hedwig [26] developed a semantic data mining algorithm that exploits this summarized knowledge for deriving efficient rules. Kastrati & Imran [27] also applied the PageRank algorithm to identify the importance value of each concept in the ontology for assisting the document classification task. The importance value is aggregated with the concept relevance score, which is the frequency of the concept in the document, to determine the final weight of each concept for the classification process. The results indicate that the proposed approach can improve classification performance.

The ontology summarization technique can determine the importance value of the concepts, which can be applied with the decision tree to obtain better classification performance.

3. Materials and methods

This section introduces the framework of our study, which is presented in Figure 1. The framework consists of four processes, including (1) Data gathering and ontology development, (2) Data preparation, (3) Decision tree model, and (4) Model evaluation. More details of each process are elaborated in the following section.

3.1 Datasets and ontologies

In this research, we utilized four publicly available datasets. The soybean dataset and the heart disease dataset are benchmark datasets from the University of California, Irvine (UCI) data repository [28]. The two other datasets, the dengue fever dataset [29] and the COVID-19 dataset [30] are available in Mendeley Data. These datasets were chosen to examine and test our approach because the related ontology of each dataset is published and available on the Internet. The experts in each domain collect and investigate the relevant knowledge to generate these ontologies.

The soybean dataset classifies 15 diseases based on the observed disease symptoms and cultivation history. This dataset comprises 35 attributes and 683 records. The heart disease dataset contains 14 attributes and 303 records. The COVID-19 dataset, which consists of 11 attributes and 3,128 records, is used to classify the COVID-19 patients based on the demographic data and the symptoms of this disease. Finally, the dengue fever dataset comprises 14 attributes and 1,104 records, and this dataset is used to classify the dengue fever patients.

Four different ontologies are applied to examine our proposed framework. The Protégé [31], an ontology editing tool, is used to implement these ontologies. The process of designing and selecting ontologies consists of the following steps. First, the domain and scope of ontology are defined. Second, the published ontology will be applied if its knowledge relates to the studied area. Third, the concepts and the relationships between concepts will be added to the selected ontology for covering the studied dataset.

For classifying the soybean disease, the soybean ontology [32] is used and the relevant information, such as knowledge of soybean disease symptoms [33], together with the expert rules derived from Michalski’s work [34], were used to construct our soybean disease ontology. The other three ontologies, which are in the medical domain, which include the Heart Failure Ontology [35], the Dengue Fever Ontology [36, 37], and the COVID-19 Ontology [38, 39], were utilized for classifying the related datasets. To cover all aspects in the studied datasets, we added patient demographic data into the medical ontologies. The example concepts in each ontology are presented in Figure 2, and the structure of each ontology is presented in the appendix section.

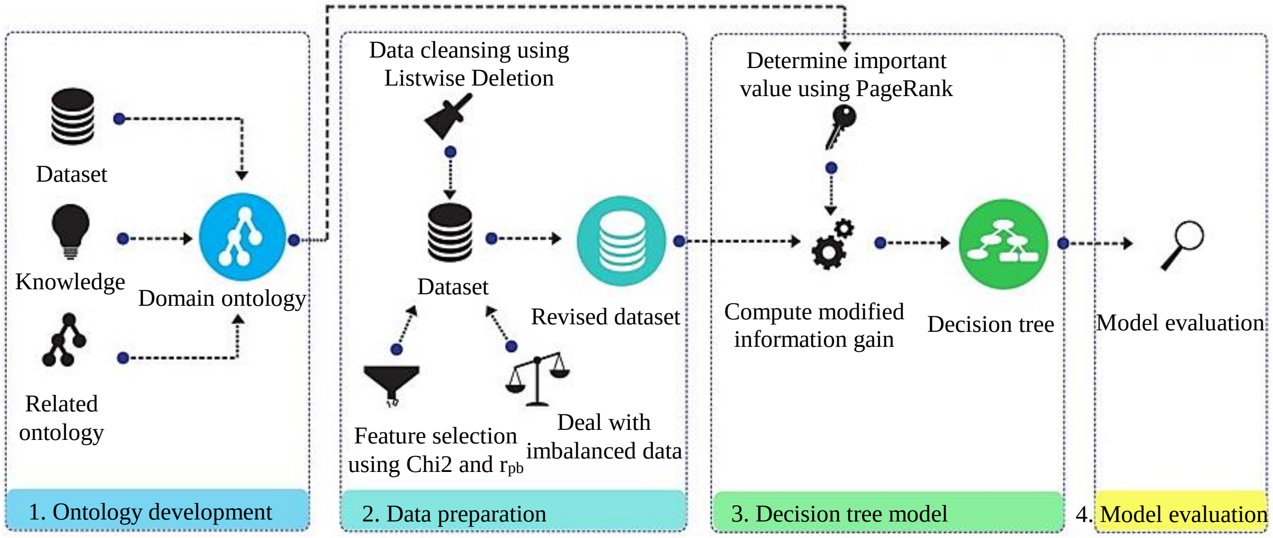


Figure 1 Framework of the knowledge-based decision tree

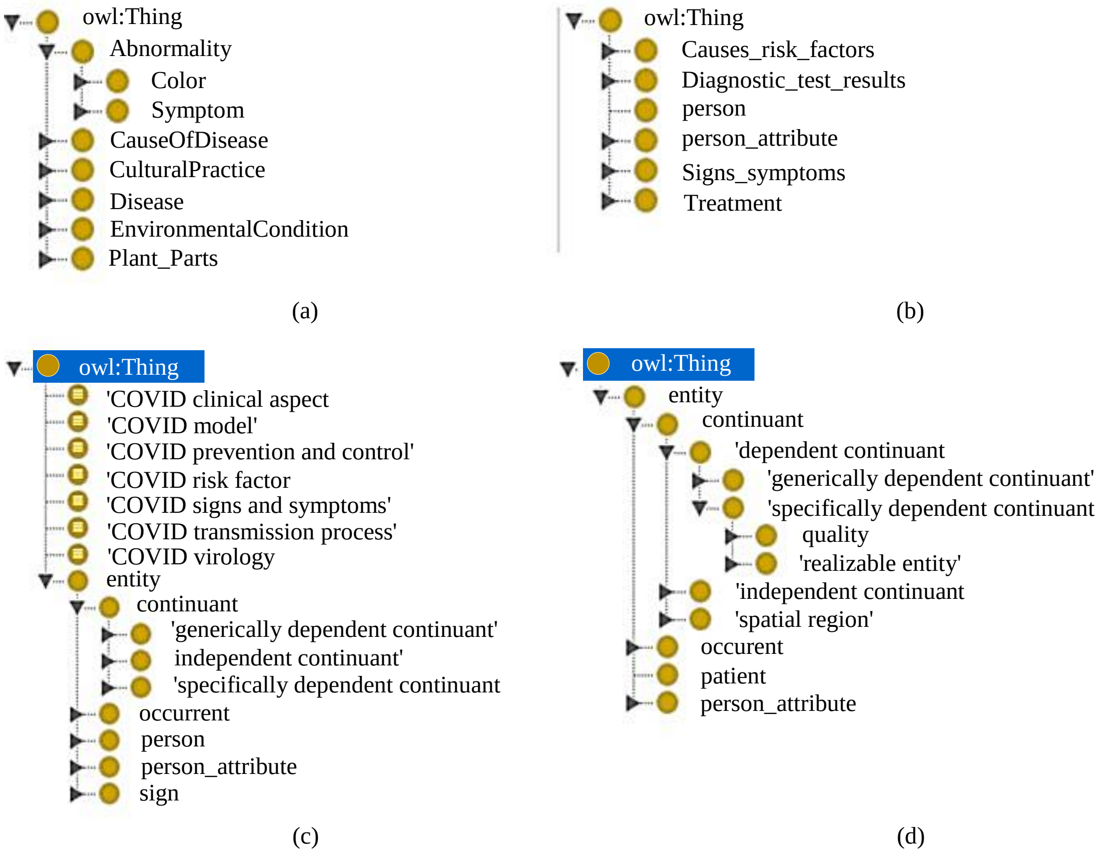


Figure 2 The example concepts in the test ontologies; (a) soybean disease ontology (b) heart disease ontology (c) COVID-19 ontology and (d) dengue fever ontology

3.2 Data exploring and preparation

All datasets are first explored to find errors existing in the datasets. All errors that are found will be resolved in the data preparation phase before the decision tree construction process. Data preparation is a process of transforming the raw data into the form appropriate for the analysis process, and this process also assists in improving data quality. In our study, several tasks were applied to enhance the data quality in the datasets, such as data cleansing, features selection, and handling an imbalanced dataset which refers to the situation where the number of samples per class is unequally distributed.

After pre-processing the data in this way, we found that two datasets, the soybean, and heart disease datasets, contained missing data. The Listwise Deletion Technique [40] is applied to handle those missing data by removing the records with incomplete data. Having cleaned the data, the soybean dataset is reduced to 562 records from the initial 683 records and 297 records remained in the heart disease dataset from the initial 303 records.

Feature selection is the process of reducing the number of attributes used for constructing the model. This process can help avoid the dimensionality problem [41] that arises when analyzing the dataset with a large number of attributes. Analyzing high-dimensional data will affect the efficiency of the model on several aspects, such as obtaining unclear data patterns, achieving unsatisfactory classification results, and an exponential increase in time for the model construction [42]. In our study, the Chi-square statistic (X^2) [43] is applied to identify the correlation between the categorical attributes and the target classes, and the point-biserial correlation (r_{pb}) [44] is used to identify the relationships between numerical features and the defined classes. A *p*-value in either statistic equal to or less than 0.05 indicates that there is a correlation between the test attribute and the target classes. Finally, the unrelated attributes are filtered out from the dataset.

Another problem in the classification task is that of imbalanced data [45], referred to before, where the number of samples per class is unequally distributed. In this case, there is a large number of instances for one class with much fewer instances for other classes. When classifying the imbalanced dataset, the algorithm will learn based on the biased data and usually obtain high classification accuracy. Unfortunately, this result only reflects the accuracy of the majority class, while the performance of the minority classes may be unsatisfactory. To tackle this problem, the over-sampling technique called SMOTE [46] and the under-sampling technique [45] are applied to balance the dataset. SMOTE is used to increase the size of the minority class of the soybean dataset because this dataset is a small dataset of 562 records. In contrast, the under-sampling technique decreases the number of instances of the majority class for balancing the dengue fever dataset.

3.3 Decision tree model

Modeling is the process of generating the data model. ID3 is a standard decision tree algorithm used to classify the data into the defined classes. It is a simple but high-speed algorithm for constructing the decision tree. The ID3 applies Information Gain as the splitting measure to identify the best attributes to be nodes of the decision tree, which causes the multi-value bias problem. This problem affects the classification performance.

To solve the multi-value bias problem, we introduce the "*Knowledge-based Decision Tree (KDT) algorithm*" to classify the dataset. The KDT is developed based on the ID3 algorithm collaborating with the domain ontology. We employ the attribute importance value to adjust the traditional Information Gain to avoid bias on node selection. The ontology model, which presents the relationships between concepts, and the ontology summarizing technique, were applied to determine the attribute importance value in our study. For calculating the attribute importance value, the ontology is transformed into a directed graph model. The vertex represents each concept in the ontology, and the graph edge represents the relationship between concepts. The PageRank algorithm is then used to determine the importance value of each concept which can be defined as in (4)

$$PR(r_i) = d\sum_{j\to i}\frac{1}{N_j}\times PR(r_j) + (1-d) \tag{4}$$

where $PR(r_i)$ refers to the importance value of concept *i*, and $PR(r_j)$ refers to the importance value of concept *j*. The term N_j is the number of outgoing links of the concept *j*, and *d* refers to the damping factor of the PageRank algorithm that is set to 0.85.

The simple idea for calculating the importance value using the PageRank algorithm is illustrated in Figure 3. For a part of the soybean ontology, when the importance value of the *Fruit-Pods* concept and *Fruit Spots* concept are computed, these values will be assigned to the *FruitpodSymptom* concept. Similarly, when the importance values of the *FruitpodSymptom* concept and the *Diseased* concept are calculated, these values will be assigned to the *Symptom* concept for determining the importance value of the *Symptom* concept.

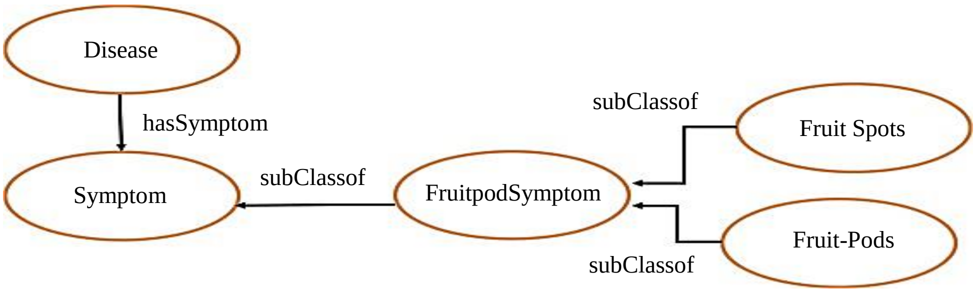


Figure 3 A part of soybean disease ontology graph

For the decision tree construction process, the concept importance value derived from the ontology is used as the importance value of each attribute for modifying the traditional Information Gain. For the ID3 algorithm, the attribute with the maximum Information Gain value is chosen to be a node of the decision tree, so there is a chance that an unimportant feature will be selected, leading to generating complicated classification rules. For the KDT, the importance value is applied to modify the Information Gained value, and the attribute with the highest modified Information Gain is selected as a node of the decision tree. As a result, there is a greater chance that the significant attribute with a low Information Gain value will be selected in the decision tree. The classification rules may be easy to understand, and the performance may increase. The modified Information Gain is defined in (5)

$$MGain(A) = \left(Info(D) - Info_A(D)\right)\times PR(r_A) \tag{5}$$

where *MGain*(*A*) is the modified Information Gain value of attribute *A*, and $PR(r_A)$ is the importance value of attribute *A*.

The differences between our KDT and IADT [22] are the source of knowledge and the method of determining the attribute importance values. For the KDT, the knowledge in the ontology is applied to enhance the performance of the decision tree. This knowledge is used to determine the importance values using the PageRank algorithm. In contrast, experts would provide the knowledge and the feature importance values for the IADT algorithm. Determining the importance values using an ontology is a convenient method since the ontology, which is an explicit knowledge in the domain, is published on the internet and available for reuse.

3.4 Model evaluation

To measure the performance of the KDT, an accuracy value, the standard measurement, is applied. The accuracy is defined as in (6)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \tag{6}$$

where TP (True positive) is the number of positive samples that are assigned into the positive class, TN (True negative) refers to the number of negative samples that are categorized into the negative class, FP (False positive) is the number of negative samples that are classified into the positive class, and FN (False negative) indicates the number of positive samples that are classified into the negative class.

4. Results and discussion

The purpose of this section is to evaluate and validate the KDT algorithm performance. The KDT is utilized to construct the decision tree for each of the soybean, heart disease, dengue fever, and COVID-19 datasets. The experimental results are presented in the following sections.

4.1 Performance evaluation

The objective of this experiment is to evaluate the performance of the classification model when the ontology is used to assist in decision tree construction. Having conducted data cleansing, as explained in Section 3.2, the Chi-square test and the point-biserial correlation were applied to evaluate the relationships between the attribute and the defined classes. A *p*-value of the test attribute higher than 0.05 would indicate that a relationship between this attribute and the defined classes does not exist and it should therefore be removed. After identifying the relationships between data, we removed unrelated attributes, leaving 31 attributes for the soybean dataset, 11 attributes for the heart disease dataset, and 11 attributes for the dengue fever dataset. For the COVID-19 dataset, the statistical results indicate that all attributes correlate with the target classes and, therefore, we kept all attributes of the COVID-19 for the next process.

Table 1 The results of data preparation process

Dataset	Method for identifying data relationship	Number of related attributes	Number of remaining samples	Number of classes
Soybean	Chi-square	31	562	15
Heart disease	Chi-square, point-biserial correlation	11	297	2
COVID-19	Chi-square	10	3,128	2
Dengue fever	Chi-square, point-biserial correlation	11	1,286	2

The final process of data preparation in our study is handling imbalanced data. The SMOTE technique is employed to balance the soybean dataset, and the under-sampling technique is applied for the dengue fever dataset. The results of data preparation are presented in Table 1.

In this experiment, each dataset is randomly separated into training and testing data at a ratio of 70:30 and the experiment is conducted with 30 repetitions. The classification results of the KDT were compared to the traditional ID3 and the MIDT [14] for evaluating the effect of using the knowledge base in the decision tree construction. The classification accuracies of various datasets using the three algorithms, KDT, ID3, and MIDT, are presented in Figure 4.

The classification accuracy of the KDT is greater than the ID3 and the MIDT in all datasets. For the soybean dataset, the ID3 achieved an average accuracy of 87.08%, while the average accuracy of the MIDT is 89.50%. When the KDT is applied, the average accuracy increased to 89.76%. For the heart disease dataset, there are quite different accuracies between each sample test set in all algorithms because this dataset is quite small. Therefore, when we randomly generate the training set, the useful information to allow the algorithm to learn the data pattern in the training set may not be adequate to derive satisfactory classification performance. However, The KDT obtains a better average accuracy at 77.37%, while the average accuracy of the ID3 is 73.78% and the MIDT is 73.59%. For the COVID-19 dataset, the KDT achieved the highest average accuracy at 88.30%, while the ID3 and the MIDT obtained the same average accuracy at 87.68%. For the dengue fever dataset, our approach is slightly better than the MIDT algorithm. The average accuracy increased by 0.21% to 88.82% when the KDT is applied.

As shown in Figure 4, the same pattern is shown in both results of the COVID-19 dataset and the dengue fever dataset. The accuracy of the KDT is similar in approach. As shown in Table 2, which presents the example of the attribute importance value of the studied ontologies, most of the attribute importance values derived from each ontology are similar. For the KDT, the decision tree nodes were selected based on the modified information gain, which is computed using equation (5). When the attribute importance values are similar, the rank of the attributes in the node selection process of the KDT may be identical to the rank of the attributes of the traditional decision tree. Therefore, both algorithms may select the same attribute to use as the nodes of the decision tree, indicating that the classification accuracy of the KDT is similar to the accuracy of the ID3. On the other hand, the KDT performed better than the other algorithms when classifying the soybean dataset. The attribute importance values derived from the soybean disease ontology are different, and these values can help change the rank of attribute for the node selection process. As a result, the appropriate attribute will be selected to construct the decision tree which improved the classification performance.

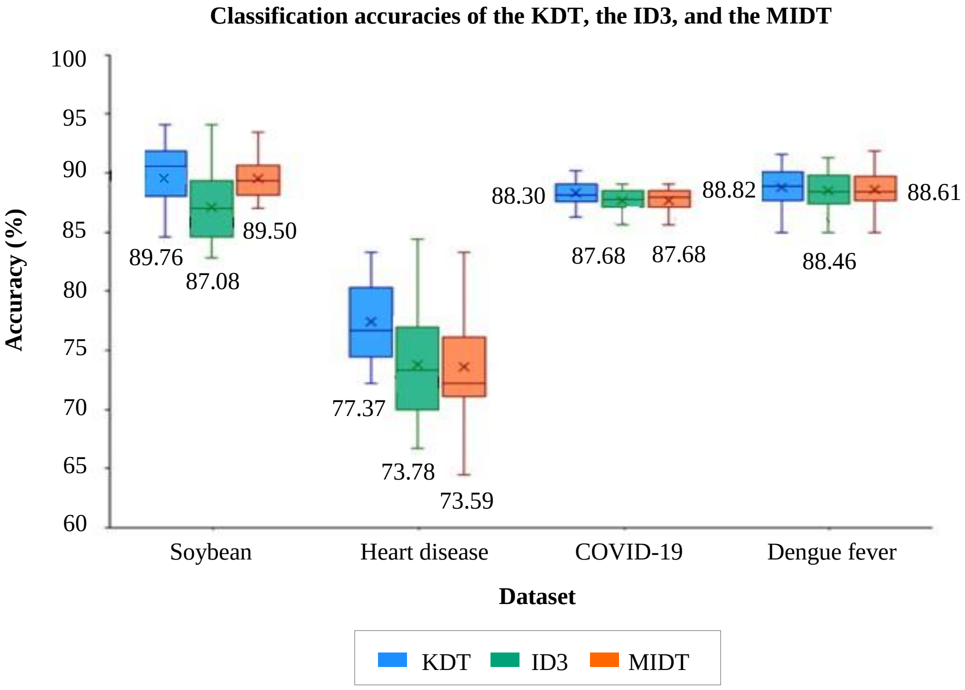


Figure 4 Classification accuracies of the KDT, the ID3, and the MIDT

Table 2 The example of the attribute importance value derived from ontologies

Soybean disease ontology		**Heart disease ontology**		**COVID-19 ontology**		**Dengue fever ontology**	
Attribute (Concept)	**Importance value**	**Attribute (Concept)**	**Importance value**	**Attribute (Concept)**	**Importance value**	**Attribute (Concept)**	**Importance value**
leaves	1.12	Cp	0.38	Olfactory disorders	0.41	fever	0.28
stem	0.98	sex	0.28	dyspnea	0.28	headache	0.21
precip	0.73	Thal	0.28	cough	0.28	gender	0.18
temp	0.68	Trestbps	0.21	gender	0.25	age	0.18
seed	0.61	Age	0.15	sore throat	0.15	rash	0.15
eafspots-halo	0.50	Restecg	0.15	fever	0.15	pruritus	0.15
leafspots-marg	0.49	Thalach	0.15	headache	0.15	myalgia	0.15
leafspot-size	0.49	Exang	0.15	taste disorders	0.15	arthralgia	0.15
seed-size	0.45	Oldpeak	0.15	coryza	0.15	arthritis	0.15
stem-cankers	0.45	Slope	0.15	health professional	0.15	conjunctivitis	0.15
date	0.44	Ca	0.15				
...	…						
shriveling	0.15						
S.D.	***0.25***	***S.D.***	***0.08***	***S.D.***	***0.09***	***S.D.***	***0.04***

Based on these results, we can conclude that using a knowledge-based model in the form of an ontology, to adjust the Information Gain value, can improve the decision tree performance. The structure and relationship between concepts can be used to determine the vital attributes that help to reduce the problem of selecting the unimportant attributes with multi-values as a node in the decision tree construction process, thus overcoming the decrease in classification performance experienced when the insignificant attribute is used.

4.2 Uncertainty of using importance value

We also investigated the effect of ontology complexity on the classification result. The ComplexOnto [47], which is the ontology complexity metric, is applied to determine the complexity score of each ontology. The ComplexOnto is related to various metrics, including link density, link per concept, link richness, and cyclomatic complexity. The complexity of the ontologies used in the experiment is shown in Table 3, which illustrates that the link per concept value of the soybean disease ontology is much higher than other ontologies but its accuracy is not significantly different from the COVID-19 and Dengue fever datasets. Therefore, the complexity of the ontology is not directly correlated with the classification accuracy, as shown in Figure 4.

The more important factor that affects the decision tree performance is the importance value, which is directly related to the link per concept. This is because the link per concept represents the frequency of incoming and outgoing links, which are used by the PageRank to determine the importance value of an attribute. The high value of link per concept means that the concepts in an ontology have several links to other associated concepts. When using the PageRank, each concept will assign its importance value to other connected concepts resulting in almost all concepts obtaining the high importance value accumulated from several connected concepts.

For the KDT, the importance values of the concepts related to the attributes in the dataset are used, as shown in Table 2. The importance values of related attributes in the Heart disease, the COVID-19, and the Dengue fever ontologies are not much different, measured by the standard deviation value (*S.D.*), which were 0.08 for the Heart disease ontology, 0.09 for the COVID-19 ontology, and 0.04 for the Dengue fever ontology. When the importance values are not different, the rank of candidate attributes for the node selection is also similar to the ID3. As a result, the decision rules derived from the decision tree structure generated by the KDT and the ID3 are similar and, thus, the accuracies of both approaches are comparable, which occurred in both of the COVID-19 and Dengue fever datasets, as shown in Figure 4. Therefore, we have concluded that the complexity of an ontology is not directly related to the classification performance of a decision tree. However, the link per concept is more important to the attribute importance values because they are used to find more appropriate information gain, the variable MGain in equation (5), than the traditional decision tree.

Table 3 The complexity of various ontologies

Ontology	Link density	Link per concept	Link richness	Cyclomatic complexity	ComplexOnto score
Soybean disease	305×10^{-6}	1604×10^{-4}	1349×10^{-4}	22.00	5.57
Heart disease	0.7×10^{-6}	6×10^{-4}	5×10^{-4}	413.00	103.25
COVID-19	4.6×10^{-6}	53×10^{-4}	46×10^{-4}	355.00	88.75
Dengue fever	2×10^{-6}	50×10^{-4}	42×10^{-4}	888.00	222.00

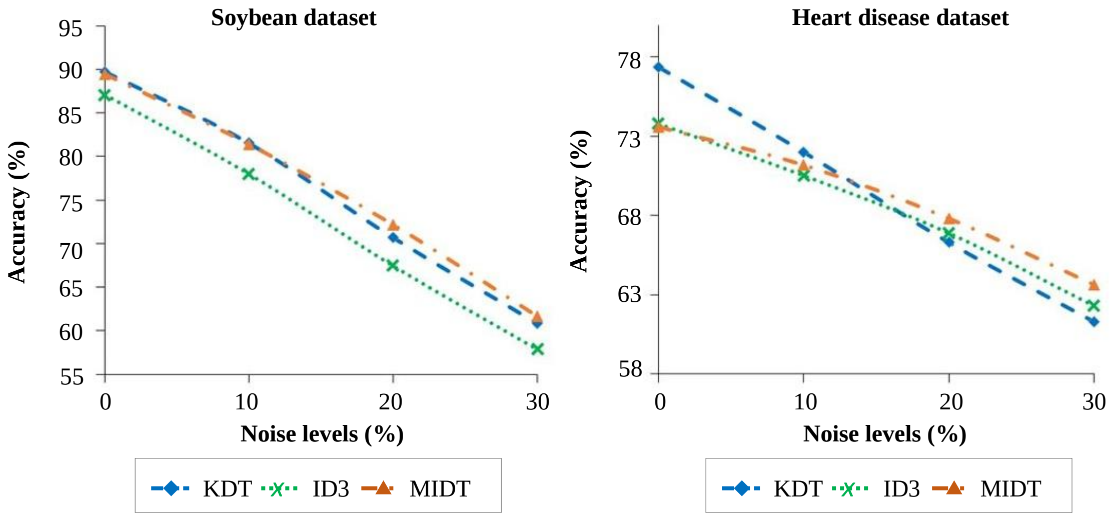
Table 4 Number of decision rules derived from the decision tree of the KDT

Ontology	*S.D.* of Information Gain values of top 3 attributes	Number of identical rules with rules of ID3	Number of new rules discovered by the KDT	Total rules
Heart Disease	0.01	11	26	37
COVID-19	0.05	136	20	156
Dengue fever	0.13	60	18	78

We found that the Heart disease ontology, which has a low link per concept value and low *S.D.* of the importance value, obtained higher classification accuracy than the classical decision tree. This is because the information gain value of the attribute which achieves the first order in the attribute ranking is not different from those of other attributes in the same dataset. Therefore, when adjusting the Information Gain by multiplying it by an importance value ($PR(r_A)$), as shown in equation (5), there is a chance that another attribute can become the first order in the attribute ranking. As a result, the KDT algorithm can generate many new decisions rules, as shown in Table 4. For example, the KDT algorithm can discover 26 new rules that are different from the ID3. These new rules can represent a new decision tree structure, which better classifies data than the ID3. This is because the attribute importance value can assist the node selection process to obtain the more appropriate split nodes for the decision tree, resulting in enhanced classification performance as shown in Figure 2.

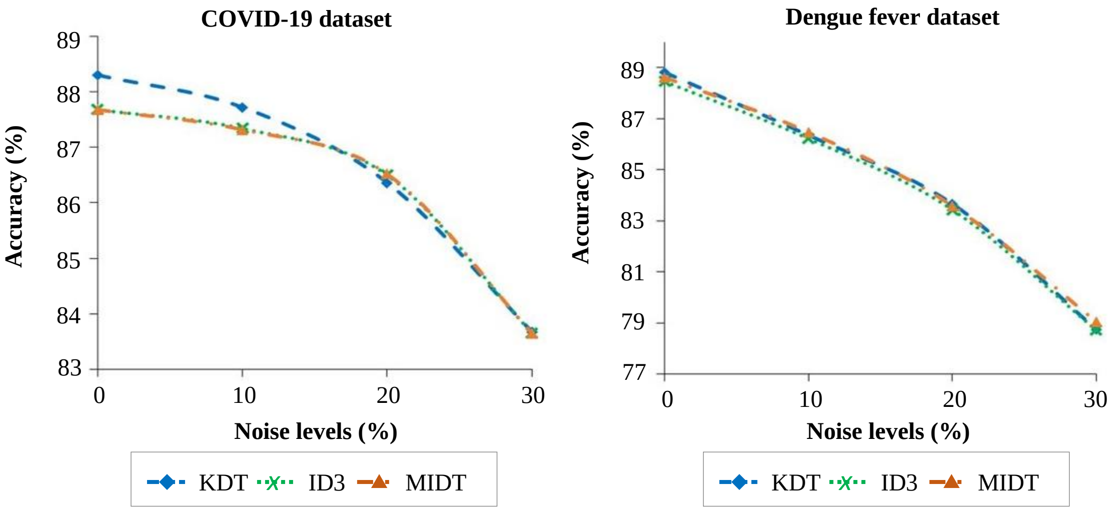
4.3 Effect of noise on knowledge-based decision tree

The purpose of this experiment was to examine the impact of noise on the KDT. We investigated the performance of our proposed approach on analyzing the dataset with different levels of noise in the training data. In our study, the class noise/label noise [48], which refers to the class of each instance being assigned incorrectly, was generated in the following manner. For the datasets with binary classes, the heart disease, the dengue fever, and the COVID-19 dataset, we switched a positive class to a negative class and a negative class to a positive class. For the soybean dataset, we replaced one class with another class that had a similar data distribution, to obtain the new training set with the same class distribution as the original training data. The percentage of data noise versus correct data used in this experiment is 0%, 10%, 20%, and 30%. Figure 5 presents the effect of noise on the accuracy of the three algorithms; the KDT, the ID3, and the MIDT.



(a) (b)

Figure 5 The effect of noises on the different algorithms



(c) (d)

Figure 5 (continued) The effect of noises on the different algorithms

The experimental results showed that classification accuracy is affected by different amounts of noise in the datasets. When the percentage of noise data increases, the classification accuracy decreases. When using a dataset with noise for constructing the classification model, the algorithm will learn the fluctuation patterns and generate the model that fits those patterns. As a result, the classification performance will not be satisfied when this model is then applied to the unseen data. As shown in Figure 5(a) and Figure 5(d), there is the same pattern in the accuracies of the soybean dataset and the dengue fever dataset. The accuracies of the KDT and the MIDT continuously decreased when more noise is added to the datasets. For the heart disease dataset (Figure 5-b) and the COVID-19 dataset (Figure 5-c), the accuracies of the KDT dramatically decreased when adding 10% noise. However, these accuracies are still higher than other algorithms. The MIDT and the ID3 slightly outperformed the KDT when the data contain a noise value of about 20%.

As shown in Figure 5, the KDT is more sensitive to noise data than the ID3 and MIDT algorithms. The accuracy of the KDT decreased faster and is overtaken by the MIDT when the noise data is more than 20%. Since the Information Gain is computed, based on the observed data in the dataset, the obtained Information Gain may be inaccurate when more noise exists. The KDT used the attribute importance values to adjust the Information Gain, and thus, these importance values may not be good enough to obtain the better splitting value for generating the proper ranking of the attribute for the node selection process. Therefore, the accuracy of the KDT is insufficient when analyzing noisy data. However, the performance of KDT on analyzing noise-free data is superior to the ID3 and the MIDT. Therefore, the data preparation phase is necessary for the KDT. If the noisy data is appropriately handled, the KDT can achieve superior classification to the state-of-the-art techniques.

4.4 Comparison of knowledge-based decision tree and other classification algorithms

We compared our proposed approach to other well-known classification methods like k-Nearest Neighbor (*k*-NN) and Support Vector Machine (SVM) [49]. The classification algorithms were applied to the same datasets as in our previous experiments to evaluate their performance. The grid search technique [50], a method used to identify the optimal parameters of the algorithm to obtain the best performance, was applied to determine the best value of each related parameter for classification algorithms. To determine the proper depth for obtaining the best performance with KDT, the maximum tree depth was varied between 1 to the full depth of the decision tree. For achieving the best result from the *k*-NN, the parameter *k* was varied from 1 to 30. The two parameters of the SVM, kernel, and C, were also optimized; the set of kernel parameters used in this experiment consisted of '*linear kernel*,' '*polynomial kernel (poly)*,' '*radial basis function (RBF)*,' and '*sigmoid kernel*,' and the set of parameter *C* was {0.1,1, 10, and 100}. The optimal parameters of each algorithm are presented in Table 5.

These optimal parameters were set for each algorithm to classify the studied datasets. The classification results are presented in Figure 6, which illustrates that our approach is superior to the *k*-NN algorithm in all datasets. To compare our approach to the SVM algorithm, the accuracies of the KDT were better than the SVM for the soybean datasets and the dengue fever dataset. Contrarily, the SVM achieved higher accuracy than our approach on classifying the heart disease dataset and the COVID-19 dataset. For the soybean dataset, the KDT achieved the highest accuracy at 90.69%. The accuracy of our approach is slightly better than the SVM; the accuracy improved by 0.88% to 90.18% when analyzing the dengue fever dataset with the KDT. For the heart disease dataset, the SVM achieved the highest accuracy at 84.7% and 89.03% for the COVID-19 dataset. This result indicates that the SVM performs well in many circumstances, such as analyzing the heart disease dataset, which is a small dataset. This is because the SVM classifies data using the support vector, which is the data point falling on either side of a hyperplane. The position and orientation of the hyperplane depend on this support-vector. As a result, the size of the training set does not impact the algorithm as long as the dataset consists of the support vector [51]. In contrast, the performance of the decision tree will be insufficient when analyzing a small sample size because of the lack of informative patterns for constructing the decision tree [52].

Table 5 The optimal parameters of various algorithms

Dataset	**Tree depth (KDT)**	***k* (*k*-NN)**	**Kernel (SVM)**	***C* (SVM)**
Soybean	6	2	RBF	10
Heart disease	3	24	Linear	1
COVID-19	9	3	RBF	100
Dengue fever	5	3	Linear	1

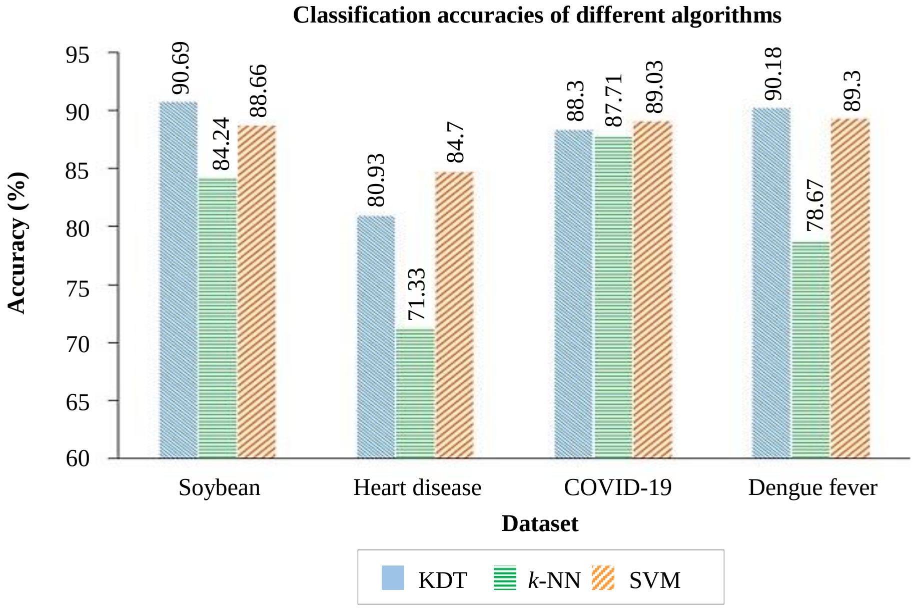


Figure 6 The classification accuracy of the different algorithms

The results of this experiment showed that our approach, the KDT, could outperform the k-NN algorithm. For the soybean dataset, the SMOTE technique is applied to handle the imbalanced data problem. This method may generate noise which results in loss of classification performance [53]. However, the KDT uses the knowledge to assist the node selection process of the decision tree, so the KDT may be less affected by noise than the SVM, which is sensitive to noise and outliers [54]. Therefore, the accuracy of the KDT is superior to the SVM when analyzing the soybean dataset. Moreover, the classification accuracies of the KDT were lower than the SVM when analyzing both the heart disease dataset and the COVID-19 dataset. However, one advantage of the decision tree-based algorithm is that the result is easier to understand than the results of the SVM, which has been proved by Chaitra & Kumar [49]. As such, the performance and the interpretable result could be inverted, and the classification method should be carefully designed to achieve the work's objective.

4.5 Knowledge-based decision tree algorithm

This section provides the process to identify the importance value of each concept in the ontology and the process to construct the decision tree. The method for determining the concept importance value is shown as Algorithm 1, and the process of decision tree construction is shown as Algorithm 2.

For determining the concept importance value, an ontology is transformed into a directed graph model. The concept in the ontology is the vertex, and the relationships between concepts were the graph edge. The ontology graph is loaded to Algorithm 1 to determine the importance value of each concept using the PageRank technique.

For constructing the decision tree, the dataset, a target attribute, and the importance value of each concept were input into Algorithm 2. The process of the KDT is similar to the process of the ID3 algorithm. The modified Information Gain of each attribute is computed, and the attribute with the highest modified information Gain is chosen as the decision tree node. Next, the dataset is split based on the attribute which yields the best modified Information Gain. The process repeats to generate further nodes until the stop condition is met.

Algorithm 1 : Concept importance value identification

```
Input: Ontology graph ($G$)
Output:  Concept importance value (CI)
Initial empty set for concept importance value {$CI$}
d = 0.85
$N$ = number of concepts in $G$
// Initial default importance value of each concept
FOR each vertex $c_i$ where $C_i \in G$
    $PR(c_i) = 1/N$
ENDFOR
// Identify concept importance value
REPEAT
    FOR each vertex $c_i$ where $C_i \in G$
        sum = 0
        FOR each vertex $c_j$ that has outbound link to vertex $c_i$
            sum = sum + ((1/ number of outbound links of $c_j$) × $PR(c_j)$)
        ENDFOR
        $PR(c_i)$ = d × sum + (1-d)
    ENDFOR
UNTIL  importance value $PR(c_i)$ of all concept are not change
Update {$CI$} with all $PR(c_i)$
RETURN {$CI$}
```

For defining the time complexity of Algorithm 1, we determined the worst-case scenario for any input of size n. As shown in Algorithm 1, lines 1 to 3 are the simple statements that perform at once. The first FOR loop (lines 5 to 7) is executed n times, so the total number of times for lines 1 to 7 can be defined as in (7).

$$T_1 = 3 + n \tag{7}$$

As shown in Algorithm 1, the REPEAT loop will execute until the stopping condition is invoked, so we assume that it performs k times. The nested FOR loop, lines 10 to 16, will execute $n \times (n+2)$ times. Lines 18 and 19 are simple statements that perform at once. As a result, lines 9 to 19 will require time to run, as shown in (8)

$$T_2 = k \times \left(n \times (n+2)\right) + 3 = kn^2 + 2kn + 3 \tag{8}$$

Therefore, the total time of Algorithm 1 is shown as (9).

Algorithm 2 : Knowledge-based Decision Tree

```
Input: Dataset (D), Target attribute ($a_{target}$), Attribute importance values (CI)
Output: Decision tree
Initial empty set for decision tree = {}
IF samples in D are all the same class
    Create leaf node that correspond to the most frequency class of D
ENDIF
FOR each attribute $a_i$ where $a_i \in D$ and $a_i \neq a_{target}$
    //Compute modified information gain of attribute $a_i$
    // $PR(a_i) \in CI$
    $MGain(a_i) = (Info(D) - Info_{a_j}(D)) \times PR(a_i)$
ENDFOR
$a_{best}$ = Attribute that obtains the highest $MGain(a_i)$
Tree = Create a node of the decision tree based on attribute $a_{best}$
$D_j$ = Create sub-dataset from D based on attribute $a_{best}$
FOR each attribute $a_j$ where $a_i \in D_j$ and $a_i \neq a_{target}$
    //call recursive algorithm: Knowledge-based Decision Tree
    $Tree_j$ = call algorithm Knowledge-based Decision Tree($D_j$, $a_{target}$, CI)
    Attach $Tree_j$ to the corresponding branch of tree
ENDFOR
RETURN Tree
```

$$T_{algorithm1} = T_1 + T_2 = kn^2 + 2kn + n + 6 \tag{9}$$

Therefore, the complexity of Algorithm 1 is *$O(n^2)$*, where n^2 is the highest order of growth of the function. Since the procedure of KDT is similar to the traditional decision tree algorithm, the time complexity of Algorithm 2 is *$O(mn \log n)$*, which is the complexity of the decision tree algorithm. Finally, the complexity of our algorithm is *$O(n^2)+O(mn \log n)$*. The computation time is a limitation of our approach compared to the traditional decision tree algorithm. When the ontology consists of many concepts and relationships, the algorithm will take a long time to determine the attribute importance values and classify the data. Therefore, the data scientist should consider selecting between classification performance and time complexity for achieving the goal of each task.

5. Conclusion

We propose an approach that utilizes an ontology to improve the decision tree construction process. Since the data quality usually impacts the decision tree performance, the Listwise Deletion technique is used to handle the missing data. Chi-square statistics and the point-biserial correlation are also applied to eliminate unrelated attributes out of the dataset. The SMOTE method and the under-sampling are applied for balancing the studied dataset. This approach improves the dataset's quality, resulting in faster model construction, generating a less complicated model, and achieving better classification accuracy.

The PageRank algorithm is applied to determine the concept importance value from the ontology. These values are used as the importance value of the relevant attribute to modify the traditional Information Gain for avoiding the multi-values bias problem in the node selection process. Therefore, there is more chance that the significant attribute with the low Information Gain value will be selected as a node in the decision tree. Consequently, the performance of the decision tree will improve. However, our approach's main limitation is the quality of the ontology. When the knowledge in the ontology does not cover all essential aspects in the domain, the importance value derived from this ontology may be inaccurate. These values may lead to constructing a complicated decision tree.

For our future work, we will focus on improving the method of identifying the concept importance value by considering the type of relationship between concepts. In addition, further study of the pruning process will be undertaken to investigate the removal of insignificant sections of the decision tree that result in the overfitting problem.

6. Acknowledgements

This research is supported by Thailand Science Research and Innovation (TSRI) (Fundamental Fund Grant No. R2565B063), the Health Systems Research Institute (Grant No. 64-156), the Program Management Unit for Human Resources & Institutional Development, Research and Innovation-CU (Grant number B16F630071), and the Thailand Science Research Innovation (TSRI)-CU (Grant No. FRB640001). The manuscript is edited for correct English grammar and expression by Mr. Roy I. Morien of the Naresuan

University Graduate School. The funder had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. All authors have read and approved the final manuscript and declare that no competing interests exist.

7. References

[1] Hand DJ. Principles of data mining. Drug Saf. 2007;30(7):621-2.

[2] Dou D, Wang H, Liu H. Semantic data mining: a survey of ontology-based approaches. The 9th International Conference on Semantic Computing; 2015 Feb 7-9; Anaheim, USA. New York: IEEE; 2015. p. 244-51.

[3] Sirichanya C, Kraisak K. Semantic data mining in the information age: a systematic review. Int J Intell Syst. 2021;36(8):3880-916.

[4] Anand SS, Bell DA, Hughes JG. The role of domain knowledge in data mining. The 4th International Conference on Information and Knowledge Management; 1995 Nov 29 - Dec 2; Baltimore, USA. New York: Association for Computing Machinery; 1995. p. 37-43.

[5] Staab S, Studer R. Handbook on ontologies. 2nd ed. Heidelberg: Springer; 2009.

[6] Bytyçi E, Ahmedi L, Lisi FA. Enrichment of association rules through exploitation of ontology properties-healthcare case study. Procedia Comput Sci. 2017;113:360-7.

[7] Salguero AG, Espinilla M. Ontology-based feature generation to improve accuracy of activity recognition in smart environments. Comput Electr Eng. 2018;68:1-13.

[8] Chanmee S, Kesorn K. Data quality enhancement for decision tree algorithm using knowledge-based model. Curr Appl Sci Technol. 2020;20(2):259-77.

[9] Paul AK, Shill PC. Incorporating gene ontology into fuzzy relational clustering of microarray gene expression data. Biosystems. 2018;163:1-10.

[10] Alkahtani M, Choudhary A, De A, Harding JA. A decision support system based on ontology and data mining to improve design using warranty data. Comput Ind Eng. 2019;128:1027-39.

[11] Maimon OZ, Rokach L. Data mining with decision trees: theory and applications. 2nd ed. Singapore: World Scientific; 2014.

[12] Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2:56-67.

[13] White AP, Liu WZ. Technical note: bias in information-based measures in decision tree induction. Mach Learn. 1994;15(3):321-9.

[14] Fang L, Jiang H, Cui S. An improved decision tree algorithm based on mutual information. The 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery; 2017 Jul 29-31; Guilin, China. New York: IEEE; 2017. p. 1615-20.

[15] Wang Z, Liu Y, Liu L. A new way to choose splitting attribute in ID3 algorithm. IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference; 2017 Dec 15-17; Chengdu, China. New York: IEEE; 2017. p. 659-63.

[16] Wang Y, Li Y, Song Y, Rong X, Zhang S. Improvement of ID3 algorithm based on simplified information entropy and coordination degree. Algorithms. 2017;10(4):124.

[17] Zhou H, Zhang J, Zhou Y, Guo X, Ma Y. A feature selection algorithm of decision tree based on feature weight. Expert Syst Appl. 2020;164:113842.

[18] Soni VK, Pawar S. Emotion based social media text classification using optimized improved ID3 classifier. International Conference on Energy, Communication, Data Analytics and Soft Computing; 2017 Aug 1-2; Chennai, India. New York: IEEE; 2017. p. 1500-5.

[19] Es-Sabery F, Hair A. An improved ID3 classification algorithm based on correlation function and weighted attribute. International Conference on Intelligent Systems and Advanced Computing Sciences; 2019 Dec 26-27; Taza, Morocco. New York: IEEE; 2019. p. 1-8.

[20] Dietrich D, Heller B, Yang B. Data science and big data analytics: discovering, analyzing, visualizing and presenting data. Indianapolis: John Wiley & Sons; 2015.

[21] Han J, Pei J, Kamber M. Data mining: concepts and techniques. 3rd ed. Burlington: Elsevier; 2011.

[22] Iqbal MRA, Rahman S, Nabil SI, Chowdhury IUA. Knowledge based decision tree construction with feature importance domain knowledge. The 7th International Conference on Electrical and Computer Engineering; 2012 Dec 20-22; Dhaka, Bangladesh. New York: IEEE; 2012. p. 659-62.

[23] Pouriyeh S, Allahyari M, Liu Q, Cheng G, Arabnia HR, Atzori M, et al. Graph-based methods for ontology summarization: a survey. IEEE First International Conference on Artificial Intelligence and Knowledge Engineering; 2018 Sep 26-28; Laguna Hills, USA. New York: IEEE; 2018. p. 85-92.

[24] Brin S, Page L. Reprint of: the anatomy of a large-scale hypertextual web search engine. Comput Netw. 2012;56(18):3825-33.

[25] Kralj J, Vavpetič A, Dumontier M, Lavrač N. Network ranking assisted semantic data mining. In: Ortuño F, Rojas I, editors. International Conference on Bioinformatics and Biomedical Engineering; 2016 Apr 20-22; Granada, Spain. Cham: Springer; 2016. p. 752-64.

[26] Vavpetič A, Novak PK, Grčar M, Mozetič I, Lavrač N. Semantic data mining of financial news articles. In: Fürnkranz J, Hüllermeier E, Higuchi T, editors. Discovery science; 2013 Oct 6-9; Singapore. Berlin: Springer; 2013. p. 294-307.

[27] Kastrati Z, Imran AS. Performance analysis of machine learning classifiers on improved concept vector space models. Future Gener Comput Syst. 2019;96:552-62.

[28] Dua D, Karra Taniskidou E. UCI Machine learning repository [Internet]. University of California, Irvine, School of Information and Computer Sciences; 2017 [cited 2019 Feb 12]. Available from: https://archive.ics.uci.edu/ml/index.php.

[29] Vianna Cardozo S, Maniero V, Rangel P, Camargo T, Souza M, Forte J, et al. Databases of a clinico-ecological study of a triple epidemic [Internet]. Mendeley Data; 2018 [cited 2021 May 10]. Available from: https://data.mendeley.com/datasets/2drcj8mtbc/1.

[30] Viana dos Santos Santana Í, CM da Silveira, Sobrinho A, Chaves e Silva L, Dias da Silva L, Freire de Souza Santos D, et al. A Brazilian dataset of symptomatic patients for screening the risk of COVID-19 [Internet]. Mendeley Data; 2021 [cited 2021 May 28]. Available from: https://data.mendeley.com/datasets/b7zcgmmwx4/5.

[31] Knublauch H, Fergerson RW, Noy NF, Musen MA. The protégé OWL plugin: an open development environment for semantic web applications. In: McIlraith SA, Plexousakis D, van Harmelen F, editors. The Semantic Web-ISWC; 2004 Nov 7-11; Hiroshima, Japan. Berlin: Springer; 2004. p. 229-43.

[32] Crop ontology curation tool. Soybean ontology [Internet]. 2011 [cited 2018 Aug 24]. Available from: http://www.cropontology.org/ontology/CO_336/Soybean.

[33] Markell S, Malvick D. Soybean disease diagnostic series-publications [Internet]. 2018 [cited 2019 Feb 13]. Available from: https://www.ag.ndsu.edu/publications/crops/soybean-disease-diagnostic-series.

[34] Michalski RS. Learning by being told and learning from examples: an experimental comparison of the two methods of knowledge acquisition in the context of development an expert system for soybean disease diagnosis. Int J Policy Anal Inf Syst. 1980;4(2):125-61.

[35] Wang L. Heart failure ontology. BioPortal [Internet]. 2015 [cite 2021 May 11]. Available from: https://bioportal.bioontology.org/ontologies/HFO.

[36] Mitraka E, Topalis P, Dritsou V, Dialynas E, Louis C. Describing the breakbone fever: IDODEN, an ontology for dengue fever. PLoS Negl Trop Dis. 2015;9(2):e0003479.

[37] Mitraka E. Dengue fever ontology. BioPortal [Internet]. 2014 [cited 2021 Jul 5]. Available from: https://bioportal.bioontology.org/ontologies/IDODEN.

[38] Sargsyan A, Kodamullil AT, Baksi S, Darms J, Madan S, Gebel S, et al. The COVID-19 ontology. Bioinformatics. 2020;36(4):5703-5.

[39] Kodamullil AT. COVID-19 Ontology. BioPortal [Internet]. 2021 [cited 2021 Jul 6]. Available from: https://bioportal.bioontology.org/ontologies/COVID-19.

[40] McCarthy RV, McCarthy MM, Ceccucci W, Halawi L. Know your data-data preparation. In: Applying predictive analytics: finding value in data. Cham: Springer; 2019. p. 27-56.

[41] Debie E, Shafi K. Implications of the curse of dimensionality for supervised learning classifier systems: theoretical and empirical analyses. Pattern Anal Applic. 2019;22(2):519-36.

[42] Shroff KP, Maheta HH. A comparative study of various feature selection techniques in high-dimensional data set to improve classification accuracy. International Conference on Computer Communication and Informatics; 2015 Jan 8-10; Coimbatore, India. New York: IEEE; 2015. p. 1-6.

[43] Verma JP. Non-parametric tests for psychological data. In: Statistics and research methods in psychology with excel. Singapore: Springer; 2019. p. 477-521.

[44] Verma JP. Non-parametric correlations. In: Statistics and research methods in psychology with excel. Singapore: Springer; 2019. p. 523-65.

[45] Kaur H, Pannu HS, Malhi AK. A systematic review on imbalanced data challenges in machine learning: applications and solutions. ACM Comput Surv. 2019;52(4):1-36.

[46] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16(1):321-57.

[47] Kumar S, Baliyan N. Quality evaluation of ontologies. In: Semantic web-based systems: quality assessment models. Singapore: Springer; 2018. p. 19-50.

[48] Gupta S, Gupta A. Dealing with noise problem in machine learning data-sets: a systematic review. Procedia Comput Sci. 2019;161:466-74.

[49] Chaitra PC, Kumar DRS. A review of multi-class classification algorithms. Int J Pure Appl Math. 2018;118(14):17-26.

[50] Shekar BH, Dagnew G. Grid search-based hyperparameter tuning and classification of microarray cancer data. Second International Conference on Advanced Computational and Communication Paradigms; 2019 Feb 25-28; Gangtok, India. New York: IEEE; 2019. p. 1-8.

[51] Althnian A, AlSaeed D, Al-Baity H, Samha A, Dris AB, Alzakari N, et al. Impact of dataset size on classification performance: an empirical evaluation in the medical domain. Appl Sci. 2021;11(2):796.

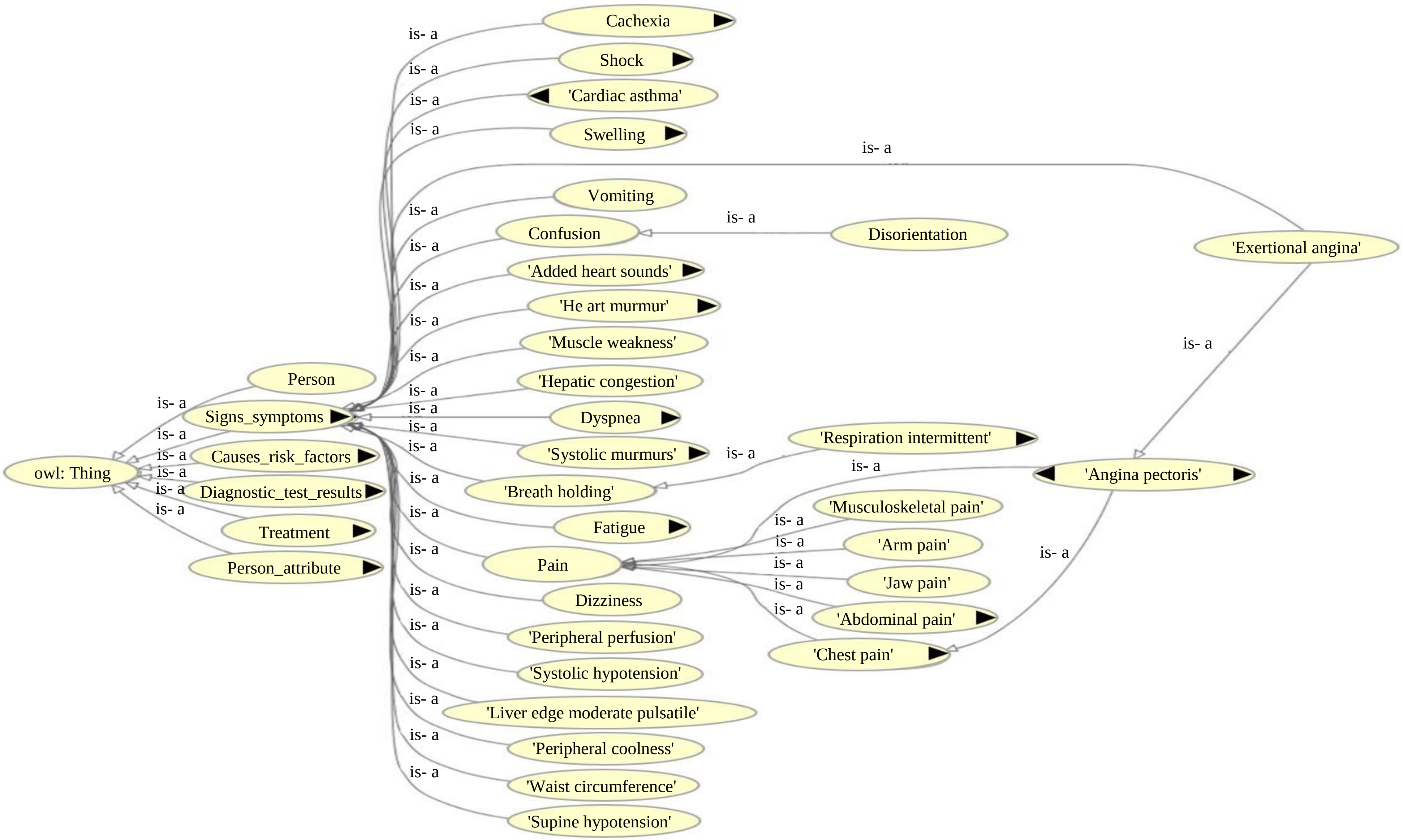
[52] Mehta P, Bukov M, Wang CH, Day AGR, Richardson C, Fisher CK, et al. A high-bias, low-variance introduction to machine learning for physicists. Phys Rep. 2019;810:1-124.

[53] Jiang Z, Pan T, Zhang C, Yang J. A new oversampling method based on the classification contribution degree. Symmetry. 2021;13(2):194.

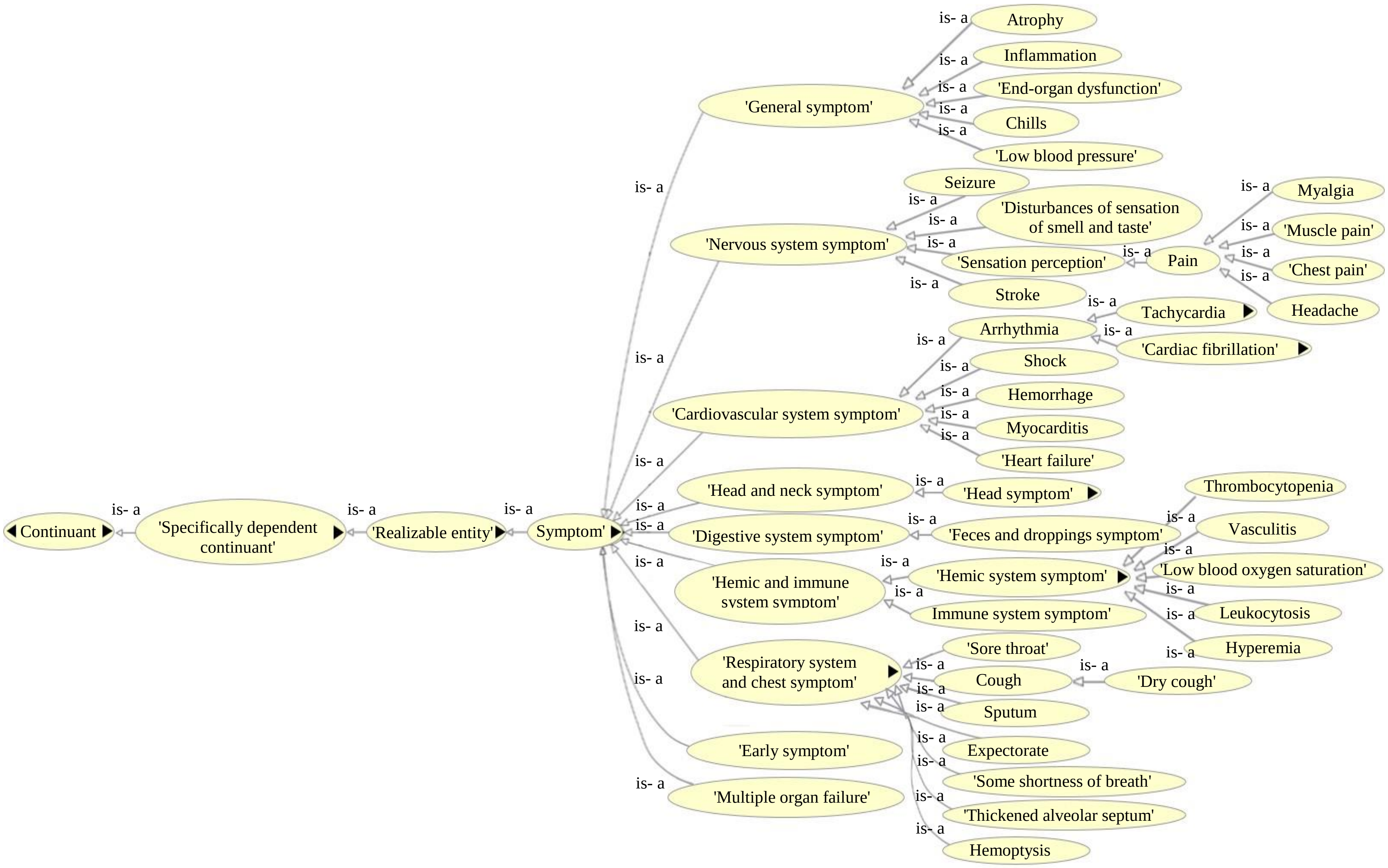
[54] Gaye B, Zhang D, Wulamu A. Improvement of support vector machine algorithm in big data background. Math Probl Eng. 2021;2021:5594899.

8. Appendix

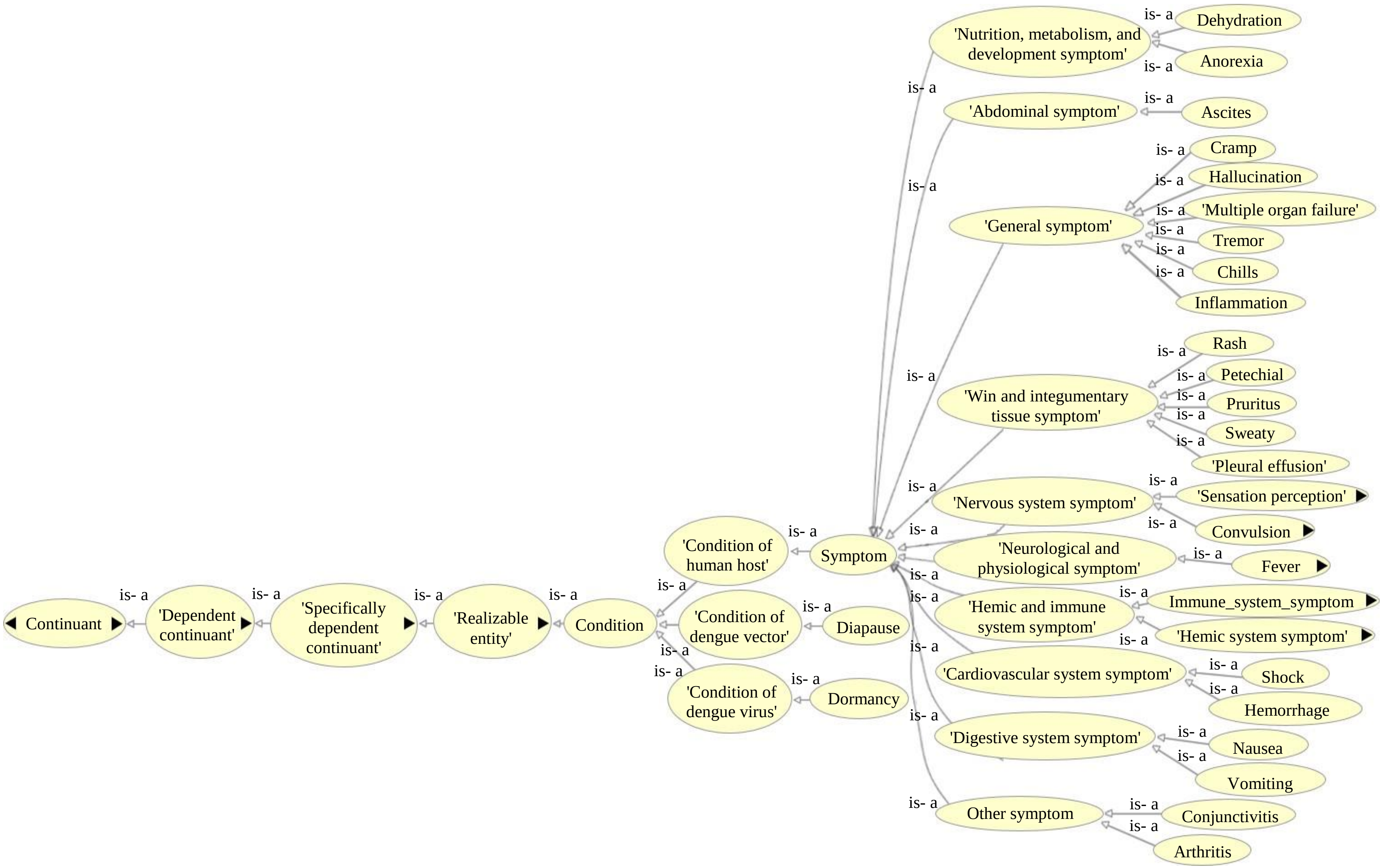
Appendix A: A part of the heart disease ontology



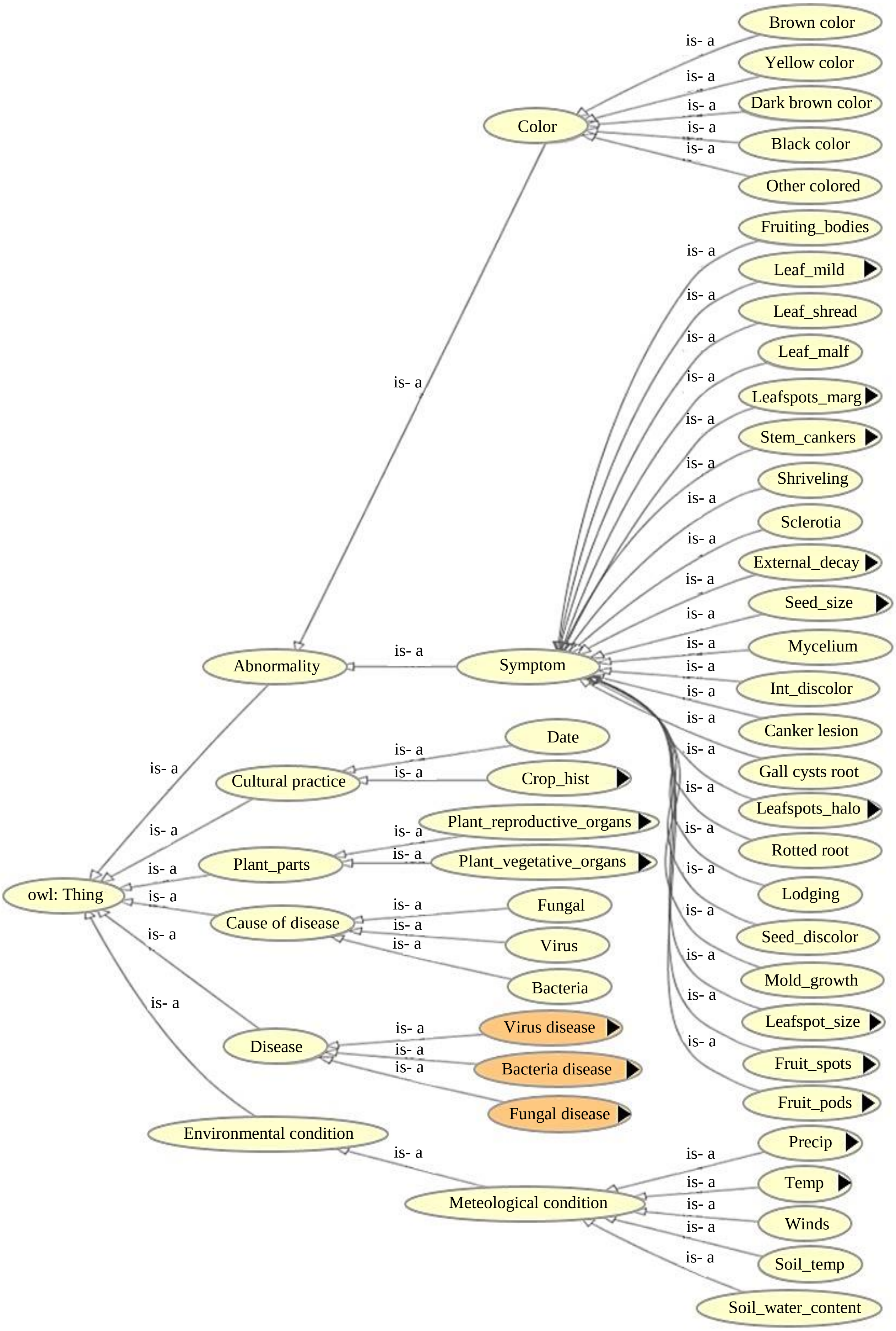
Appendix B: A part of COVID-19 ontology



Appendix C: A part of the dengue fever ontology



Appendix D: A part of the soybean disease ontology


owl: Thing
Abnormality
Cultural practice
Plant_parts
Cause of disease
Disease
Environmental condition
Color
Symptom
Brown color
Yellow color
Dark brown color
Black color
Other colored
Fruiting_bodies
Leaf_mild
Leaf_shread
Leaf_malf
Leafspots_marg
Stem_cankers
Shriveling
Sclerotia
External_decay
Seed_size
Mycelium
Int_discolor
Canker lesion
Gall cysts root
Leafspots_halo
Rotted root
Lodging
Seed_discolor
Mold_growth
Leafspot_size
Fruit_spots
Fruit_pods
Date
Crop_hist
Plant_reproductive_organs
Plant_vegetative_organs
Fungal
Virus
Bacteria
Virus disease
Bacteria disease
Fungal disease
Meteological condition
Precip
Temp
Winds
Soil_temp
Soil_water_content
is- a