Data Quality Enhancement for Decision Tree Algorithm using Knowledge-Based Model

Article *in* Current Journal of Applied Science and Technology · March 2020

DOI: 10.14456/cast.2020.15

Kraisak Kesorn
Naresuan University
35 PUBLICATIONS **178** CITATIONS

Data Quality Enhancement for Decision Tree Algorithm using Knowledge-Based Model

Sirichanya Chanmee and Kraisak Kesorn*

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

Received: 9 March 2020, Revised: 13 March 2020, Accepted: 19 March 2020

Abstract

Data mining is an approach to discovering knowledge or unrevealed patterns from huge data sets by using several methods, such as statistics, machine learning and other data analysis techniques. However, the main limitation of these conventional techniques is that they ignore data relationships and semantics. The data are considered as meaningless numbers with statistical methods being used for model building. For example, the decision tree, a classification method of data mining, is produced from a given set of labeled data, and those data are classified without understanding the semantics of the data or the relationships between attributes. To understand the inherent meaning in the data and to take advantage of the relationships between data elements, we introduce a knowledge-based approach to improve data quality. The proposed approach uses the ontology as the background knowledge to assist the decision tree classification in the process of data preparation. The ontology is used to infer the relationships between attributes and concepts in an ontology. This relationship information can assist the system in identifying related attributes which could assist in the classification process. Two datasets in different domains; agriculture and economics, were used to evaluate the generalization of the proposed approach. Accuracy was the standard measure of success, and was tested in the evaluation of the model. The experimental results showed that the proposed approach can efficiently enhance the performance of the data classification process.

Keywords: data analytics, data mining, ontology, semantic, classification, decision tree
DOI……………………………….

1. Introduction

Data mining [1] is an analytic approach that applies various conventional techniques such as statistical and machine learning to discover hidden knowledge within a set of data. Such datasets can be enormous in this age of Big Data. Ignoring the semantics inherent but undiscovered in the data and the inability to identify relationships between data elements is the limitation of this approach. Previously, for the purpose of building models, data was considered only as numerical values. This was, and is, a limitation when data can, in fact, be categorical or other. To overcome this limitation, an approach called *Semantic Data Mining* has been proposed. Semantic Data Mining [2] refers to an approach in which domain knowledge is incorporated into the data mining tasks to assist in the analysis process.

*Corresponding author: Tel.: +66 81 555 7499
E-mail: kraisakk@nu.ac.th

The domain knowledge can be used to constrain the search space and to reveal more visible patterns in the data [3], and also to identify data relationships. To illustrate, Kuo *et al*. [4] applied medical knowledge in an ontology to categorize 85 attributes relevant to cardiovascular disease, into seven groups for identifying the association rules governing cardiovascular disease resulting in death. Their experimental results found that the use of domain knowledge could help to reveal meaningful rules.

An Ontology [5] is an explicit specification of a shared conceptualization. The knowledge in an ontology is presented as a hierarchical structure of entities and relationships between them. The basis of ontology is a generalization/ specialization of concepts. For example, when focusing on aspects related to plant pathology, the terms such as fungal disease, powdery mildew, and downy mildew might be relevant concepts where the first is a general concept of the latter two. For semantic data mining, an ontology is used to assist several tasks of data mining such as association rules [6, 7], classification [8, 9], and clustering [10, 11].

Classification is a common task in data mining for categorizing prior data into the defined class. The class of each test case is considered on the observed patterns of each training data. The performance of classification depends on the quality of data to be classified, such as the number of missing values, the number of irrelevant attributes, and the size of data. For dealing with the size of data, the notion of data abstraction that denotes the general concept of each value can be applied to obtain the smaller and more general data. Also, the use of abstract data can be obtained the more meaningful data, for instance, the grade point average (GPA) attribute which is the numerical values can be generalized by the higher-level concepts for categorizing these values into several levels such as excellent, good and fair which help to present the student’s academic performance.

To take advantage of the abstraction, Tang and Fong [12] proposed an approach that used the abstract values of each attribute for building the compact decision tree. The concept hierarchy was used to identify the general concepts related to primitive values. The proposed approach was evaluated by using 8,000 online auction instances. The experimental results found that the tree was simpler, and the accuracy also improved. An ontology is one approach that has been used to acquire the abstract values of each attribute, and these values are employed to improve the decision tree performance. For example, Zhang *et al*. [13] presented an ontology-based decision tree algorithm that used abstraction at multiple levels for tree induction. A customer purchase database was used to evaluate the proposed approach. The results showed that the used of abstract values could guide the decision tree induction process and help to enhance the performance of the decision tree. Vieira and Antunes [14] proposed an algorithm that the decision tree incorporated with knowledge in an ontology. The existing attributes and the abstract values which inferred from an ontology were used for a tree induction. The results showed that the proposed approach produced a compact decision tree, and accuracy also increased. As mentioned before, the use of abstraction can improve the decision tree’s performance by handling the variety of attribute’s values, and this approach can help to generate a smaller and more accurate decision tree. The ontology was also used to infer the related abstract concepts of each attribute and identify the association between data elements.

In this paper, we present the ontology-based approach for two disparate domains; soybean disease and census data classification. The abstract concepts in the ontology are used for the data preparation process to improve the quality of the dataset which is later used in the decision tree algorithm. Our study makes a contribution to existing research by examining whether the use of the abstract concepts in an ontology for the data preparation processes can improve the performance of the classification algorithms.

2. Materials and Methods

The framework of the knowledge-based approach for the data quality enhancement is presented in Figure 1. The details of the materials and the processes of this approach are described as follows.

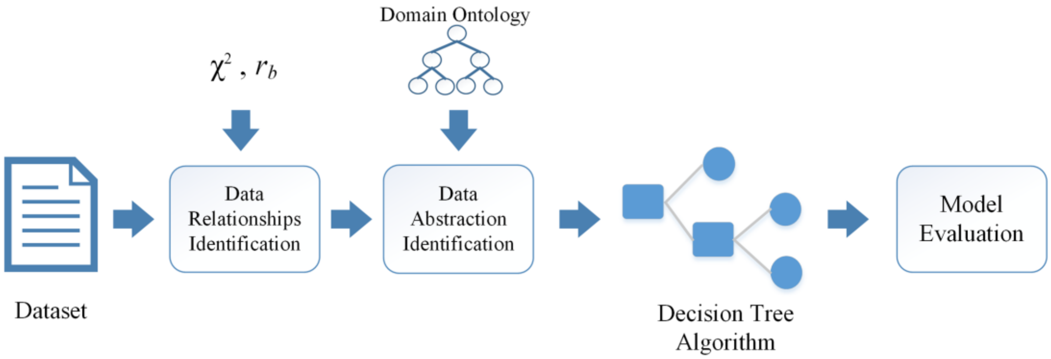


Figure 1. Framework of the data quality enhancement for decision tree algorithm using knowledge-based model.

2.1 Dataset Gathering and Ontology Designing

The datasets used for the experiment consisted of the soybean dataset and the census bureau dataset. The soybean cultivation and diseased indication dataset [15] was used in our experiment. This dataset included 35 attributes, 683 records and 15 classes of disease. There are 121 records with missing value being eliminated at the pre-processing phase for improving data quality, and only 562 remained which were used to evaluate the proposed approach.

The other dataset used in the experiment was the census bureau dataset [15]. This dataset included 13 attributes, and it were used to classify the household income in the United States into two classes. The records with missing data were eliminated then this dataset would reduce from 32,560 records to 30,161 records.

Protégé [16] was used to design and construct the ontologies used in this experiment. The soybean disease ontology consisted of the concepts related to soybean diseases such as the disease name, the indicators of disease and disease type, and the environmental factors, as shown in Figure 2. The development of the soybean disease ontology adopted some ideas from the relevant existing ontologies such as the Soybean Ontology [17] and the ontology of rice diseases in Thailand [18]. The knowledge of disease symptoms extracted from the soybean disease diagnostic series [19] published by North Dakota State University and the expert-derived rules in Michalski’s research [20]. For the census data processing, the personal information ontology necessary was designed to obtain the related abstract values for the attributes of the census dataset. This ontology was adapted from the ontology called OntoLife [21]. The designed personal information ontology consisted of the sociodemographic of a sampling group such as gender, education, marital status, and hometown

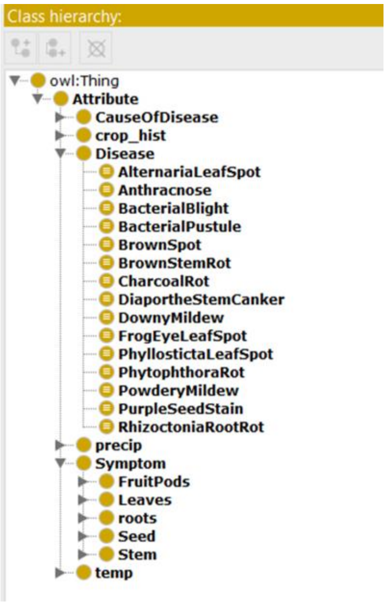


Figure 2. The structure of soybean ontology.

2.2 Data Relationshionships Identification

Data preparation is the process for dealing with the dataset before using that data in the modeling and analysis process. The data preparation process includes data cleansing, data selection, and data transformation. In our research, the Listwise Deletion technique [22] was used for handling the missing values of a dataset by excluding the entire record with any missing variable values. This technique helped to produce the complete dataset for using in the analysis process. After this step, 562 records remained for the soybean dataset and 30,161 records for the census dataset.

For identifying the relationships between attributes and defined classes, biserial correlation (r_b) and Chi-square (χ^2) were used. The biserial correlation is a measure used to estimate the association between a dichotomous nominal variable and an interval variable [23]. The biserial correlation define as in (1) [24].

$$r_b = \frac{r_{pb\sqrt{P_1P_2}}}{y} \tag{1}$$

where r_{pb} is the point-biserial coefficient, P_1 is the fraction of case in the first category, P_2 is the fraction of case in the second category, and y is the ordinates of the normal distribution at the point of P_1 and P_2.

Also, Chi-square is the statistic used to test the independence between variables when those variables are nominal. The formula for calculating the Chi-square value is shown as (2), derived from [25]

$$\chi^2 = \frac{(O-E)^2}{E} \tag{2}$$

where O is the frequency of the observed values and E is the frequency of the expected values.

The null hypothesis (H_0) for the Chi-square independent test is that two categorical variables are not associated. On the other hand, the alternative hypothesis (H_a) state that the two categorical variables are associated. When the p-value of the Chi-square test is less than 0.05, the null hypothesis is rejected, and the alternative hypothesis is accepted.

Thus we can conclude that the relationship between those two variables exists. When the significant results of the Chi-square test were obtained, the Cramer's V was then used as a measure of their relationships. The value of the Cramer’s V is between 0 and 1 without any negative values and the interpretation of these values as shown in Table 1 [26]. After this process, the irrelevant attributes are discarded.

Table 1. Interpretation of Cramer’s V

Cramer’s V	Interpretation
> 0.25	Very strong
> 0.15	Strong
> 0.10	Moderate
> 0.05	Weak
> 0	No or very weak

2.3 Data Abstraction Identification

Data transformation is a process to transform the raw data into the appropriate format, such as the numerical values of temperature would be replaced by the defined categories: ‘normal’, ‘greater than normal’, and ‘lower than normal’. In our study, the designed ontologies were used to identify the association between concepts (so called knowledge in each domain) and values in datasets, and the abstract values derived from the ontologies would replace the primitive values of the datasets in the related attributes for the data transformation process. The use of abstract values allowed the data scientists to view these data more meaningfully.

For example, as shown in Figure 3, the near ground node was the superclass in the soybean ontology, and the below soil node and the above soil node were subclasses. In the 1st process, the values of the stem canker attribute would be mapped with the knowledge in the ontology for identifying the related concepts. In the 2nd process, the related concepts obtained from the ontology would replace the original values of the stem canker attribute, and the revised dataset were used as input of the classification algorithms.

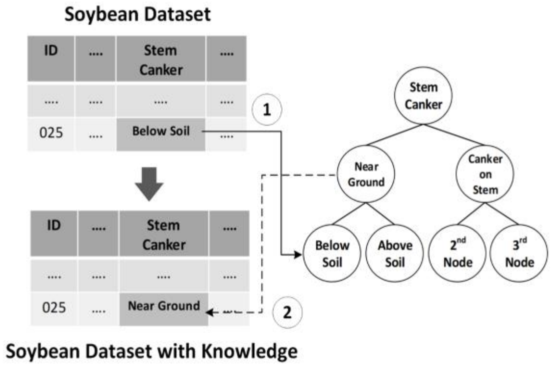


Figure 3. The process of mapping dataset with ontology.

2.4 Classification Process

The classification algorithm used for identifying soybean disease is a decision tree. The decision tree [27] is a classification algorithm that can handle both numerical data and categorical data, and the results are easy to interpret. This algorithm has a tree structure included nodes and branches. Each node of the tree is a decision node, and the leaf nodes are the classes/results of the classification. The decision tree induction algorithms are based on the recursive partitioning approach, and the impurity measures such as entropy and information gain were used as splitting criteria to select the most informative attribute for growing the tree.

The entropy and the information gain are defined as in (3) and (4) [28]

$$Entropy(S) = \sum_{i=1}^{c} - p_i \log_2 p_i \tag{3}$$

where S is an attribute that used to compute the entropy and p_i is the probability of the instances belonging to the i^{th} class.

$$IG(S,A) = Entropy(S) - \sum_{v \in Value(A)} \frac{|Sv|}{|S|} Entropy(S) \tag{4}$$

where A is an attribute in dataset, $|Sv|$ is the number of attribute A's instances which has value v, and $|S|$ is the total number of instances.

Typically, a data scientist prefers the less complex decision tree because the more complex models (larger size of the decision tree) may lead to insufficient performance [29]. The tree complexity is controlled by the stopping criteria and the pruning method. The metrics that are used to measure the tree complexity consist of the total number of nodes, the total number of tree leaves, the depth of the tree, and the number of attributes used. Also, the stopping criteria of the tree growing state include the following condition:

- All the attribute values belong to a single class.
- The tree was growth to the maximum tree's depth.
- The minimum number of instance in the parent nodes is more than the number of instances in the terminal nodes.
- The number of instances in one or child node is less than the minimum number of instances for child node when the node was split.
- The best splitting criteria is less than a threshold [29].

2.5 Evaluation Methodology

The accuracy measure is used to estimate the overall success rate of classification. The accuracy is defined as (5) [30].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \tag{5}$$

where TP is the number of positive instances which are classified as positive, TN is the number of negative instances which are classified as negative, FP is the number of negative instances which are classified as positive, and FN is the number of positive instances which are classified as negative.

3. Results and Discussion

The decision tree algorithm was used to classify both datasets. The results of the classifications using knowledge in the ontology are given in this section.

3.1 Relationships between Attributes

The purpose of this experiment was to examine the effect of the data dimensionality and the relationships between data on the classification. One indicator used to estimate the performance of algorithms is the time complexity of the decision tree algorithm which is estimated according to the number of attributes and the size of training set. For instance, $\mathrm{O}(m \cdot n^2)$ is the time complexity of C4.5 algorithms [31] and $\mathrm{O}(m \cdot n)$ is the time complexity of Su and Jang's algorithm [32], where m denotes the size of the training dataset and n is the number of attributes. Thus, the decrease in the number of attributes used for tree induction can improve the performance of the algorithm.

To test the efficiency of the classification, we prepared the soybean dataset by removing all records with missing data leaving 562 records remaining. Chi-square tests were used to identify the relationships between attributes and soybean diseases by considering the p-values. If the *p-value* of each attribute is greater than 0.05, we concluded that this attribute is not associated with the diseases. Cramer's V value was then used to measure the strength of the association between the attributes and the diseases. The results of the Chi-square test are shown in Table 2. There were four unrelated attributes, indicated by the *bold-italic* letters, including hail, crop-hist, germination, and roots. These attributes were excluded, and then the number of remaining attributes of the soybean dataset was 31.

Table 2. The measurement of the association between attributes using Chi-square and Cramer’s V of soybean dataset

Attribute	χ^2	***p*-value**	**Cramer’s V**	**Association Level**	**Attribute**	χ^2	***p*-value**	**Cramer’s V**	**Association Level**
date	188.18**	0.00	0.334	●	stem	227.02**	0.00	0.636	●
plant-stand	70.76**	0.00	0.355	●	lodging	94.14**	0.00	0.409	●
precip	149.78**	0.00	0.365	●	stem-cankers	594.05**	0.00	0.727	●
temp	176.80**	0.00	0.397	●	canker-lesion	374.16**	0.00	0.471	●
hail	***3.02***	***0.39***	***0.073***	□	fruiting body	148.06**	0.00	0.513	●
crop-hist	***4.39***	***0.88***	***0.051***	□	external decay	89.03**	0.00	0.398	●
area-damage	121.63**	0.00	0.269	●	mycelium	79.15**	0.00	0.375	●
severity	198.04**	0.00	0.42	●	int-discolor	274.74**	0.00	0.699	●
seed-tmt	12.66**	0.04	0.106	◆	sclerotia	118.01**	0.00	0.458	●
germination	***3.85***	***0.70***	***0.059***	□	fruit-pods	595.80**	0.00	0.728	●
plant-growth	256.60**	0.00	0.676	●	fruit-spots	568.06**	0.00	0.58	●
leaves	149.96**	0.00	0.516	●	seed	41.65**	0.00	0.272	●
leafspots-halo	416.37**	0.00	0.609	●	mold-growth	39.98**	0.00	0.267	●
leafspots-marg	415.72**	0.00	0.608	●	seed-discolor	49.24**	0.00	0.296	●
leafspot-size	431.18**	0.00	0.619	●	seed-size	85.72**	0.00	0.391	●
leaf-shread	58.19**	0.00	0.322	●	shriveling	122.89**	0.00	0.468	●
leaf-malf	10.97**	0.02	0.14	◆	***roots***	***4.02***	***0.21***	***0.085***	□
leaf-mild	21.65**	0.00	0.196	○					

Note: ** p-value < 0.05 , ● Very Strong, ○ Strong, ◆ Moderate, □ Weak

The records with missing values in the census bureau dataset were also removed. Then, biserial correlation and Chi-square tests were used to measure the association between each attribute and the defined classes of this dataset. The results of the biserial correlation are shown in Table 3 and the Chi-square tests are shown in Table 4. The result indicated that all attributes correlated with the assigned classes, so all attributes of the census bureau dataset were used for the analysis process.

Table 3. The measurement of the association between attributes using biserial correlation of census bureau dataset

Attribute	r_{pb}	*p*-value	r_b
age	0.24^{**}	0.00	0.33
capital-gain	0.22^{**}	0.00	0.30
capital-loss	0.15^{**}	0.00	0.20
education-num	0.34^{**}	0.00	0.45
hour-per-week	0.23^{**}	0.00	0.31

Note: ** p-value < 0.05

Table 4. The measurement of the association between attributes using Chi-square and Cramer's V of census bureau dataset

Attribute	χ^2	*p*-value	Cramer's V	Association Level
workclass	804.20^{**}	0.00	0.16	○
education	4070.91^{**}	0.00	0.37	●
marital-status	6061.30^{**}	0.00	0.45	●
occupation	3687.30^{**}	0.00	0.35	●
relationship	6233.43^{**}	0.00	0.46	●
race	304.28^{**}	0.00	0.10	◆
sex	1416.52^{**}	0.00	0.22	○
native-country	317.74^{**}	0.00	0.10	◆

Note: ** p-value < 0.05 , ● Very Strong, ○ Strong,◆ Moderate, □ Weak

The results of the classification of soybean diseases that used different numbers of attributes are shown in Table 5. The accuracy of the model that used only some related attributes was higher than the accuracy of the model that used all attributes in the soybean dataset, increasing by 1.78% from 89.94% to 91.72%. The processing time of the classifications also reduced when only related attributes were used. The processing time for analysis of all attributes was 0.0217 seconds, which decreased by 0.0012 seconds to 0.0205 seconds for the analysis of related attributes only.

Table 5. The results of decision tree classification that used different numbers of attributes

Results	All attributes	Only some related attributes
Numbers of attribute	35	31
Accuracy	89.94%	91.72%
Processing time (Seconds)	0.0217	0.0205

The results in Table 5 illustrate that knowing the relationships between data can help to improve the classification performance by identifying the irrelevant attributes which these attributes will be eliminated from the dataset. This lowers the processing time of data mining when analyzing the reduced dataset.

We compared our approach to the Recursive Feature Elimination (RFE) [33] in terms of the elimination of irrelevant attributes. RFE is a well-known wrapper approach for feature selection. Even though Principal Component Analysis (PCA) [34] is a familiar method to reduce the dimension of a dataset, it was not used to compare with our approach because it does not eliminate the attributes in the dataset. On the other hand, PCA will create a new feature set that consists of several inter-correlated input features for dimensionality reduction. The comparative result of our approach and RFE method is shown in Table 6. The results showed that our approach and RFE perform insignificantly different on irrelevant attributes identification and accuracy. The number of related attributes when using the RFE was slightly lower (29 attributes) than that when using our proposed approach (31 attributes). The classification accuracies of both datasets when using the attributes derived from the RFE method were not significantly different (about 0.59% and 0.66% for the soybean and census bureau dataset respectively) compared to the accuracies of our approach. However, the major drawbacks of the wrapper methods are that they required computations to acquire the feature subset, and the classifier used to obtain the feature subset will tend to overfit [35]. To avoid these problems, the use of biserial correlation and Chi-square as a method to identify the unrelated attributes could help to obtain the number of related attributes and accuracy that close to those obtained from the existing feature selection methods.

Table 6. The comparative result of two different feature selection method

Dataset	Our approach		RFE	
	Related attributes	Accuracy	Related attributes	Accuracy
Soybean	31	91.72%	29	92.31%
Census Bureau	13	80.72%	12	81.38%

3.2 Using Knowledge in an Ontology to Assist the Decision Tree Algorithm

The purpose of this experiment was to examine the use of abstract values (inferred data) from an ontology in the classification process of the decision tree algorithm. The data in the datasets are mapped with an ontology for inferring the related concepts/abstraction of each value.

Table 7. The abstract values inferred from the ontologies.

Dataset	Attribute	Primitive Data	Abstract Data
Soybean	stem-canker	below-soil, above-soil	near ground
		above-sec-nde	canker on stem
	fruit-pods	diseased, few-present	presented symptom on fruit pod
	fruit spots	colored, brown-w/blk-specks	colored fruit spots
Census Bureau	age	Numerical values	< 16, 16-19, 20-24, 25-34, 35-44, 45-54, 55-64, > 64
	education	Preschool	Preschool
		1st-4th, 5th-6th	Primary Education

Table 7. (cont.)

	7th-8th, 9th, 10th, 11th, 12th,HS-grad	Secondary Education
	Bachelors, Some-college, Prof-school, Assoc-acdm, Assoc-voc, Masters, Doctorate	Post-Secondary Education
marital-status	Never-married, Widowed, Divorced	Single
	Married-civ-spouse, Separated, Married-spouse-absent, Married-AF-spouse	Married
native-country	Cambodia, India, Japan, China, Iran, Philippines, Vietnam, Laos, Taiwan, Thailand, Hong Kong	Asia
	England, Germany, Greece, Italy, Poland, Portugal, Ireland, France, Hungary, Scotland, Yugoslavia, Holand-Netherlands.	Europe
	United-States, Puerto-Rico, Canada, Outlying-US(Guam-USVI-etc), Cuba, Honduras, Jamaica, Mexico, Dominican-Republic, Haiti, Guatemala, Nicaragua, El-Salvador, Trinadad&Tobago,	North America
	Ecuador, Columbia, Peru	South America

As shown in Table 7, the inferred concepts from the soybean disease ontology were the values of three attributes including stem-canker, fruit-pods, and fruit-spots. Furthermore, the inferred concepts from the personal information ontology were the values of four attributes of the census bureau dataset including age, education, marital-status, and native-country. These inferred concepts will replace the primitive values of each related attribute in the dataset. Next, the datasets with the abstract values are used as the input to the decision tree for identifying the soybean's disease and the class of the U.S income. The accuracy was used to measure the performance of the classification. Also, the depth of the tree was used to estimate the classification efficiency because the depth of the tree is one metric to measure the tree complexity.

Table 8. The classification accuracies and tree's depth.

Dataset	**Primitive Data Classification**		**Abstract Data Classification**	
	Accuracy	**Maximum Depth**	**Accuracy**	**Maximum Depth**
Soybean	91.72 %	14	92.31 %	14
Census Bureau	80.72%	42	82.31%	38

The results of the classification are indicated in Table 8. When the dataset with the abstract values was used, the accuracy of the classification increased from 91.72% on original dataset of the soybean dataset to 92.31% on the revised dataset. The maximum depth of the tree of classification on both the primitive soybean dataset and the soybean dataset with abstract values was 14. The depth of tree is equivalent on both soybean datasets because there were few differences in the attribute values in the original dataset and the revised dataset. To illustrate, as shown in Table 7, the values of the attribute stem-canker were changed from three values (below-soil, above-soil, and above-sec-nde) into two values (near ground, and canker on stem), so the frequency of the observed values that were used as criteria for tree growth of the revised dataset was similar to the original dataset, and it might not be enough to effect the tree's depth. For the classification on the census bureau data, the accuracy of the original census dataset was 80.72% which rose to 82.31% when the related abstract values were used. On the contrary, the maximum depth of the revised dataset was lower than the original dataset because there are more differences in the attribute values of both datasets. As presented in Table 7, the variety of primitive values could be categorized into small groups, for instance, the attribute named native-country showed 41 countries, and these countries could be categorized into four groups: Asia, Europe, North America, and South America. By categorizing the attribute values in this way reduces the variety of attributes values resulting in, for the tree construction process, the tree growing phase terminating faster because of the fewer attributes values to consider, and identifying which of these values belongs to which class [29]. Therefore, the size of the decision tree was smaller when using the dataset with abstract values. The maximum tree depth for the primitive value was 42, while when the census dataset with the abstract values was used, the highest tree's depth reduced by 4 to 38.

As the results have shown, the use of abstract values in an ontology for the decision tree algorithm affects the classification performance. The inference engine of an ontology was used to infer the related concepts of each value, and these related concepts were used to transform the primitive data into the general concepts that help to reduce the wide variety of attribute values. The variety of attribute values affects the depth of the tree because one stopping criterion of the tree growing phase is when it is considered that all attribute values belong to an appropriate class. If the values of the attributes are different, the tree depth is deeper and have a higher time complexity. This would tend to produce poor classification performance.

3.3 Parameter Tuning for the Optimal Results

The purpose of this experiment was to identify the optimal classification results when using the abstract values. The grid search technique [36], which is a method to find the best parameters for the classification algorithm, was used to identify the optimal tree depth for obtaining the best performance of the classification. The important parameter of a decision tree is the maximum depth of the tree. We varied the maximum depth parameter from 1, 2, 3, ...,n where n is the maximum number of tree depth the decision tree algorithm can construct. The results of the classification with parameter tuning are presented in Table 9.

Table 9. The accuracies and tree's depth when applied parameter tuning.

Dataset	Primitive Data Classification		Abstract Data Classification	
	Accuracy	Optimal Depth	Accuracy	Optimal Depth
Soybean	91.72%	12	92.31%	12
Census Bureau	84.80 %	10	84.83 %	8

The value of the optimal tree depth derived from the grid search technique of the original soybean dataset and the soybean dataset with abstract values, was 12, and these parameters were used to determine the optimal performance of the model. The accuracy when using the optimal tree's depth as a parameter in the decision tree algorithm was 91.72% for classifying the original soybean dataset and was 92.31% for analysis on the soybean dataset with abstract values.

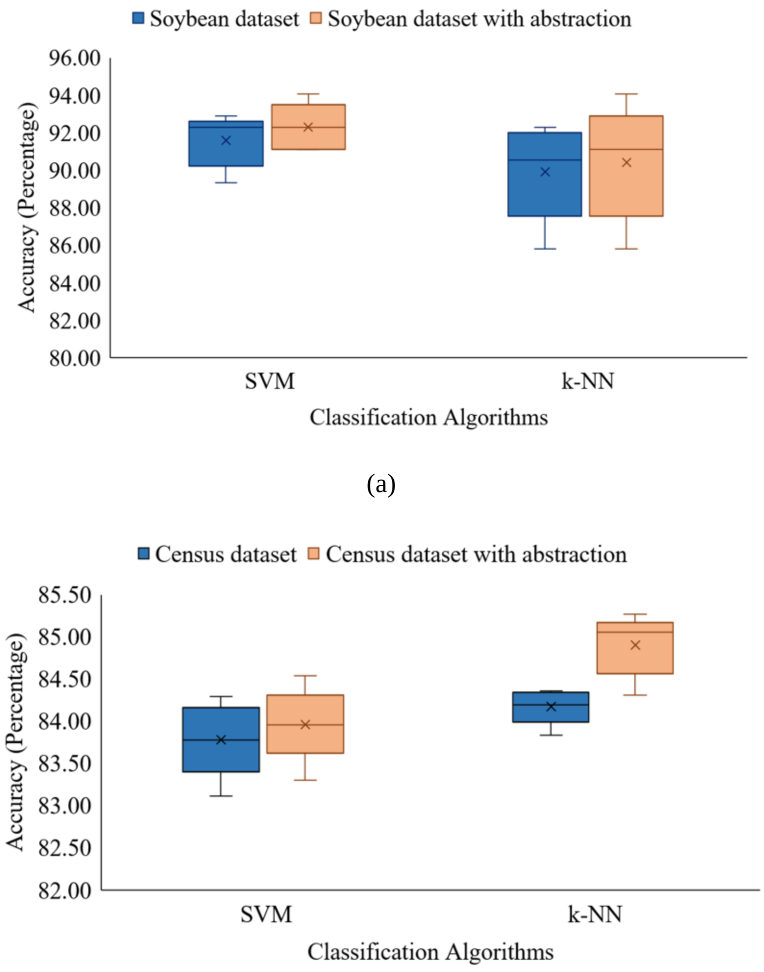
The grid search technique was also used to estimate the optimal depth of the tree to acquire the optimal performance of the household income classification. The grid search parameters were set from 1, 2, 3, ...,n where n is the maximum number of tree depth the decision tree algorithm can construct. The optimal tree depth for analysis of the original census dataset was 10 and for the revised census dataset, 8. The accuracy of the classification with parameter tuning was 84.80% for the classification of the original dataset and 84.83% for the classification of dataset with related abstraction.

The results in this experiment found that the use of abstract values does help to obtain efficient classification performance when parameter tuning was applied. Lowering the depth of the tree when using the abstract values is helpful to data scientists when considering the decision rules

3.4 Using Abstract Values with Various Classification Algorithms

The purpose of this experiment was to examine the use of abstraction values from an ontology in various classification algorithms. Initially, Support Vector Machine (SVM) [27], a supervised learning algorithm that uses the hyperplane, was used to perform the classification. This algorithm requires a small number of samples for training. It provides high accuracy, but the performance depends on parameter selection. In this experiment SVM with a linear kernel was used for the classification process. Also, k-Nearest Neighbor (k-NN) [27] was applied in this experiment. The classification process of the k-NN algorithm is based on the similarity between sample data. The similarity between data is measured by computing the distance between them, and then the data will be grouped into the nearest class. The performance of k-NN is dependent on the size of the data and the number of 'k' which doesn't have a principle of selection. Also, in this experiment, both the soybean dataset and the census dataset were used to evaluate the performance of various algorithms. For the classification of the soybean dataset, the parameter *k* of *k*-NN was set to 2, and the parameter *C* of SVM was set to 1. For the analysis of the census bureau dataset, the parameter *k* of *k*-NN was set to 10, and the parameter *C* of SVM was also set to 1.

The results of the soybean classification process using the different algorithms are presented in Figure 4(a). The accuracy of all algorithms that used the data from an ontology was higher than those used in the original dataset. The SVM's average accuracy increased from 91.60% to 92.31% when analyzing the dataset using the ontology knowledge. Also, the average accuracy of *k*-NN increased from 89.94% to 90.41% when analyzing the dataset with abstraction values.



(a)

(b)

Figure 4. The Accuracy of various classification algorithms that used the primitive dataset and the dataset with abstract values

The results of the household income classification with various classification algorithms are shown as Figure 4(b). The accuracy of the classification of both the primitive census dataset and the revised census dataset also increased. When using the SVM algorithm, the average accuracy of the original census dataset classification was 83.78%, and the average accuracy improved minimally to 83.96% for the classification of the revised census dataset. Furthermore, the average accuracy of the *k*-NN algorithm on the analysis of the primitive census dataset was 84.18%, and the accuracy also climbed marginally to 84.91% when the abstraction data were used.

The results of this experiment showed that the abstract values derived from an ontology could improve the classification results of various classification algorithms [13,14]. For the soybean disease classification, the accuracy of the *k*-NN was lower than the accuracy of SVM. In contrast, the classification accuracy of *k*-NN on the census dataset was greater than the accuracy of the SVM algorithm. Since the performance of the SVM depend on the parameter selection, and the *k*-NN's performance depend on the good value of '*k*' and the size of data [27], so the accuracies of SVM and *k*-NN on both dataset could be in different patterns.

3.5 Algorithm Complexity

In this section, we provide the details of the process to identify the abstract value of each attribute in the dataset using a domain ontology. The algorithm to define the ontology's concept related to the attribute's value is shown as Algorithm 1, and the process of replacing the primitive values with the abstract values is shown as Algorithm 2.

In Algorithm 1, the hierarchical structure of an ontology was utilized to identify the related abstract value of each attribute. The concepts (class) in the ontology were mapped with the attribute's name. If the attribute's name matches the parent class, the attribute's value is used to identify the child class of which this value is a member. Then, this identified class is used as the abstract value for replacing the primitive values in the dataset.

Algorithm 1: Abstract value identification.

```
Input: A list of classes in an ontology ({$C$}), an attribute in dataset ($a_i$), and a value of attribute ($v_{ai}$)
Output: an abstract value
FOR each class $c_i$ where $c_i \in C$
    IF attribute $a_i$ match with parent class of $c_i$
        IF list of instance of $c_i$ ({$I$}) exist
            FOR each instance $ins_i$ where $ins_i \in I$
                IF value of attribute $v_{ai}$ match with instance $ins_i$
                    RETURN $c_i$
                ENDIF
            ENDFOR
        ENDIF
    ENDIF
ENDFOR
```

Algorithm 2 shows that all attributes of a dataset, their values and the classes of an ontology, were loaded to Algorithm 1 for determining the abstract data. The attributes values that could be matched with the concept in the ontology were duplicated. Then, the inferred data obtained from Algorithm 1 were substituted for the primitive value in each duplicate attribute. Finally, this transformed dataset was used as input for various classification algorithms.

Algorithm 2 : Replacing the primitive data

```
     Input : sample dataset (S), an ontology (O)
     Output : a dataset with abstract data
1    Initial the empty set {Ab} for abstract data
2    Find the list of class {C} of ontology
3    // Identify the abstract value of each attribute's value
4    FOR each attribute $a_i$ where $a_i \in S$
5        FOR all unique values $v_{ai}$ of attribute $a_i$
6            InferClass = call Algorithm1({C}, $a_i$, $v_{ai}$)
7            update {Ab} with InferClass
8        ENDFOR
9    ENDFOR
10   //Update the dataset with the abstract values
11   FOR each attribute $a_i$ where $a_i \in S$
12       Count_Infer = count the number of the abstract values of attribute $a_i$
13       IF Count_Infer is greater than zero
14           Duplicate attribute $a_i$ as new attribute a_infer
15       ENDIF
16       FOR each row of dataset
17           Update attribute a_infer with {Ab}
18       ENDFOR
19   ENDFOR
```

For defining the computational complexity of our algorithm, we determine the worst-case execution time of the algorithm for any input of size n. As shown in Algorithm 1, the outer FOR loop will execute 2n times and the inner FOR loop will execute n times. Therefore, Algorithm 1 will require time to run as shown in (6).

$$T_1 = 2n \times n = 2n^2 \tag{6}$$

As shown in Algorithm 2, lines 1 and 2 are the simple statements which executed at once. The outer FOR loop (lines 4 to 9) will execute *n* times. In the inner FOR loop (lines 5 to 8), Algorithm 1 is executed *n* times, so the total number of times that two loops execute are defined as in (7).

$$T_2 = 2 + n \times (n \times T_1) = 2 + 2n^4 \tag{7}$$

For Algorithm 2, lines 10 to 19, the nested FOR loop will require time to execute as shown in (8).

$$T_3 = n \times (2 + n) = 2n + n^2 \tag{8}$$

Therefore, the total time of our algorithm is shown as (9).

$$T_{total} = T_2 + T_3 = 2n^4 + n^2 + 2n + 2 \tag{9}$$

We conclude that the complexity of our algorithm is O(n^4) where n^4 is the highest order of growth of a function. This could be considered as our main limitation compared to other approaches. For example, the classical decision tree has $\mathrm{O}(m \cdot n^2)$ while SVM [37] and *k*-NN [38] have O(n^2). However, this is a trade-off between performance and complexity and the approach should be carefully selected to fit to the work.

4. Conclusions

To enhance the quality of data for the classification task, we proposed an approach that uses an ontology as the background knowledge to assist the data preparation process. For data quality improvement, the records with missing values were eliminated, and biserial correlation, Chi-square and Cramer’s V were used to measure the relationships between attributes and the defined classes. The use of the associated attributes in the classification process improves the classification accuracy by reducing the number of attributes used for the decision tree construction. The reduction of the number of used attributes means the time complexity of the algorithm can be reduced because the number of attributes are one metric that is used to compute the time complexity with Big O notation [32].

After measuring the association between attributes, the soybean ontology was used to infer the related concepts between data. The values in the soybean dataset were substituted by the related concepts from the ontology. The used of abstraction from the ontology improves the performance of the decision tree algorithm by narrowing the variety of attribute values. The consideration of all attribute values into a single class is one stopping criteria of the decision tree growing phase, so if there are quite different numbers of attributes, the model might be complex. Furthermore, when the abstract values are used in the classification process, the depth of the tree is reduced. Since the depth of the tree is one metric to measure the tree complexity, the lowering of the tree’s depth can affect the performance of the decision tree [39, 40]. Also, the notion of abstraction could be applied to the other algorithms such as SVM, *k*-NN because the classification accuracy obtained from each algorithm could improve when the dataset with abstract values were used.

In conclusion, the use of abstract values in the classification task enables better performance and allows the data scientist to view the data in more meaningful ways. However, the level of concepts in then hierarchy used for generalizing primitive data is an important aspect to consider when using this technique. If concepts used are high in the hierarchy, the data may be more general and lead to missing significant information.

There are still opportunities for the researcher to apply an ontology in the data mining process to improve the classification efficiency, such as the use of an ontology to adjust the classification algorithms and to assist the post-processing process. In the future, our work will be continued to improve the performance of the decision tree by using the knowledge in ontology as the criteria to assist the node selection of the decision tree induction process. Also, when the new incoming data are loaded to an existing decision tree, the knowledge in ontology will be used to consider which node of the current decision tree could be modified to obtain better performance.

5. Acknowledgements

This work was supported in part by a grant from Department of Computer Science and Information Technology and Faculty of Science, Naresuan University (Grant no: R2562E027). We also thank Mr. Roy I. Morien of the Naresuan University Graduate School for his assistance in editing the English grammar and expression in this paper.

6. Authors' contributions

K. Kesorn conceived of and designed the study. S. Chanmee and K. Kesorn analyzed the data. S. Chanmee developed the algorithm for Semantic Data Imputation. S. Chanmee and K. Kesorn wrote the manuscript. S. Chanmee submitted and responded to the reviewer's comments. All of the authors read and approved the final manuscript and declared that no competing interests exist.

References

[1] Hand, D. J., 2007. Principles of data mining. *Drug-Safety*, 30(7), 621-622.

[2] Dou, D., Wang, H. and Liu, H., 2015. Semantic data mining: A survey of ontology-based approaches. *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing*. Anaheim, CA, USA, February 7-9, 2015, 244-251.

[3] Anand, S. S., Bell, D. A., and Hughes, J. G., 1995, The Role of Domain Knowledge in Data Mining, *Proceedings of the 4th International Conference on Information and Knowledge Management*, Baltimore, Maryland, USA, November, 1995, 37-43.

[4] Kuo, Y.-T., Lonie, A., Sonenberg, L. and Paizis, K., 2007. Domain ontology driven data mining: A medical case study. *Proceedings of the 2007 International Workshop on Domain Driven Data Mining*, San Jose, California, USA, August 12, 2007, 11-17.

[5] Staab, S. and Studer, R., 2009. *Handbook on Ontologies*. Heidelberg: Springer Science & Business Media.

[6] Marinica, C. and Guillet, F., 2010. Knowledge-based interactive postmining of association rules using ontologies. *IEEE Transactions on Knowledge and Data Engineering*, 22(6), 784-797.

[7] Asadifar, S. and Kahani, M., 2017. Semantic association rule mining: A new approach for stock market prediction. *Proceedings of the 2nd Conference on Swarm Intelligence and Evolutionary Computation*, Kerman, Iran, March 7-9,2017, 106-111.

[8] Benites, F. and Sapozhnikova, E., 2014. Using semantic data mining for classification improvement and knowledge extraction. *Proceedings of the LWA 2014 Workshops*, Aachen, Germany, September 8-10, 2014, 8-10.

[9] Effati, M. and Sadeghi- Niaraki, A., 2015, A Semantic-based classification and regression tree approach for modelling complex spatial rules in motor vehicle crashes domain. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(4), 181-194.

[10] Wang, H., Azuaje, F. and Bodenreider, O., 2005. An ontology-driven clustering method for supporting gene expression analysis. *Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems*, Dublin, Ireland, June 23-24,2005, 389-394.

[11] Trappey, A. J. C., Trappey, C. V., Hsu, F. and Hsiao, D. W., 2009. A fuzzy ontological knowledge document clustering methodology. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(3), 806-814.

[12] Tang, A. and Fong, S., 2010. A taxonomy-based classification model by using abstraction and aggregation. *Proceedings of the 6th International Conference on Advanced Information Management and Service*, Seoul, South Korea, November 30 - December 2, 2010, 448-454.

[13] Zhang, J., Silvescu, A. and Honavar, V., 2002. Ontology-driven induction of decision trees at multiple levels of abstraction. *Proceeding of International Symposium on Abstraction, Reformulation, and Approximation*, Kananaskis, AB, Canada, August 2-4, 2002, 316-323.

[14] Vieira, J. and Antunes, C., 2014. Decision tree learner in the presence of domain knowledge. *Proceedings of Chinese Semantic Web and Web Science Conference*, Wuhan, China, August 8-12, 2014, 42-55.

[15] Dua, D. and Karra Taniskidou, E., 2017. *UCI Machine Learning Repository*. [online] Available at: http://archive.ics.uci.edu/ml

[16] Knublauch, H., Fergerson, R. W., Noy, N. F. and Musen, M. A., 2004. The Protégé OWL Plugin: An open development environment for semantic web applications. *Proceedings of the Semantic Web*, Hiroshima, Japan, November 7-11, 2004, 229-243.

[17] Crop Ontology Curation Tool, 2011. *Soybean Ontology*. [online] Available at: http://www.cropontology.org/ontology/CO_336/Soybean.

[18] Jearanaiwongkul, W., Anutariya, C., and Andres, F., 2018. An ontology-based approach to plant disease identification system. *Proceedings of the 10th International Conference on Advances in Information Technology*, Bangkok, Thailand, December 10-13, 2018, 1-8.

[19] Markell, S. and Malvick, D., 2018. *Soybean Disease Diagnostic Series-Publications*. [online] Available at: https://www.ag.ndsu.edu/publications/crops/soybean-disease-diagnostic-series.

[20] Michalski, R. S., 1980. 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. *International Journal of Policy Analysis and Information Systems*, 4(2), 125-161.

[21] Kargioti, E., Kontopoulos, E. and Bassiliades, N., 2009. OntoLife: An ontology for semantically managing personal information. *Proceedings of Artificial Intelligence Applications and Innovations III*, Thessaloniki, Greece, April 23-25, 2009, 127-133.

[22] Baraldi, A. N. and Enders, C. K., 2010. An introduction to modern missing data analyses. *Journal of School Psychology*, 48(1), 5-37.

[23] Bedrick, E. J., 2005. Biserial Correlation. In *Encyclopedia of Biostatistics.*

[24] Andy, F., 2000. *Discovering Statistics Using Spss for Windows: Advanced Techniques for the Beginner*. CA.: Sage Publications.

[25] McHugh, M. L., 2013. The Chi-Square test of independence. *Biochemia Medica*, 23(2), 143-149.

[26] Akoglu, H., 2018. User's Guide to Correlation Coefficients. *Turkish Journal of Emergency Medicine*, 18(3), 91-93.

[27] Singh, A., Thakur, N. and Sharma, A., 2016. A review of supervised machine learning algorithms. *Proceedings of the 3rd International Conference on Computing for Sustainable Global Development*, New Delhi, India, March 16-18, 2016, 1310-1315.

[28] Cios, K. J., Pedrycz, W., Swiniarski, R. W. and Kurgan, L. A., 2007. *Data Mining: A Knowledge Discovery Approach*. New York: Springer US.

[29] Kotsiantis, S. B., 2013. Decision trees: a recent overview. *Artificial Intelligence Review*, 39(4), 261-283.

[30] EMC Education Services, 2015. *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. Indianapolis: Wiley.

[31] Quinlan, J. R., 1993. *C4.5: Programs for Machine Learning*. San Francisco: Morgan Kaufmann Publishers Inc.

[32] Su, J. and Zhang, H., 2006. A fast decision tree learning algorithm. *Proceedings of the 21st National Conference on Artificial Intelligence*, Boston, Massachusetts, July 16-20, 2006, 500-505.

[33] Guyon, I., Weston, J., Barnhill, S. and Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1), 389-422.

[34] Abdi, H. and Williams, L. J., 2010. Principal component analysis. *WIREs Computational Statistics*, 2(4), 433-459.

[35] Chandrashekar, G. and Sahin, F., 2014. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1), 16-28.

[36] Syarif, I., Prugel-Bennett, A. and Wills, G., 2016. SVM parameter optimization using grid search and genetic algorithm to improve classification performance. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, 14(4), 1502-1509.

[37] Chapelle, O., 2007. Training a support vector machine in the primal. *Neural Computation*, 19(5), 1155-1178.

[38] Cai, Y. and Wang, X., 2011. The analysis and optimization of KNN algorithm space-time efficiency for Chinese text categorization. *Proceedings of Advances in Computer Science, Environment, Ecoinformatics, and Education*, Wuhan, China, August 21-22, 2011, 542-550.

[39] Breiman, L., Friedman, J., Stone, C. J. and Olshen, R., 1984. *Classification and Regression Trees*, Wardsworth, Belmount: Chapman and Hall.

[40] Rokach, L. and Maimon, O., 2005. Top-down induction of decision trees classifiers - a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 35(4), 476-487.