

Article

Not peer-reviewed version

Enhanced Feature Selection via Hierarchical Concept Modeling

[Jirapond Muangprathub](#)^{*}, [Patsita Wetchapram](#), [Apirat Wanichsombat](#), [Arthit Intarasit](#), [Jarunee Sealee](#), [Laor Boongasame](#), [Boonyarit Choopradit](#)

Posted Date: 4 November 2024

doi: 10.20944/preprints202411.0024.v1

Keywords: Formal Concept Analysis; Feature Selection Methods; Hierarchical Concept Model; The Paired-samples T-test, Classification



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Article

Enhanced Feature Selection via Hierarchical Concept Modeling

Jirapond Muangprathub^{1,*}, Patsita Wetchapram¹, Apirat Wanichsombat¹, Arthit Intarasit², Jarunee Sealee², Laor Boongasame³ and Boonyarit Choopradit⁴

¹ Faculty of Science and Industrial Technology, Prince of Songkla University, Surat Thani Campus, Surat Thani 84000, Thailand

² Faculty of Science and Technology, Prince of Songkla University, Pattani Campus, Pattani 94000, Thailand

³ Department of Mathematics, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand

⁴ Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University (Rangsit Campus), Pathumthani, Thailand

* Correspondence: jirapond.m@psu.ac.th; Tel.: +66-887-539-041

Abstract: The objectives of feature selection include simplifying modeling and making the results more understandable, improving data mining efficiency, and providing clean and understandable data preparation. With big data it also allows us to reduce computational time, improve prediction performance, and better understand the data in machine learning or pattern recognition applications. In this study, we present a new feature selection approach based on hierarchical concept models using formal concept analysis (FCA) and decision tree (DT) for selecting a subset of attributes. The presented methods are evaluated based on all learned attributes with 10 datasets from the UCI Machine Learning Repository by using three classification algorithms, namely decision trees, support vector machines (SVM), and artificial neural networks (ANN). The hierarchical concept model is built from a dataset, and it is selected by top-down considering features (attributes) node for each level of structure. Moreover, this study is considered to provide a mathematical feature selection approach with optimization based on paired-samples t-test. To compare the identified models in order to evaluate feature selection effects, the indicators used were information gain (IG) and chi-squared (CS), while both forward selection (FS) and backward elimination (BS) were tested with the datasets to assess whether the presented model was effective in reducing the number of features used. The results show clearly that the proposed models when using DT or using FCA, needed fewer features than the other methods for similar classification performance.

Keywords: formal concept analysis; feature selection methods; hierarchical concept model; the paired-samples t-test; classification

1. Introduction

Nowadays, big data are often confronted, and dataset sizes have increased dramatically in applications of classification by machine learning. Feature selection is typically applied in these classification tasks. The aim of feature selection is to improve the predictive performance of classifiers by removing redundant or irrelevant features. A single feature subset to generate a binary classifier tends to have features that are useful for distinguishing these specific classes but useless for distinguishing others [1–3]. Feature selection is a common task in applications of pattern recognition, data mining, and machine learning since it can help improve prediction quality, reduce computation time, and allow building more human-understandable models. Thus, it is used in many applications before training a classifier [3]. However, while there are a lot of state-of-the-art approaches for feature selection in standard feature space [4], only a few approaches for feature selection in hierarchical feature space have been proposed in the literature [1,5].

Hierarchical model is one approach to feature selection for enhanced predictive performance of classifiers. It is not only advantageous for identifying the hierarchical model itself but is also

helpful for selecting a feature subset for each node [1,5–8]. Classes in a hierarchical structure have both parent-children relationships and sibling relationships [2,9,10]. Moreover, classes with a parent-children relationship are similar to each other and may share common features for classification, while distinguishing between classes with a sibling relationship may require different features [1,8,11]. The authors developed a method for joint feature selection and hierarchical classifier design using genetic algorithms. Thus, this work focuses on using hierarchical structure in feature selection to improve the predictive performance in classification. In this study, we designed a feature selection algorithm based on the hierarchical information structure.

The hierarchical models were built by using two strategies: bottom-up and top-down. The bottom-up strategy starts by placing each object in its own cluster and then merges these atomic clusters into larger and larger clusters until all of the objects are in a single cluster or until some other termination conditions are satisfied. Later, the top-down strategy does the reverse by starting with all objects in one cluster. It subdivides the cluster into smaller and smaller pieces until each object forms a cluster on its own or until some other termination conditions are met, such as a desired number of clusters or that the distance between the two closest clusters is above a certain threshold [12–14]. Thus, this work applied both bottom-up and top-down approaches, which were done with formal concept analysis (FCA) and decision tree, respectively. Both FCA and decision tree approach were considered for feature selection in order to reduce the attribute count while still retaining the efficiency of data classification.

FCA [15–18], invented by Rudolf Wille in 1982, is a method for data analysis based on concept lattice. It is widely used in information science to describe attributes and objects of information that can be represented in a hierarchical structure. FCA provides relationships of generalization and specialization among concepts in the concept lattice. This method is rarely applied to feature selection in classification or other tasks. This research applied the advantage of being general knowledge through calculation and selection a root node of hierarchical structure. The alternative approach by decision trees uses these powerful and popular tools for classification and prediction. Such classifier has a tree structure, where each node is either a leaf node or a branching decision node [11,19]. This approach is widely used to perform feature selection. The advantages of decision trees are intuitive appeal for knowledge expression, simple implementation, and high classification accuracy. Thus, this work applied both FCA and decision trees to reduce the number of features that need to be collected and also to provide better classification accuracy. Due to the different patterns of selecting general knowledge in both FCA and decision trees in hierarchical concept structure, FCA will be constructed a concept lattice structure based on logic and set theory while decision trees will be built a tree structure based on information gain. Moreover, the decision tree in this study is considered to provide a mathematical selection of the features with the optimization level based on paired-samples t-test. Next, the proposed models were tested and evaluated by using three popular algorithms for classification, namely decision tree, Support Vector Machine (SVM), and Artificial Neural Network (ANN), to assess the predictive performances with the presented approach, using data from the UCI repository.

This article is organized as follows. Section 2 briefly describes the machine learning techniques applied in this work. Section 3 presents the research methodology, followed by experimental results and discussion in section 4. Finally, section 5 concludes the article.

2. Related Works

Feature selection methods have been used for a wide variety of applications [1,2]. Such methods proceed to eliminate unfavorable features that might be noisy, redundant, or irrelevant and could penalize the performance of a classifier [2]. Feature selection contributes thus to reduce the dimensionality of data and to restrict the inputs, which can contain missing values in one or several features [1,4]. The various feature selection methods fall into three main categories: filters, wrappers, and embedded methods [2–4,20]. The first category of filter methods selects the features by using weights that indicate correlations between each feature and a class. The largest weights are

selected in rank order. The calculated weights may be based on Information gain [1,3,20], Chi-square [21,22], or other saliency measures. The second category is wrapper methods that evaluate feature subsets by directly using a learning algorithm and selecting features based on their impact on the model's performance [3,20]. While this appears reasonable as a direct approach, it tends to be too expensive computationally. Finally, embedded methods incorporate feature selection within the learning algorithm during the training process [1–3]. In prior studies, many feature selection methods were developed to identify the most relevant and informative features from a dataset. Selecting the right set of features can improve model performance, reduce overfitting, enhance interpretability, and speed up the learning process. This study focuses on applying a hierarchical concept model to select the features from available data.

In recent years, feature selection based on a hierarchical information structure has been proposed because of its rational learning inside the structure. [23] proposed a hierarchical feature selection framework by considering the parent-child relationship, sibling relationship, and family relationship. These relationships were modeled and implemented by using a data matrix concept [5]. [13] presented the hierarchical feature selection with subtree based graph regularization by exploring two-way dependence among different classes. [5] designed a feature selection strategy for hierarchical classification based on fuzzy rough sets to compute the lower and upper approximations of classes organized in a class hierarchy. The authors developed an efficient hierarchical feature selection algorithm based on sibling nodes [5]. [24] proposed feature selection framework based on semantic and structural information of labels for hierarchical feature selection by transforming to semantic regularization and adapted the proposed model to a directed acyclic graph case. [25] proposed a robust hierarchical feature to reduce the adverse effects of data outliers and learn relatively robust and discriminative feature subsets for hierarchical classification. All the above mentioned studies assumed specific relationships between categories for hierarchical regularization. However, most of the existing hierarchical feature selection methods are not robust for dealing with the inevitable data outliers, resulting in a serious inter-level error propagation problem in the classification process that follows. Thus, FCA is applied in this current study because the previous experimental studies have shown that FCA is not sensitive to outliers [7]. FCA provides a well-defined mathematical framework to discover implicit and explicit knowledge in an easily understood format by using formal context and a Hasse diagram that clearly represents the concepts' generalization/specialization relationships. In addition, FCA can construct an informative concept hierarchy providing valuable information on various specific domains. To support this work, [15] also applied FCA method to present a new filter for feature selection, called H-Ratio, which can identify pertinent features from data based on the Shannon entropy, also known as the diversity Shannon index that reflects how many different types there are in a data set to select features without considering classification accuracy. However, the classification accuracy is an important aspect of selection. Thus, this work applied the classification accuracy to formulate for feature selection from concept lattice.

3. The Proposed Models for Feature Selection

The proposed approaches based on hierarchical models focused on two methods, namely using FCA and DT. FCA and DT were applied to build the hierarchical structures for feature selection from data so that each node in the structure represented a feature (attribute). An overview of the proposed models is depicted in Figure 1.

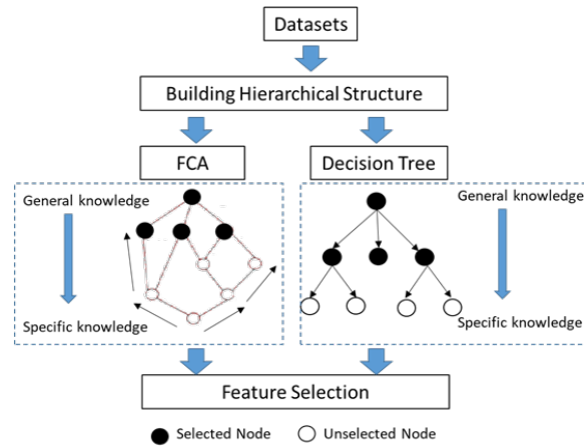


Figure 1. A conceptual overview of the proposed models

Initially, the hierarchical structure form was built, where the top node represents general knowledge, whereas the bottom represents specific knowledge. Next, the top node of structure was chosen by considering each level of structure. In this study, we selected a feature for each level of the structure, and then the selected feature was used to represent data for classification in the next step of evaluating the performance.

Afterwards, the relationship between the number count of features (attributes) and classification accuracy was considered. We found that the number of attributes for each level affects classification accuracy at some level. We attempted to optimize the feature selection from the hierarchical structure. Thus, the paired-samples t-test was applied to choose appropriate features for each level. This approach has advantages, including the estimation of mean differences between paired observations and the identification of statistically significant changes in two related measurements. The paired design helps manage individual differences, minimizing the impact of confounding variables, and optimizes data economy by leveraging the paired nature of observations. Furthermore, it serves as an efficient method for testing differences between two related conditions or measurements, offering a straightforward and practical means for comparing means within the same group. We performed feature selection based on a hierarchical structure as follows.

On level l of hierarchical structure, for any $l \in \{1, 2, \dots, k\}$, we defined the total accuracy as follows.

$$TotalAcc_l = \sum_{i=1}^n \frac{Acc_{l,i}}{n} \quad (1)$$

where $Acc_{l,i} \in \{0, 1\}$ is the i^{th} accuracy on level l of decision tree, and the values 1 and 0 mean that prediction of $Acc_{l,i}$ is true or false, respectively. To find the set of chosen levels, we denoted by $SoCL$ where level 0 is definitely chosen. Next, we consider sequentially the next level of decision tree by using the two-samples z test between total accuracy on level l and $l - 1$ defined as

$$\begin{aligned} Z_{l,l-1} &= \frac{TotalAcc_l - TotalAcc_{l-1}}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n} + \frac{1}{n}\right)}} \\ &= \sqrt{\frac{n}{2\hat{p}\hat{q}}} (TotalAcc_l - TotalAcc_{l-1}) \end{aligned} \quad (2)$$

where $\hat{p} = \frac{\sum_{i=1}^n Acc_{l,i} + \sum_{i=1}^n Acc_{l-1,i}}{n+n}$ and $\hat{q} = 1 - \hat{p}$ for any $l \in \{1, 2, \dots, k\}$. The level l will be added to set $SoCL$ depending on the level of significance α . This means we use α to consider the significant difference between total accuracy of level l and $l - 1$. If total accuracy of level l is significantly different from total accuracy of level $l - 1$, we will add l into the set $SoCL$. Otherwise, we do not add l into the

set $SoCL$ and stop level finding process. The next table shows some values of significance level α and their corresponding confidence intervals for z .

Table 1. The mapping from α to confidence interval for z

α	0.01	0.05	0.1
z	$-2.576 \leq z \leq 2.576$	$-1.96 \leq z \leq 1.96$	$-1.645 \leq z \leq 1.645$

For example, when we consider the level of significance $\alpha=0.01$, if $|Z_{(l,l-1)}| \leq 2.576$, then we add l to the set $SoCL$. Otherwise, l is not added to set $SoCL$, and we stop the process. We denote the set of chosen attributes by $SoCA$. If m is the last level which is added to the set $SoCL$, we have set $SoCL = \{0, 1, 2, \dots, m\}$ and set $SoCA = \bigcup_{i=0}^m Attr_i$ where $Attr_i$ is the set of all attributes on level i . We present the proposed feature selection algorithm using FCA and decision tree, presented as Algorithm 1.

Algorithm 1: Feature selection for FCA and decision tree.

Input : The training dataset comprises all attributes and n instances.

: Tree in decision tree or concept lattice including k level.

: α and z // for example: $\alpha = 0.01, z = [-2.576, 2.576]$

Output : Set of selected attributes ($SoCA$).

Method :

1. $SoCL = \{0\}, l = 0$
 2. $SoCA = \{Attr_l\}$
 3. $TotalAcc[l] = \sum_{i=1}^n \frac{Acc_{l,i}}{n}$ // Accuracy of using selected attributes in level l
 4. For $++l$ to k // $l=1$ to k
 5. $\hat{p} = (\sum_{i=1}^n Acc_{l,i} + \sum_{i=1}^n Acc_{l-1,i}) / (n + n)$
 6. $\hat{q} = 1 - \hat{p}$
 7. $Z_{l,l-1} = \sqrt{\frac{n}{2\hat{p}\hat{q}}} (TotalAcc[l] - TotalAcc[l-1])$
 8. If $(|Z_{(l,l-1)}| \leq 2.576)$
 9. $SoCL = \{0, l\}$
 10. $SoCA = \bigcup_{i=0}^l Attr_i$
 11. Else Break();
 12. End For
 13. Return ($SoCA$)
-

4. Research Methodology

4.1. Dataset Description and Tools

This study took all needed datasets from the UCI machine learning repository [26]. These datasets belong to many different fields, such as engineering, social science, and business. Among them, we selected 11 datasets covering many areas for our evaluation. Table 2 shows the different samples of data used. We give also, respectively, the number of attributes, instances, and classes. Moreover, in the final column, we describe the attribute data type in each data set. These data are used to demonstrate the applicability and performance of feature selection. Afterwards, classification is used to evaluate the accuracy when each dataset is randomly split into training and test sets.

Table 2. Datasets used in this study

Dataset	No. Attributes	No. Instances	No. Classes	Data Type
Dermatology	33	366	6	Integer
Glass	10	214	7	real
Iris	4	150	3	real
Lung-cancer	56	32	3	Integer
Movement	91	360	15	Integer
Pageblocks	10	5,473	5	Integer, real
Segmentation	19	2,310	7	Integer, real
Soybean	35	307	19	Integer
Tunadromd	242	4,465	2	Integer
Wine	13	178	3	Integer, real
Zoo	17	101	7	Nominal, Integer

This study used the R program for data management and analysis, specifically R-3.5.1 for Windows and RStudio Version 1.1.456. They were used to classify with the classifiers of DT, SVM, and ANN types. We also used Weka software to provide the previous feature selection algorithms compared to our models. The feature selection methods tested were filters methods (Information gain and Chi-square), wrappers methods, and forward selection. In addition, in the proposed model using FCA for feature selection, we applied Concept Explorer (ConExp) [27] programs to generate the hierarchical structure for FCA. The ConExp program has the functionality of creating and editing contexts, generating lattice from contexts, finding the relative attribute that actually occurs in the context, and the basic relationship that is true in the context.

4.2. The Experimental Design

We designed experiments for evaluating the performance in classification tasks. The overall scheme is illustrated in Figure 2.

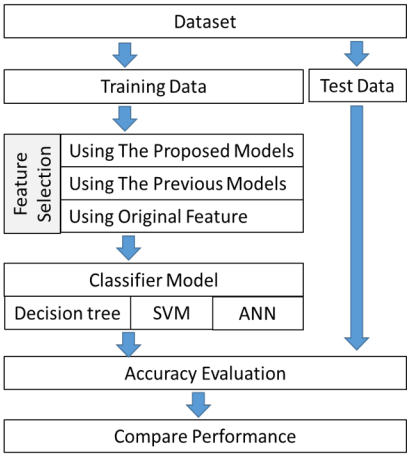


Figure 2. The experimental design

In these numerical experiments, datasets from the UCI Machine Learning Repository (Asuncion, 2007) were employed, each divided into training and testing groups. The training set was processed feature selection using three methods: the proposed models, prior models, and the original features. Initially, the proposed model involves the utilization of FCA and decision tree within a hierarchical concept model. These structures are created, and significant features are selected, as explained in

Section 3. Subsequently, traditional feature selection methods such as IG, CS, FS, and BS are employed to compare the proposed feature selection model and the original features. These conventional methods are widely utilized for feature selection and operate based on filter (e.g., IG and CS) and wrapper (e.g., FS and BS) methods. Finally, the original features are used as a baseline for comparison. Next, the selected feature(s) for each dataset were used to train a classifier, namely DT, SVM, and ANN. These models are popular classifier models. Afterwards, the test data were used to evaluate the classification accuracies of these classifiers. The accuracies served as the performance measures for each model.

Subsequently, the chosen feature(s) for each dataset were employed to train classifiers, specifically DT, SVM, and ANN. These models are well-known in the field of classification. Following the training phase, the test data were utilized to assess the classification accuracies of these classifiers. The accuracies then served as performance metrics for each model.

5. Results and Discussion

5.1. The Original Performance Results

Practically, the datasets in Table 2 were run using 10-fold cross-validation, which is the most used method. For this, each dataset was divided into complementary subsets, performing the analysis on one subset, called the training set, and validating the analysis on the other subset, called the testing set. The results from our experiments for classification accuracy are shown in Table 3.

Table 3. Summary of datasets and classifier performances

Dataset	No. attributes	The classification accuracy (%)		
		Org-DT	Org-SVM	Org-ANN
Dermatology	33	84.99	84.50	85.09
Glass	10	100.0	95.26	99.05
Iris	4	78.67	76.98	77.00
Lung-cancer	56	80.83	80.56	81.00
Movement	91	73.06	72.69	72.88
Pageblocks	10	97.94	96.05	97.88
Segmention	19	89.09	88.63	88.70
Soybean	35	96.00	95.98	96.00
Tunadromd	242	96.64	99.10	99.22
Wine	13	83.17	80.82	82.98
Zoo	17	96.09	95.53	95.72

In this experiment, the input data had not been reduced, instead all original attributes were used for training. In Table 3, the results show the standard classification accuracy for each data set when using DT, SVM, and ANN. Mostly, DT was the superior one. The accuracy results are used as benchmarks for evaluation the feature selection in the next section.

5.2. The Hierarchical Concept Model Performance

This section describes the proposed model for feature selection based on hierarchical concept model by using DT or FCA. Thus, the results will be divided into three parts as follows.

5.2.1. Feature Selection Using Decision Tree

In order to evaluate our models, we used the well-known software RapidMiner Studio to generate a tree structure using the DT algorithm. The datasets in Table 2 were used to create tree structures.

Next, we selected the attributes from each level of the tree structure. In this experiment, we selected attributes by filtering from the first to the fifth level of the tree. Table 4 shows the classification accuracy (%) in each level of the decision tree. We compared classification accuracy with our feature selection using DT to the benchmark with original features (without using feature selection at all) in Section 5.1, as shown in Figure 3.

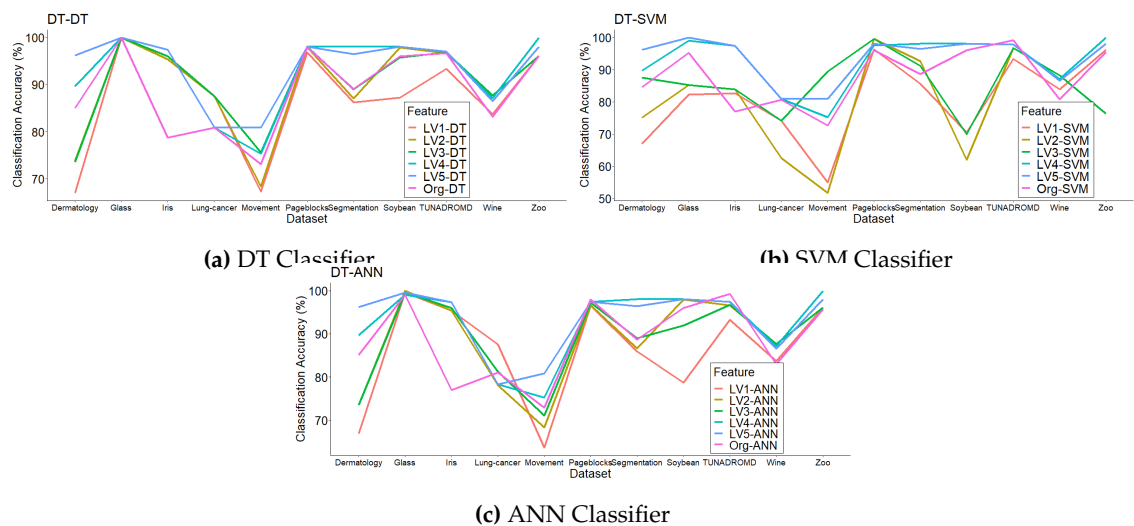


Figure 3. The comparison DT-SVM-ANN classifiers with features selected using DT from the first to the fifth level; and with all original features

From Figure 3, we see that the classification accuracy did not differ between the original and the proposed models for each level. Indeed, if we add the number of attributes from each level, it leads to a trend of improving accuracy.

Table 4. The classification accuracy on using each level of decision tree for feature selection

Dataset	The classification accuracy (%) using decision tree based feature selection																			
	LV.1				LV.2				LV.3				LV.4				LV.5			
	No. Attribute	DT	SVM	ANN	No. Attribute	DT	SVM	ANN	No. Attribute	DT	SVM	ANN	No. Attribute	DT	SVM	ANN	Attribute	DT	SVM	ANN
Dermatology	1	66.94	66.95	66.94	3	73.77	75.15	73.50	5	73.50	87.45	73.50	6	89.65	89.65	89.65	8	96.16	96.16	96.16
Glass	1	100.00	82.24	99.53	3	100.00	85.19	100.00	4	100.00	85.19	99.53	5	100.00	99.05	99.05	7	100.00	100.00	99.52
Iris	1	95.33	82.67	95.33	2	95.33	83.83	95.33	3	96.00	83.83	96.00	4	97.33	97.33	97.33	4	97.33	97.33	97.33
Lung-cancer	1	87.50	74.17	87.50	3	87.50	62.50	78.12	5	87.50	74.17	81.25	9	80.83	80.83	78.33	11	80.83	80.83	78.33
Movement	1	67.22	55.00	63.61	2	68.33	51.67	68.33	5	75.56	89.44	71.11	8	75.28	75.28	75.28	17	80.83	80.83	80.83
Page-blocks	1	96.86	96.29	96.53	2	97.94	99.60	96.55	5	98.06	99.60	97.20	7	98.06	97.44	97.44	9	98.05	98.05	97.44
Segmentation	1	86.15	85.67	85.89	2	87.06	92.64	86.67	5	89.00	91.21	89.00	9	98.00	98.00	98.00	10	96.41	96.41	96.41
Soybean	1	87.23	70.50	78.72	3	97.87	62.00	97.87	5	95.74	70.00	91.87	11	98.00	98.00	98.00	15	98.00	98.00	98.00
Tunadromd	1	93.28	93.28	93.28	2	96.64	96.64	96.64	4	96.86	96.75	96.75	6	96.98	97.76	97.42	6	96.98	97.76	97.42
Wine	1	83.71	83.82	83.71	3	87.64	88.24	87.64	5	87.64	88.24	87.64	7	87.16	87.16	87.16	10	86.50	86.50	86.50
Zoo	1	96.04	96.09	96.04	3	96.04	76.27	96.04	5	96.04	76.27	96.04	7	100.00	100.00	100.00	8	98.00	98.00	98.00

5.2.2. Feature Selection Using FCA

The other hierarchical method, FCA, was then applied to generate the hierarchical construct. We applied the ConExp program to build the concept lattice structure. Similarly, we select the attributes from each level of concept lattice from the first to the fifth level. The classification accuracy results are shown in Table 5.

Likewise, we compared classification accuracy of our feature selection using FCA to the benchmark original features (without using feature selection at all) in section 5.1, as shown in Figure 4.

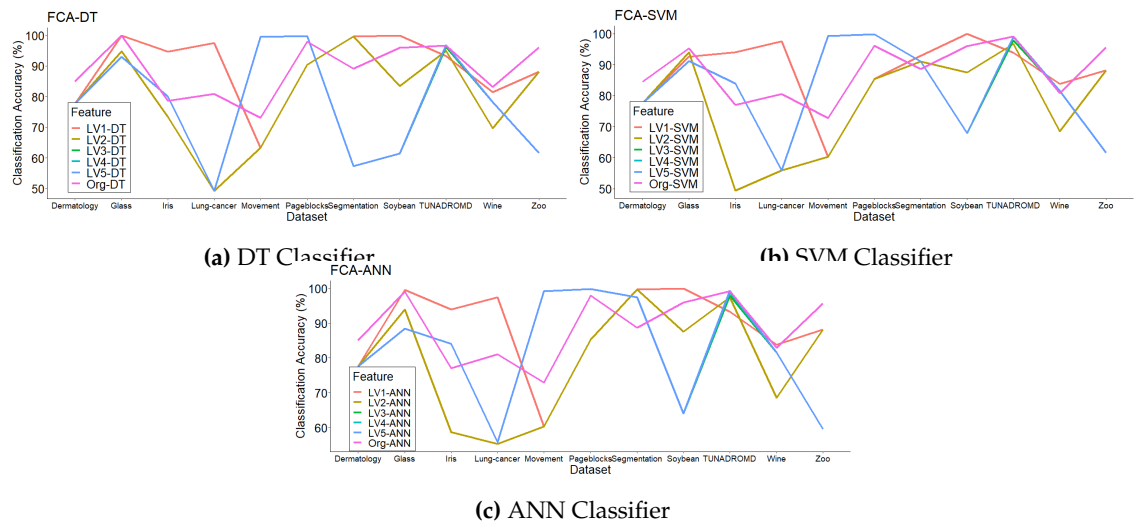


Figure 4. A comparison DT-SVM-ANN classifiers with feature selection using FCA from the first to the fifth level; and with all original features

From Figure 4, we see that the classification accuracy is not different between the original and the proposed model at each level. Indeed, if we add the number of attributes from each level, the trend in accuracy is increasing.

From the testing above, we find that the classification accuracy is not different between the original and the proposed model for each level. Thus, we can compare using paired sample t-tests as mentioned in section 3.

On considering the relationship between the number of selected attributes and the average classification accuracy for datasets, as shown in Figure 5, we found that the proposed model using decision tree gave a better classification result than using FCA.

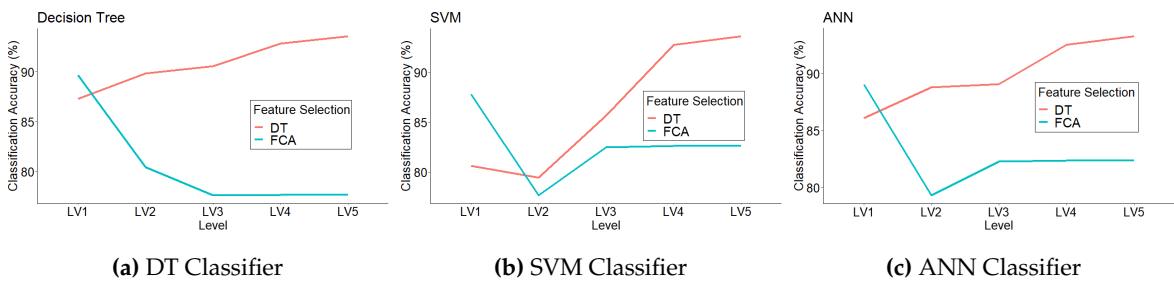


Figure 5. A comparison of average classification accuracies with each classifier model between features selected using DT or FCA

Table 5. The classification accuracy on using each level of FCA concept lattice for feature selection

Dataset	The classification accuracy (%) using FCA based feature selection																			
	LV.1				LV.2				LV.3				LV.4				LV.5			
	No. attribute	DT	SVM	ANN	No. attribute	DT	SVM	ANN	No. attribute	DT	SVM	ANN	No. attribute	DT	SVM	ANN	No. attribute	DT	SVM	ANN
Dermatology	13	77.58	77.58	77.58	21	77.58	77.58	77.58	24	77.58	77.58	77.58	28	77.58	77.58	77.58	33	77.58	77.58	77.58
Glass	1	100.00	92.60	99.55	3	94.87	93.96	93.96	5	93.01	91.17	88.38	10	93.01	91.17	88.38	10	93.01	91.17	88.38
Iris	2	94.67	94.00	94.00	3	73.33	49.33	58.67	4	80.00	84.00	84.00	4	80.00	84.00	84.00	4	80.00	84.00	84.00
Lung-cancer	5	97.50	97.50	97.50	8	49.17	55.83	55.25	10	49.17	55.83	55.83	16	49.17	55.83	55.83	32	49.17	55.83	55.83
Movement	7	63.33	60.28	60.28	12	63.33	60.28	60.28	19	99.44	99.17	99.17	28	99.44	99.17	99.17	32	99.44	99.17	99.17
Page-blocks	4	90.39	85.31	85.31	4	90.39	85.31	85.31	7	99.80	99.85	99.85	10	99.80	99.85	99.85	10	99.80	99.85	99.85
Segmentation	8	99.70	92.81	99.70	10	99.70	91.02	99.70	19	57.36	91.02	97.49	19	57.36	91.02	97.49	19	57.36	91.02	97.49
Soybean	2	100.00	100.00	100	12	83.50	87.50	87.50	16	61.50	68.00	64.00	35	61.50	68.00	64.00	35	61.50	68.00	64.00
Tunadromd	5	93.28	93.84	93.28	13	94.96	96.64	97.42	19	96.19	97.87	97.98	34	96.75	98.88	98.88	44	96.75	99.10	99.33
Wine	3	81.50	83.76	83.76	6	69.71	68.56	68.56	9	78.14	81.50	81.50	13	78.14	81.50	81.50	13	78.14	81.50	81.50
Zoo	7	88.18	88.18	88.18	10	88.18	88.18	88.18	12	61.55	61.55	59.55	17	61.55	61.55	59.55	17	61.55	61.55	59.55

5.2.3. An Example of Feature Selection Optimization

In our feature selection, we select the top node of structure to be representative of dataset. However, the proper selection of relevant attribute leads to an efficient classification. Thus, we experimented with using paired-samples t-test to optimize the selected level. We use the experimental results from Table 4 to demonstrate the proper selection of node attribute in each level. The Dermatology dataset is investigated to show the calculation for selecting attribute in each level of the tree. We assume the hierarchical structure of Dermatology dataset and the classification accuracies shown in Figure 6.

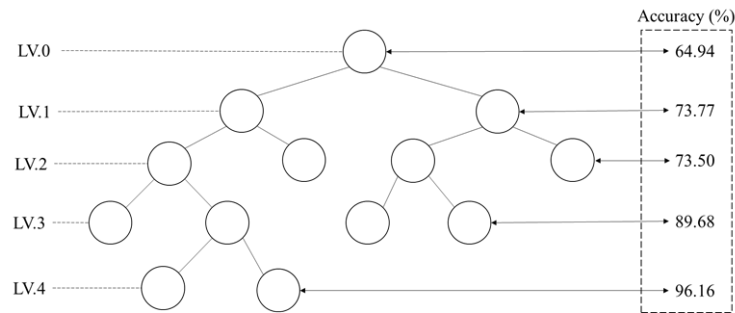


Figure 6. An example hierarchical structure using decision tree and the accuracies with selected features

From Figure 6, we can apply the proposed feature selection in an example calculation. The initial attribute node (or root) is selected with an accuracy of 64.94%. Next, we consider the next step for selection or not by using the equation 2 where $n = 366$, $\hat{p} = \frac{\sum_{l=i}^n Acc_{l,i} + \sum_{l=i}^n Acc_{l-1,i}}{n+n} = \frac{245+270}{366+366} = 0.704$, $\hat{q} = 1 - \hat{p} = 0.296$, thus

$$Z_{l,l-1} = \sqrt{\frac{366}{22 \cdot 0.704 \cdot 0.296}} (73.77 - 64.94) = 42.18.$$

From Table 1, if we consider the level of significance 0.01, if $|Z_{l,l-1}| \leq 2.576$, then we select the level 1 (LV.1 in Figure 6). Next, we still consider the next level (LV.2). We calculate $z_{2,1} = -1.61$ and find that this value is less than z in Table 1. Thus, LV.2 is not selected, and we stop the process. Thus, in this structure, we choose attributes from the structure in LV.0 to LV.2.

In addition, we estimate for all datasets which levels should be chosen. We applied the statistical paired-samples t-test to estimate this across all the datasets shown in Table 6. We find that the feature selection from each level depends on the classifier model. Namely, we can use two levels in a decision tree classifier, whereas we should select four levels for using SVM or ANN classifiers.

Table 6. The paired samples t-tests results

Methods	Level	$\pm SD$	Paired t-test	Sig.
DT	LV.1	86.69 $\pm$ 8.39	2.97	0.016
	LV.2	89.14 $\pm$ 6.74	2.55	0.031
	LV.3	91.00 $\pm$ 7.13	1.59	0.146
	LV.4	92.43 $\pm$ 3.07	2.22	0.053
	LV.5	92.51 $\pm$ 2.17	3.03	0.014
SVM	LV.1	79.34 $\pm$ 3.12	4.88	0.001
	LV.2	77.70 $\pm$ 4.07	4.15	0.002
	LV.3	84.54 $\pm$ 3.52	2.85	0.019
	LV.4	92.27 $\pm$ 0.94	2.46	0.036
	LV.5	93.21 $\pm$ 0.88	1.56	0.153
ANN	LV.1	86.28 $\pm$ 12.67	2.55	0.031
	LV.2	88.00 $\pm$ 6.50	3.20	0.011
	LV.3	89.51 $\pm$ 6.20	2.59	0.029
	LV.4	92.02 $\pm$ 3.45	2.35	0.043
	LV.5	92.85 $\pm$ 3.33	1.65	0.133

5.3. The Performances of Prior Feature Selection Approaches

This section evaluates the classification performance of feature selection using IG, CS, FS, and BS. IG and CS algorithms function as filter methods, while FS and BS algorithms operate as wrappers. These algorithms were implemented in the RapidMiner Studio program, representing traditional feature selection methods for comparison with our proposed model. The 10 datasets listed in Table 2 were employed for training and testing through a 10-fold cross-validation approach. The experimental results, presented in Table 7, presented the use of DT, SVM, and ANN as classifiers, with the "No." column indicating the number of selected attributes. The classification accuracies reported in this table will be further compared in the subsequent section.

Table 7. The classification performances of prior feature selection approaches

Dataset	IG				CS				FS				BS			
	DT	SVM	ANN	No.	DT	SVM	ANN	No.	DT	SVM	ANN	No.	DT	SVM	ANN	No.
Dermatology	92.35	93.00	92.35	12	86.07	87.29	86.07	7	85.79	87.85	84.64	9	85.79	87.85	85.79	33
Glass	99.53	100.00	100.00	4	99.53	99.82	100.00	3	100.00	100.00	100.00	9	100.00	100.00	100.00	3
Iris	95.33	95.33	95.00	2	95.33	94.22	95.33	2	96.00	96.04	96.00	1	96.00	96.04	96.67	3
Lung-cancer	68.50	68.50	69.00	2	87.50	86.50	86.50	16	87.50	87.50	83.50	55	87.50	87.50	87.50	2
Movement	66.64	66.65	65.45	16	75.00	74.51	76.02	34	76.94	82.50	77.85	88	79.44	82.50	83.33	6
Pageblocks	97.92	97.92	97.80	4	98.19	97.63	98.02	5	98.14	98.06	97.94	8	98.32	98.12	98.06	4
Segmentation	90.91	89.33	90.87	10	88.61	88.61	88.52	7	89.22	89.48	80.78	6	89.48	89.52	97.36	10
Soybean	97.87	100.00	98.87	14	100.00	100.00	98.87	10	100.00	100.00	97.87	2	100.00	100.00	100.00	34
Wine	82.58	82.50	82.00	5	83.15	83.15	84.22	6	85.96	88.22	90.45	9	87.08	88.48	89.89	10
Zoo	99.00	99.50	100.00	8	95.09	96.56	96.00	4	96.04	96.08	100.00	6	96.04	96.08	96.04	16

5.4. Comparison of Performances

We compared the classification accuracies of our feature selection using Decision Trees in Section 5.2 with others in Section 5.3, as illustrated in Figure 7. The figure depicts the classification performance after utilizing different numbers of attributes at each level (from Level 1 to 3, from top to bottom respectively) for each classification method denoted as Figure 7 (a) to (c), respectively. Detailed data

can be found in Tables 4 and 7. From the graphs in each dataset, considering the accuracy of attribute selection at each level, it appears that the performance is relatively consistent with other feature selection methods. For some datasets, the presented method may exhibit superior accuracy from the initial level, such as the Soybean dataset and Segmentation dataset. Conversely, the presented method may have lower accuracy for certain datasets, as observed in the Dermatology dataset.

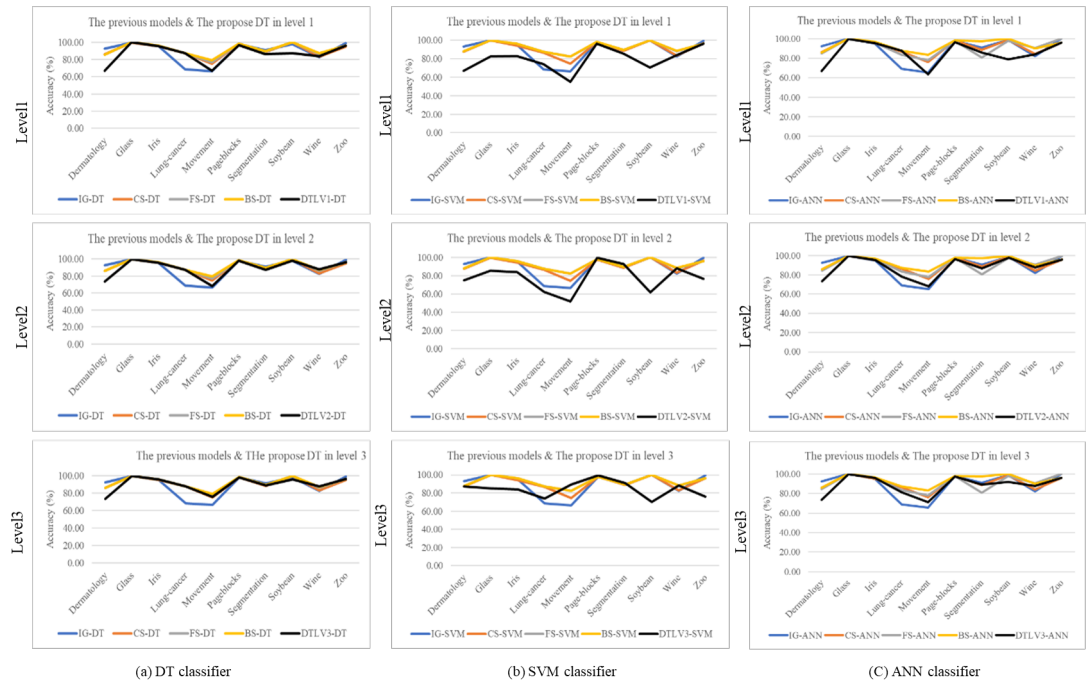


Figure 7. Comparison of proposed feature selection using DT on each level and others with DT-SVM-ANN classifiers.

Similarly, we conducted a comparison of the classification accuracy of our models using FCA in Section 5.2 with others in Section 5.3, illustrated in Figure 8. The figure demonstrates the classification performance after employing varying numbers of attributes at each level, with detailed data available in Tables 5 and 7. Upon examining the accuracy trends from this graph, we observe a similar pattern to use the decision tree approach, where some datasets may exhibit comparable performance to other methods, such as the Movement dataset, while others may show comparatively lower classification accuracy.

From above results, the outcomes obtained by employing a hierarchical structure for feature selection before the classification task indicate that it could lead to either favorable or less favorable results. This corresponds to the observed patterns seen in other feature selection methods used for comparison. Importantly, the classification performance of the proposed method does not consistently appear inferior. However, the presented key point in this study is its ability to significantly reduce the number of attributes compared to previous methods.

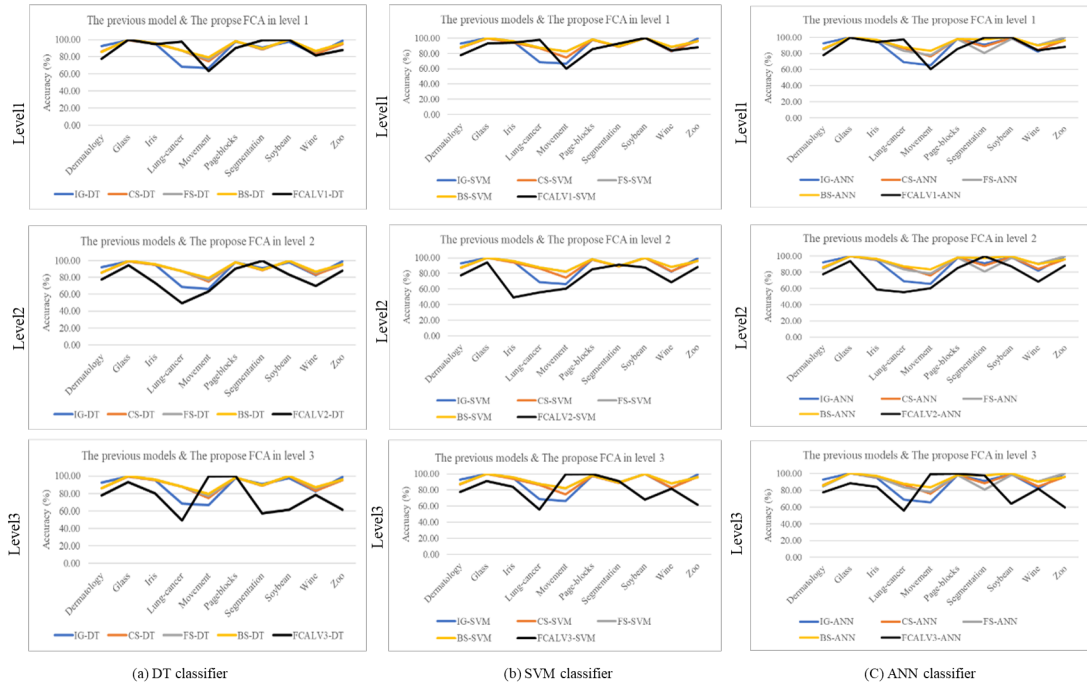


Figure 8. Comparison of proposed feature selection using FCA in each level to others on using DT-SVM-ANN classifiers.

Considering the number count of selected features, the results show clearly that the DT kept less features than the other methods, as shown in Figure 9. This figure will illustrate the number of selected attributes at each level from 1 to 3 in comparison with other methods that exhibit classification performance as shown in Figures 7 and 8 mentioned earlier. Figure 9 (a) displays the count of selected attributes from the decision tree structure, while Figure 9 (b) shows the count of selected attributes from the FCA structure. From as the results, the hierarchical structure using decision tree method can be applied to select the feature for data dimension reduction yields favorable outcomes. This is due to the decision tree’s method, which involves a top-down data entry process using specific filtering and selecting root node. However, FCA produces less satisfactory results due to the creation of a layer-wise structure designed for bidirectional ordering—both from top to bottom and from bottom to top. This may lead to effect of feature selection.

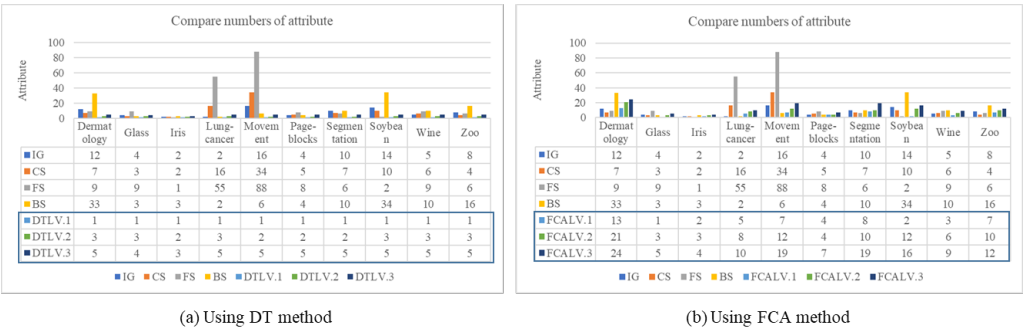


Figure 9. The number of selected features by (a) DT and (b) FCA methods in each level, compared to the other selection methods.

6. Conclusions

Feature selection methods are used in a wide variety of applications. Such methods proceed to eliminate unfavorable features that may be noisy, redundant or irrelevant, as these can penalize the performance of a classifier. Feature selection reduces the dimensionality of data and restricts the

inputs to a subset of the original, but the features can have missing values. In this paper, we propose a new feature selection method based on hierarchical concept model to improve the performance of supervised learning. We applied decision tree and FCA to build the hierarchical structures from data to generate knowledge based systems. In these hierarchical structures that embed knowledge, the top and bottom nodes represent general and specific knowledge, respectively. For this reason, we selected the top node of each structure to be representative of the dataset. However, the proper selection of relevant attributes leads to an efficient classification. Thus, we also used paired-samples t-tests to optimize the selected levels in the hierarchical structures. In this study, numerical experiments used available software, RapidMiner Studio and ConExp programs, to generate the hierarchical structures for decision tree and FCA, respectively. Afterwards, the attributes in top node of the obtained structure were selected to use in learning classifier models. Three popular classifier algorithms were used in this study: decision trees, SVM, and ANN. To compare the performances of our feature selection methods with available alternatives: all original features (without using feature selection) and the prior feature selection methods. We used 10 datasets from the UCI Machine Learning Repository for classification. The prior feature selection methods tested were IG, CS, FS, and BS. The results show clearly that the proposed models with DT and FCA gave smaller set of retained attributes than the other methods, while still classification performance remained comparable. In other words, the new approach provided superior data reduction to the least dimensionality without performance sacrifice.

Acknowledgments: The authors are deeply grateful to the Faculty of Science and Industrial Technology, Prince of Songkla University, Surat Thani Campus, Thailand. This research was supported by National Science, Research and Innovation Fund (NSRF) and Prince of Songkla University (Ref. No. (SIT6601016S). The authors gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

References

1. Dhal, P.; Azad, C. A comprehensive survey on feature selection in the various fields of machine learning. *Applied Intelligence* **2022**, *52*, 4543–4581. doi:10.1007/s10489-021-02550-9.
2. Khaire, U.M.; Dhanalakshmi, R. Stability of feature selection algorithm: A review. *Journal of King Saud University - Computer and Information Sciences* **2022**, *34*, 1060–1073. doi:https://doi.org/10.1016/j.jksuci.2019.06.012.
3. Solorio-Fernández, S.; Carrasco-Ochoa, J.A.; Martínez-Trinidad, J.F. A new hybrid filter-wrapper feature selection method for clustering based on ranking. *Neurocomputing* **2016**, *214*, 866–880. doi:https://doi.org/10.1016/j.neucom.2016.07.026.
4. Cai, J.; Luo, J.; Wang, S.; Yang, S. Feature selection in machine learning: A new perspective. *Neurocomputing* **2018**, *300*, 70–79. doi:https://doi.org/10.1016/j.neucom.2017.11.077.
5. Zhao, H.; Wang, P.; Hu, Q.; Zhu, P. Fuzzy Rough Set Based Feature Selection for Large-Scale Hierarchical Classification. *IEEE Transactions on Fuzzy Systems* **2019**, *27*, 1891–1903. doi:10.1109/TFUZZ.2019.2892349.
6. Bolón-Canedo, V.; Alonso-Betanzos, A. Ensembles for feature selection: A review and future trends. *Information Fusion* **2019**, *52*, 1–12. doi:https://doi.org/10.1016/j.inffus.2018.11.008.
7. Zebari, R.; Abdulazeez, A.; Zeebaree, D.; Zebari, D.; Saeed, J. A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction. *Journal of Applied Science and Technology Trends* **2020**, *1*, 56 – 70. doi:10.38094/jastt1224.
8. Wan, C.; Freitas, A.A. An empirical evaluation of hierarchical feature selection methods for classification in bioinformatics datasets with gene ontology-based features. *Artificial Intelligence Review* **2018**, *50*, 201–240. doi:10.1007/s10462-017-9541-y.
9. Wetchapram, P.; Muangprathub, J.; Choopradit, B.; Wanichsombat, A. Feature Selection Based on Hierarchical Concept Model Using Formal Concept Analysis. 2021 18th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2021, pp. 299–302. doi:10.1109/ECTI-CON51831.2021.9454732.
10. Hancer, E.; Xue, B.; Zhang, M. A survey on feature selection approaches for clustering. *Artificial Intelligence Review* **2020**, *53*, 4519–4545. doi:10.1007/s10462-019-09800-w.

11. Cerrada, M.; Sánchez, R.V.; Pacheco, F.; Cabrera, D.; Zurita, G.; Li, C. Hierarchical feature selection based on relative dependency for gear fault diagnosis. *Applied Intelligence* **2016**, *44*, 687–703. doi:10.1007/s10489-015-0725-3.
12. Guo, S.; Zhao, H.; Yang, W. Hierarchical feature selection with multi-granularity clustering structure. *Information Sciences* **2021**, *568*, 448–462. doi:https://doi.org/10.1016/j.ins.2021.04.046.
13. Tuo, Q.; Zhao, H.; Hu, Q. Hierarchical feature selection with subtree based graph regularization. *Knowledge-Based Systems* **2019**, *163*, 996–1008. doi:https://doi.org/10.1016/j.knosys.2018.10.023.
14. Zheng, J.; Luo, C.; Li, T.; Chen, H. A novel hierarchical feature selection method based on large margin nearest neighbor learning. *Neurocomputing* **2022**, *497*, 1–12. doi:https://doi.org/10.1016/j.neucom.2022.05.016.
15. Trabelsi, M.; Meddouri, N.; Maddouri, M. A New Feature Selection Method for Nominal Classifier based on Formal Concept Analysis. *Procedia Computer Science* **2017**, *112*, 186–194. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 21st International Conference, KES-20176-8 September 2017, Marseille, France, doi:https://doi.org/10.1016/j.procs.2017.08.227.
16. Azibi, H.; Meddouri, N.; Maddouri, M. Survey on Formal Concept Analysis Based Supervised Classification Techniques. In *Machine Learning and Artificial Intelligence*; IOS Press, 2020; pp. 21–29.
17. Wang, C.; Huang, Y.; Shao, M.; Hu, Q.; Chen, D. Feature Selection Based on Neighborhood Self-Information. *IEEE Transactions on Cybernetics* **2020**, *50*, 4031–4042. doi:10.1109/TCYB.2019.2923430.
18. Wille, R. Formal concept analysis as mathematical theory of concepts and concept hierarchies. *Lecture notes artificial intelligent (LNAI)* **2005**, *3626*, 1–33. doi:https://doi.org/10.1016/j.neucom.2021.03.002.
19. Zhou, H.; Zhang, J.; Zhou, Y.; Guo, X.; Ma, Y. A feature selection algorithm of decision tree based on feature weight. *Expert Systems with Applications* **2021**, *164*, 113842. doi:https://doi.org/10.1016/j.eswa.2020.113842.
20. Venkatesh, B.; Anuradha, J. A Review of Feature Selection and Its Methods. *Cybernetics and Information Technologies* **2019**, *19*, 3–26. doi:doi:10.2478/cait-2019-0001.
21. Bahassine, S.; Madani, A.; Al-Sarem, M.; Kissi, M. Feature selection using an improved Chi-square for Arabic text classification. *Journal of King Saud University - Computer and Information Sciences* **2020**, *32*, 225–231. doi:https://doi.org/10.1016/j.jksuci.2018.05.010.
22. Trivedi, S.K. A study on credit scoring modeling with different feature selection and machine learning approaches. *Technology in Society* **2020**, *63*, 101413. doi:https://doi.org/10.1016/j.techsoc.2020.101413.
23. Zhao, H.; Hu, Q.; Zhu, P.; Wang, Y.; Wang, P. A Recursive Regularization Based Feature Selection Framework for Hierarchical Classification. *IEEE Transactions on Knowledge and Data Engineering* **2021**, *33*, 2833–2846. doi:10.1109/TKDE.2019.2960251.
24. Huang, H.; Liu, H. Feature selection for hierarchical classification via joint semantic and structural information of labels. *Knowledge-Based Systems* **2020**, *195*, 105655. doi:https://doi.org/10.1016/j.knosys.2020.105655.
25. Liu, X.; Zhao, H. Robust hierarchical feature selection with a capped ℓ_2 -norm. *Neurocomputing* **2021**, *443*, 131–146. doi:https://doi.org/10.1016/j.neucom.2021.03.002.
26. Asuncion, A.; Newman, D. UCI machine learning repository, 2007.
27. Yevtushenko, S. Concept Explorer, Open source java software. <http://sourceforge.net/projects/conexp> **2009**.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.