Universal Feature Selection Tool

User Manual, Version 1.1

Sina Tabakhi, Parham Moradi

September 2022

Contents

Acknowledgments

The UniFeat is implemented and developed principally by Sina Tabakhi and Parham Moradi, but other people make essential contributions. We are grateful to Mr. Shahin Salavati, who implemented most of the graphical user interfaces. We also thank the professors who provided us with feedback about the tool, especially Dr. Fardin Akhlaghian at the University of Kurdistan and Dr. Haiping Lu at the University of Sheffield.

About the Authors

Sina Tabakhi is currently a Ph.D. student in the Department of Computer Science at the University of Sheffield, United Kingdom, working in the Machine Learning Research Group. Before joining the University of Sheffield, he had worked in the industry as a software engineer for five years. He graduated with a Master's degree as the first honor student in Computer Engineering, Artificial Intelligence from the University of Kurdistan, Iran. He has been working on efficient feature selection methods for integrative multi-omics analysis to tackle the curse of the dimensionality phenomenon of these data types.

Parham Moradi received the M.Sc. and Ph.D. degrees in Computer Science from Amirkabir University of Technology, Iran, in 2005 and 2011, respectively. He conducted a part of his Ph.D. research work in the Laboratory of Nonlinear Systems, Ecole Polytechnique Federal de Lausanne (EPFL), Lausanne, Switzerland, from September 2009 to March 2010. He is currently an associate professor in the Department of Computer Engineering at the University of Kurdistan, Iran. His research focuses on machine learning, social network analysis, recommender systems, and deep learning.

1 Introduction

The field of data mining is concerned with knowledge discovery from data through the development of computer programs. During the last two decades, the rapid advances in computer and database technologies have led to the production of datasets with large numbers of features in many fields [13]. Most of the features are irrelevant and redundant, and these unnecessary features have stimulated a phenomenon in the data mining algorithms called the curse of dimensionality [5, 9, 42]. The dimensionality reduction technique plays an essential role in addressing this issue by mitigating the dimensions of features and retaining informative features of the original data. This technique is performed based on either feature selection or feature extraction approaches [21]. Feature selection is the process of identifying a subset of relevant features in an original feature set, while feature extraction methods transform the original data into a lower dimensional space to derive informative and non-redundant features. Feature selection is regarded as an important and active research area in data preprocessing, and numerous methods have been developed based on this technique [40].

From one aspect, feature selection methods can be classified into four approaches, including filter, wrapper, embedded, and hybrid [35, 27, 33, 38, 26]. The filter approach estimates the relevance of features using intrinsic properties of the data without the need for any learning algorithms. This approach can be subdivided into univariate and multivariate [38, 31, 2]. The univariate filter approach examines the relevance of each feature individually based on a given criterion. In contrast, the multivariate filter approach selects a subset of features by considering the dependencies between features. The wrapper approach integrates a specific learning algorithm to evaluate different subsets of selected features within the feature selection process. The embedded approach considers the feature selection process as part of constructing a given learning algorithm. Finally, the hybrid approach uses filter-based methods to reduce the original feature set in the first step and then applies wrapper-based techniques to select the final feature set.

From another aspect, the feature selection methods can be classified into supervised and unsupervised modes [13, 26]. In the supervised mode, the class labels of data are applied in the feature selection process as a guide, but in the unsupervised mode, the process of feature selection is done without using class labels.

Several tools and libraries have been developed for the feature selection task. A collection of existing feature selection tools is presented and compared in Table 1.1. In this collection, we have incorporated the open-source tools that have been recently updated, have been maintained for several years, or have been endorsed by the number of stars on their repository hosts (e.g., GitHub and SourceForge). Several vital metrics in the development of research software have been considered for comparing tools in Table 1.1:

- Programming language: This metric indicates developers' most used programming languages in the feature selection research area.
- Documentation: This criterion demonstrates how well the software is documented

for end-users and developers in terms of installation instructions, illustrative examples of use cases of the software, complete API documentation, and development tutorials to extend the software.

- Availability of graphical user interfaces (GUI): This feature shows how easy and quick the tool is to use for end-users.
- Data format: This measure points out the tool's support for various input data formats.
- Creation and last update: These two criteria indicate the age and maintenance status of the software.
- Main focus: This column specifies the primary goal of software development, which can be classified into two categories: data mining and feature selection tools. Data mining tools provide a general-purpose environment for machine learning models with different features such as data preprocessing, classification, regression, clustering, and visualization. In these tools, feature selection can only be considered a small module. On the other hand, feature selection tools are designed specifically for the feature selection task and provide a wide range of feature selection methods.
- Coverage of feature selection approaches: This indicator figures out the support of the software in implementing baseline, well-known, and advanced feature selection methods in different categories.

The analysis of Table 1.1 shows that the development guide is lacking in most software, and the API documentation has not been provided with several tools. Therefore, developers find it challenging to modify and extend the existing software. Another essential drawback with some tools, such as Weka [48], is that the documentation is not up-to-date to cover the newest features and functionalities. Moreover, since non-expert researchers prefer to explore the software more efficiently without any coding requirements or through command-line environments, providing a high-level representation of the software functionalities via a GUI has become a standard in software development. Nonetheless, only a few available tools support a graphical interface for the users.

Another critical gap in Table 1.1 that deserves attention is the tool's coverage of the feature selection methods. Most existing tools focus on only one feature selection approach and ignore the others. For example, Mlxtend [37] is a data mining library whose feature selection module only contains baseline greedy wrapper-based methods. Another example is Weka which is general-purpose software for data mining, but it provides only a few conventional feature selection methods based on the filter and wrapper approaches. Besides, a small number of tools have incorporated all feature selection approaches in their codebase with the implementation of simple, baseline, and well-known methods. However, the community still demands new advanced feature selection methods. RapidMiner [17], as an example, is an integrated platform for the generation of machine learning models in which many representative filter-based feature selection methods are implemented, but advanced ones are missing. Furthermore, only a few baseline methods based on the wrapper and embedded approaches are available in the RapidMiner repository. Among the software provided in Table 1.1, only scikit-feature

Table 1.1: Comparison of UniFeat to existing feature selection tools.

Tool name	**Programming language**	**License**	**Documentation**	**GUI**	**Data format**	**Creation**	**Last update**	**Main focus**	**Feature selection approach**		
									Filter	**Wrapper**	**Embedded**
Weka [48]	Java	GNU GPL	Installation instructions Illustrative examples API documentation Development guide	✓	ARFF, CSV, XRFF, XML	1993	2022	Data mining	Few baseline methods are implemented	A small number of conventional methods are included	N/A
RapidMiner [17]	Java	GNU AGPL A proprietary license	Installation instructions Illustrative examples API documentation Development guide	✓	ACCDB, ARFF, CSV, DBF, DTA, HYPER, MDB, QVX, SAS, SAV, TDE, XLS/XLSX, XML, XRFF	2001	2022	Data mining	Many representative methods are implemented, but advanced ones are missing	Few baseline methods are implemented	Few baseline methods are implemented
Scikit-learn [32]	Python	BSD 3-Clause	Installation instructions Illustrative examples API documentation Development guide	N/A	CSV, XLS/XLSX, JSON, MAT, ARFF, SQL, Numpy arrays, LibSVM format	2007	2022	Data Mining	A small set of simple baseline methods are provided	Few sequential techniques are included	Some baseline and well-known methods are involved
Mlxtend [37]	Python	BSD 3-Clause	Installation instructions Illustrative examples API documentation Development guide	N/A	CSV	2014	2022	Data mining	N/A	A few greedy methods are implemented	N/A
ITMO_FS [34]	Python	BSD 3-Clause	Installation instructions Illustrative examples API documentation	N/A	Scikit-learn input formats	2018	2022	Feature selection	Many well-known and advanced methods are included	Several conventional methods are provided	A few advanced techniques are implemented
FeatureSelector [22]	Python	GNU GPL	Illustrative examples API documentation	N/A	CSV	2018	2022	Feature selection	A few traditional methods are implemented	N/A	A few simple baseline techniques are involved
Feature-engine [11]	Python	BSD 3-Clause	Installation instructions Illustrative examples API documentation Development guide	N/A	Scikit-learn input formats	2019	2022	Feature engineering and selection	Several conventional statistical methods are incorporated	A few traditional techniques are provided	A small number of baseline methods are involved
Jx-WFST [46]	MATLAB	BSD 3-Clause	Illustrative examples	N/A	MAT	2020	2021	Feature selection	N/A	Many representative and advanced algorithms are implemented	N/A
Mulan [47]	Java	GNU GPL	Installation instructions Illustrative examples API documentation Development guide	N/A	XML, ARFF	2007	2020	Multi-label learning	Few baseline methods have been implemented	N/A	N/A
Feature Selection for Machine Learning [8]	Python	N/A	Illustrative examples	N/A	Pandas input formats	2018	2020	Feature selection	Some conventional statistical methods are provided	A small number of greedy algorithms are included	N/A
FEAST [4]	MATLAB C/C++	BSD 3-Clause	Installation instructions Illustrative examples	N/A	MAT	2011	2019	Feature selection	Many standard mutual information-based methods are implemented	N/A	N/A
Scikit-feature [25]	Python	GNU GPL	Installation instructions Illustrative examples API documentation	✓	MAT, CSV	2015	2019	Feature selection	A range of well-known and advanced methods are included	A few greedy techniques are implemented	A few representative methods are provided
FeatureSelect [28]	MATLAB	MIT	Installation instructions Illustrative examples	✓	MAT, XLS, TXT	2018	2019	Feature selection	A small number of baseline methods are included	A set of well-known and advanced methods are provided	N/A
MLFeatureSelection [6]	Python	MIT	Installation instructions Illustrative examples	N/A	Pandas input formats	2018	2019	Feature selection	N/A	A small number of methods are implemented	N/A
LOFS [51]	MATLAB OCTAVE	GNU GPL	Installation instructions Illustrative examples API documentation	N/A	MAT, CSV	2016	2016	Online feature selection	N/A	Some online methods are involved	N/A
UniFeat	Java	MIT	Installation instructions Illustrative examples API documentation Development guide	✓	CSV	2022	2022	Feature selection	Many well-known and advanced algorithms are implemented	Some representative and advanced methods are included	Several baseline and popular methods are provided

[25] and ITMO_FS [34] have implemented numerous feature selection methods in all categories. However, scikit-feature has not developed advanced feature selection methods in the wrapper and embedded classes, and its codebase has not been updated for several years. Moreover, ITMO_FS has provided several conventional wrapper feature selection methods, and the implementation of advanced techniques is still needed.

Therefore, our aim in developing the Universal Feature Selection Tool (UniFeat) as a comprehensive feature selection tool includes seven aspects. (1) UniFeat implements well-known and advanced feature selection methods within a unified framework to respond to the pressing requirements of the community. (2) UniFeat can be considered as a benchmark tool due to the development of methods in all the approaches. (3) The functions presented in UniFeat provide essential auxiliary tools for performance evaluation, result visualization, and statistical analysis. (4) UniFeat has been completely implemented in Java and can be run on various platforms. (5) Researchers are able to use UniFeat through its GUI environment or as a library in their Java codes. (6) The open-source nature of UniFeat can help researchers use and modify the tool to fit their research requirements and facilitate sharing their methods with the scientific community rapidly. (7) Finally, a well-documented tutorial for developers and end-users is provided for UniFeat to support the further extension of the software.

2 Introduction to UniFeat

The **Uni**versal **Feat**ure Selection Tool (UniFeat) is an open-source Java tool for feature selection, developed at the University of Kurdistan, Iran, and distributed under the MIT License[1] terms. The project aims to create a unified framework for researchers applying feature selection.

For simplification of the development of the tool, UniFeat was divided into six main packages, including (1) featureSelection, (2) dataset, (3) classifier, (4) gui, (5) result, and (6) util (as shown in Figure 2.1), used for the following purposes.

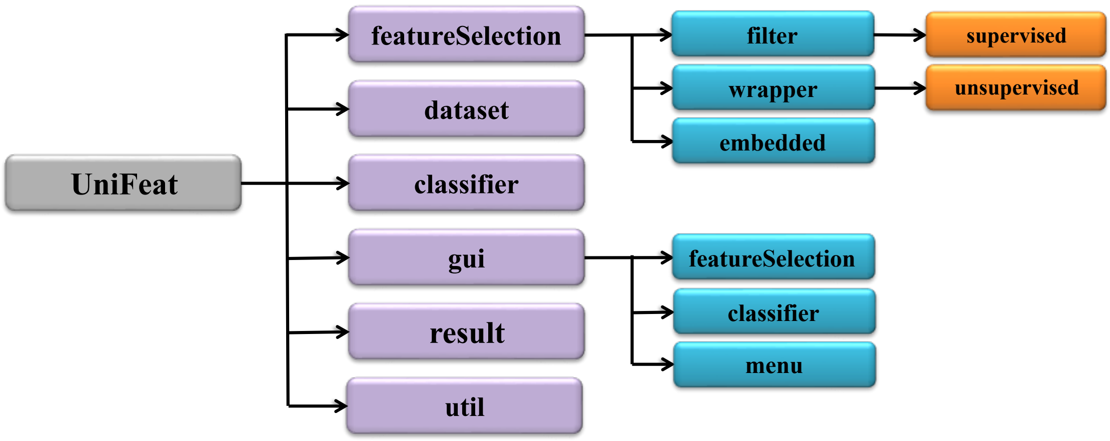


Figure 2.1: UniFeat packages.

1. **featureSelection package**: provides all the feature selection methods implemented in the tool. This package is divided into three sub-packages to cover all the feature selection approaches. Moreover, filter-based feature selection methods have been split into supervised and unsupervised packages. The current feature selection methods in the UniFeat repository are based on the filter, wrapper, and embedded approaches. The unified interface of the package allows researchers to implement their feature selection methods and share them with the other researchers in the feature selection community.
2. **dataset package**: is used for loading, saving, editing, and exporting different types of dataset files.
3. **classifier package**: collects several well-known and frequently used classifiers from the Weka software package [48].
4. **gui package**: provides GUIs that display the entire graphical representation of the panels for interaction with the user. Moreover, the package reports the results visually. It should be noted that this package has been separated from the others.
5. **result package**: reports the performance results of feature selection methods based on several criteria such as accuracy and execution time.

[1] https://github.com/UniFeat/unifeat/blob/main/LICENSE

6. **util package**: presents various utility methods for manipulating arrays and matrices and performing basic statistical operations that can be used in the feature selection methods.

UniFeat is entirely implemented in Java, and it can thus be run on any platform where Java Runtime Environment (JRE) is installed.

3 Download and run

Two types of files are provided for the UniFeat:

1. The executable file of UniFeat (version 0.1.1) that can be downloaded from the project website[2].
2. The source-code of UniFeat (version 0.1.1) which is available in the GitHub repository[3].

After downloading the tool, you must have the Java Runtime Environment (JRE) on your system to run it. Also, if you want to use the source codes and modify them, you need the Java Development Kit (JDK) to compile the modified source codes again.

You can start the UniFeat as a graphical user interface by clicking the UniFeat-v0.1.1.jar file or by typing the following command from the command prompt:

```
java -jar UniFeat-v0.1.1.jar
```

Figure 3.1 shows the initial panel of the UniFeat. This panel is used to select a workspace path for the tool. It should be noted that some essential files will be created in this path.

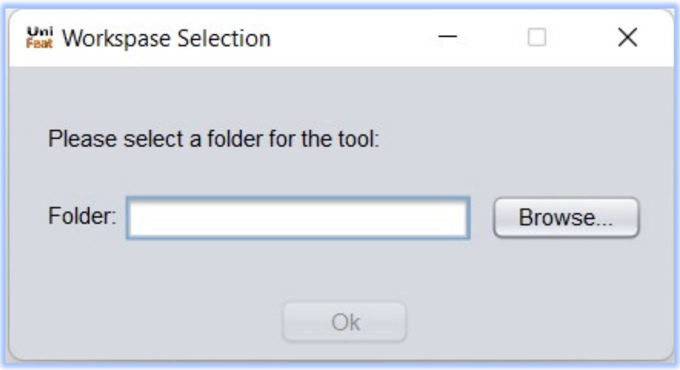


Figure 3.1: Workspace selection panel.

The easiest way to use UniFeat is through its graphical user interface. Another way is to use UniFeat as a library (using UniFeat-v0.1.1.jar) for researchers who use feature selection as a part of their own methods and, therefore, prefer to embed the UniFeat feature selection methods in their Java codes. The detailed information is described in Section 5.

[2] https://unifeat.github.io/software.html
[3] https://github.com/UniFeat/unifeat

4 The UniFeat exploration

After running UniFeat and selecting a workspace for the tool, you will see the main panel of the tool, which is illustrated in Figure 4.1. This panel gives all the tool facilities access using form filling and menu items.

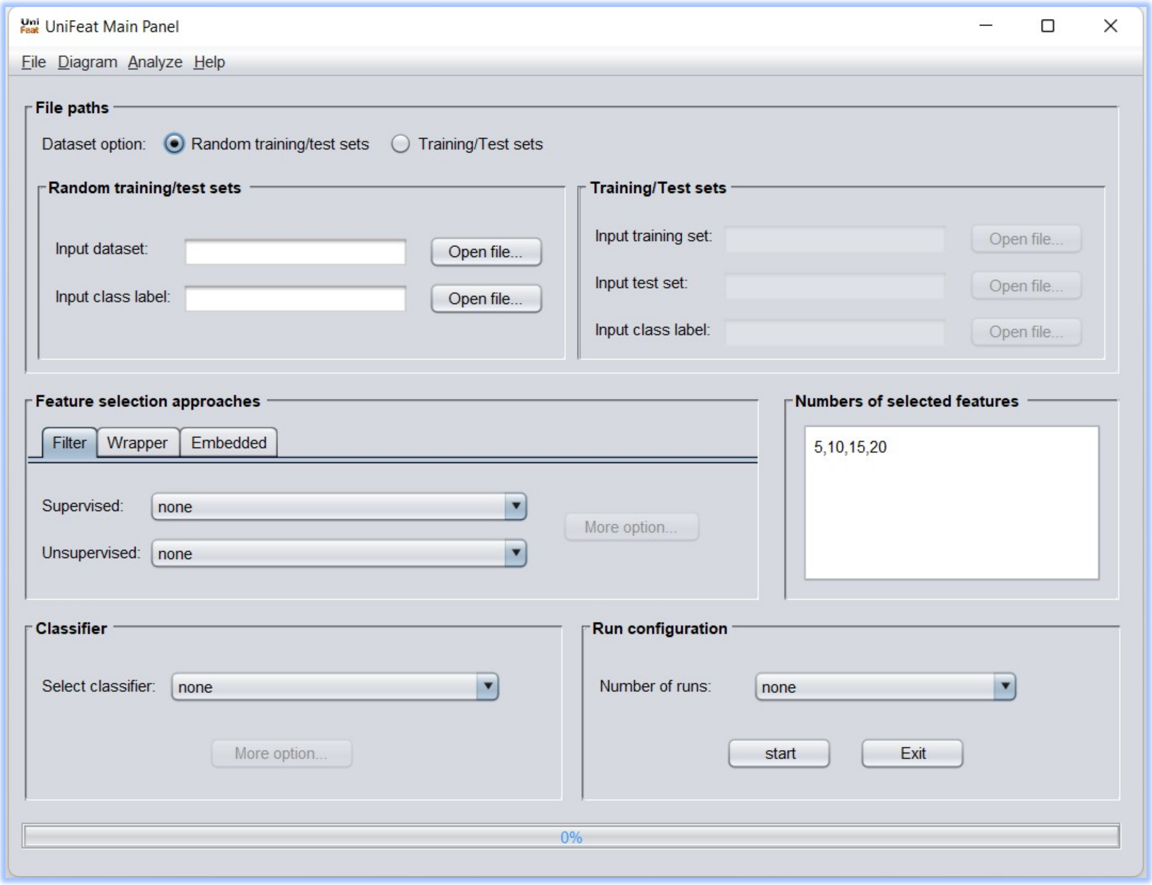


Figure 4.1: Main panel of UniFeat.

4.1 Panel description

There are five different parts corresponding to the specific task of the UniFeat tool. Each of the five parts is described in the following subsections.

4.1.1 Loading the dataset files

In the feature selection methods, datasets are split into training and test sets. The training set is a portion of the data that is used in the feature selection process. Moreover, this set is employed for training the learning algorithm. On the other hand, the test set is the unseen data that is used to evaluate the performance of the feature selection methods. Generally speaking, benchmark datasets which are provided for the feature selection domain are available in two types. In the first type, the dataset file consists of all the samples. In the second type, the datasets are divided originally into two portions: training and test sets. UniFeat provides support to both types of dataset files.

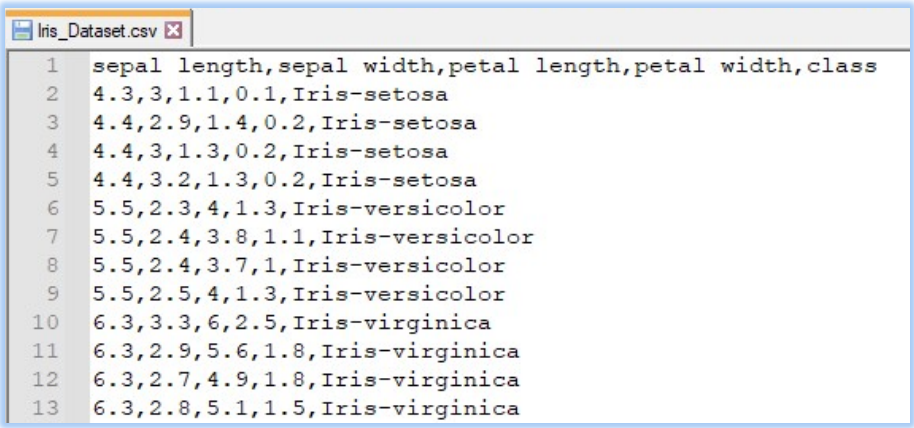
The “File paths” panel allows you to load all dataset files in the tool for future purposes:

1. “Random training/test sets” section: you can easily import a file of the dataset, and then the training and test sets are drawn randomly by the tool from the input dataset file (2/3 of the data are considered as a training set, and the other portion is used as a test set).
2. “Training/Test sets” section: if you want to use the datasets that are divided originally into training and test sets, this section can be used.

A specific way of representing datasets is needed for the tool. In UniFeat format, the representation of the input dataset consists of the following parts:

- The first row of the datasets must have the names of all features.
- The next rows contain all data, with each row corresponding to a sample. A vector of feature values describes each sample with a class label separated by commas. Also, the class labels of all samples must be available in the last column of the dataset.

You can easily import the dataset into the UniFeat tool as a file in comma-delimited format (i.e., CSV file format). Figure 4.2 shows an example of a dataset in the UniFeat format. In Figure 4.2, the input dataset contains 12 samples and four features within the class feature. In this case, the class feature has three values, including *Iris-setosa*, *Iris-versicolor*, and *Iris-virginica*.




```
sepal length,sepal width,petal length,petal width,class
4.3,3,1.1,0.1,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.4,3,1.3,0.2,Iris-setosa
4.4,3.2,1.3,0.2,Iris-setosa
5.5,2.3,4,1.3,Iris-versicolor
5.5,2.4,3.8,1.1,Iris-versicolor
5.5,2.4,3.7,1,Iris-versicolor
5.5,2.5,4,1.3,Iris-versicolor
6.3,3.3,6,2.5,Iris-virginica
6.3,2.9,5.6,1.8,Iris-virginica
6.3,2.7,4.9,1.8,Iris-virginica
6.3,2.8,5.1,1.5,Iris-virginica
```

Figure 4.2: An example of the UniFeat format of a dataset.

In addition to the dataset files, a separate file that contains the values of the class feature must be imported into the tool. Each value of the class feature is presented in a row. For example, for the dataset in Figure 4.2, the class label file contains three rows, each representing a class label, including *Iris-setosa*, *Iris-versicolor*, and *Iris-virginica*. It should be noted that rows of the class label file must be compatible with the class

Table 4.1: Filter-based feature selection methods in the UniFeat repository.

Method	Supervised/Unsupervised	Multivariate/Univariate
Information gain [29]	Supervised	Univariate
Gain ratio [29]	Supervised	Univariate
Symmetrical uncertainty [26]	Supervised	Univariate
Fisher score [12]	Supervised	Univariate
Gini index [39]	Supervised	Univariate
mRMR [33]	Supervised	Multivariate
Laplacian score [16]	Supervised & unsupervised	Univariate
RRFS [9]	Supervised & unsupervised	Multivariate
Term variance [45]	Unsupervised	Univariate
Mutual correlation [15]	Unsupervised	Multivariate
RSM [23]	Unsupervised	Multivariate
UFSACO [43]	Unsupervised	Multivariate
RRFSACO_1 [42]	Unsupervised	Multivariate
RRFSACO_2 [42]	Unsupervised	Multivariate
IRRFSACO_1 [42]	Unsupervised	Multivariate
IRRFSACO_2 [42]	Unsupervised	Multivariate
MGSACO [44]	Unsupervised	Multivariate

feature values in the dataset file.

A list of some benchmark datasets from different sources that were converted to UniFeat format is available on the project website[4].

4.1.2 Choosing a feature selection method

In the “Feature selection approaches” panel, you can simply access to the well-known and state-of-the-art feature selection methods in the literature. In this panel, there are three tabs corresponding to the different feature selection approaches including the “Filter,” “Wrapper,” and “Embedded” tabs.

In the “Filter,” “Wrapper,” and “Embedded” tabs, the UniFeat repository has involved the feature selection methods, the details of which are listed in Tables 4.1 to 4.3, respectively.

Some feature selection methods have adjustable parameters that need to be set. The “More option...” button is provided in the tool to set these parameters. An example is presented in Figure 4.3 for the Laplacian score method. It should be noted that to prevent unwanted errors, the tool automatically checks the values. In other words, when the value of a parameter is empty or incorrect, a star symbol ‘*’ immediately appears in front of the parameter to alert the user. In Figure 4.3, this issue arises for the “k-nearest neighbor” parameter.

[4]https://unifeat.github.io/datasets.html

Table 4.2: Wrapper-based feature selection methods in the UniFeat repository.

Method	Supervised/Unsupervised
Binary particle swarm optimization (BPSO) [49]	Supervised
Continuous particle swarm optimization (CPSO) [49]	Supervised
Particle swarm optimization version 4-2 (PSO(4-2)) [50]	Supervised
HPSO-LS [30]	Supervised
Simple GA [18]	Supervised
HGAFS [19]	Supervised
Optimal ACO [1]	Supervised

Table 4.3: Embedded-based feature selection methods in the UniFeat repository.

Method	Supervised/Unsupervised
Decision tree based method [24]	Supervised
Random forest [3]	Supervised
SVM_RFE [14]	Supervised
MSVM_RFE [20]	Supervised
OVO_SVM_RFE [7]	Supervised
OVA_SVM_RFE [7]	Supervised

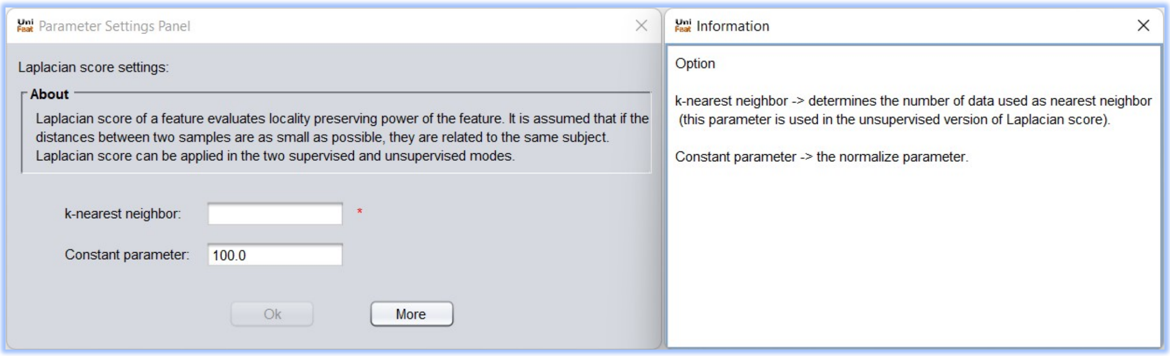


Figure 4.3: An example of the parameter settings panel.

4.1.3 Feature subset sizes

In some feature selection methods, the number of selected features is a parameter that needs to be set. The "Numbers of selected features" panel is designed to set this parameter. Therefore, users are able to enter the different numbers of selected features altogether. These values must be separated by commas. Figure 4.1 shows an example of how 5, 10, 15, and 20 features should be selected by the given feature selection method.

4.1.4 Selecting classifier

In the "Classifier" panel, you can select a classifier for evaluating the subsets of features chosen by a given feature selection method. Four frequently used classifiers, including support vector machine (SVM) [14], decision tree (DT) [36], naïve Bayes (NB) [45], and k-nearest neighbors (KNN) [45], are provided to UniFeat induced from the Weka software package [48]. Also, the "More option..." button is embedded in this panel to adjust the parameters of the classifiers.

4.1.5 Run configuration

The "Run configuration" panel is designed for two purposes. (1) While input datasets are divided randomly into the training and test sets, the division process should repeat several times. This process reduces the effect of the random nature of the dataset and improves the estimation of the performance of a feature selection method. (2) Some feature selection methods embed randomness into their search processes and thus provide stable results when they run several times independently.

Finally, you can click on the "Start" button to start the feature selection process. If the user provides all the requirements of the tool with form filling, the resulting interface will be shown. In this interface, some necessary information will be reported to the user. This includes some information about the dataset, weights of features, classification accuracies, execution times, and subsets of selected features in each iteration.

Figure 4.4 shows an example of the output results generated by the tool. From Figure 4.4, it is clear that two different subsets of features are selected, and the method has been run for six independent iterations. Note that each column corresponds to a specific subset of selected features.

Three buttons are embedded in the resulting interface, each of which is described in the following:

1. "View subsets" button: by clicking this button, you can see the different subsets of selected features obtained by a given feature selection method in each iteration. An example of this issue is shown in Figure 4.5.
2. "View training/test sets" button: by clicking this button, you can see the two different folders with the names CSV and ARFF. The reduced datasets based on different subsets of selected features in each iteration are saved in these folders. The ARFF folder represents the reduced dataset in the attribute-relation file format (i.e., ARFF, which is the standard format of the Weka software), and the CSV folder represents the reduced datasets in the comma-delimited format (i.e., CSV file format). Each file in these folders is saved with the format "name[i-j].format", where:
   - name: is the type of dataset with two different values: trainSet and testSet;
   - i: represents the ***i***-th iteration of the tool;
   - j: shows the number of selected features;

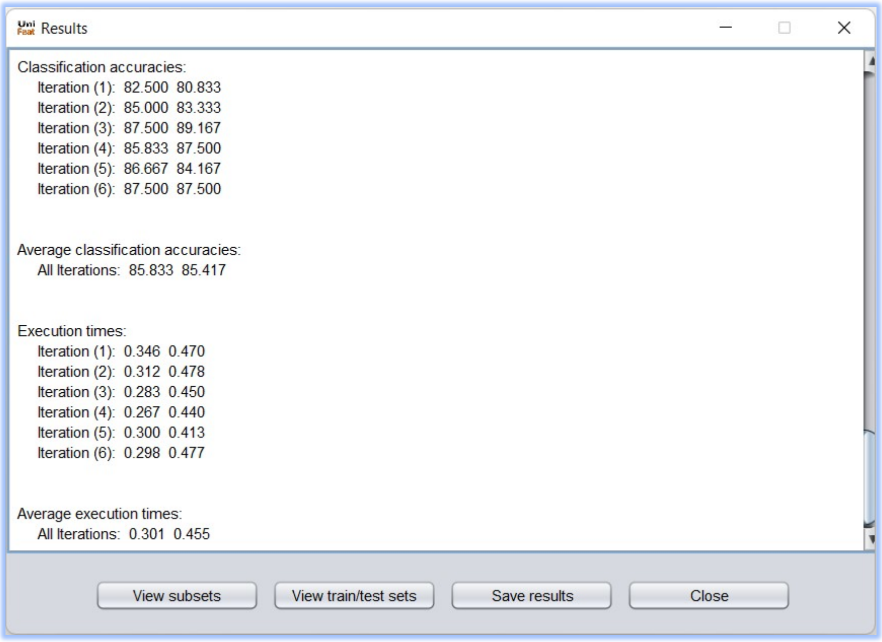
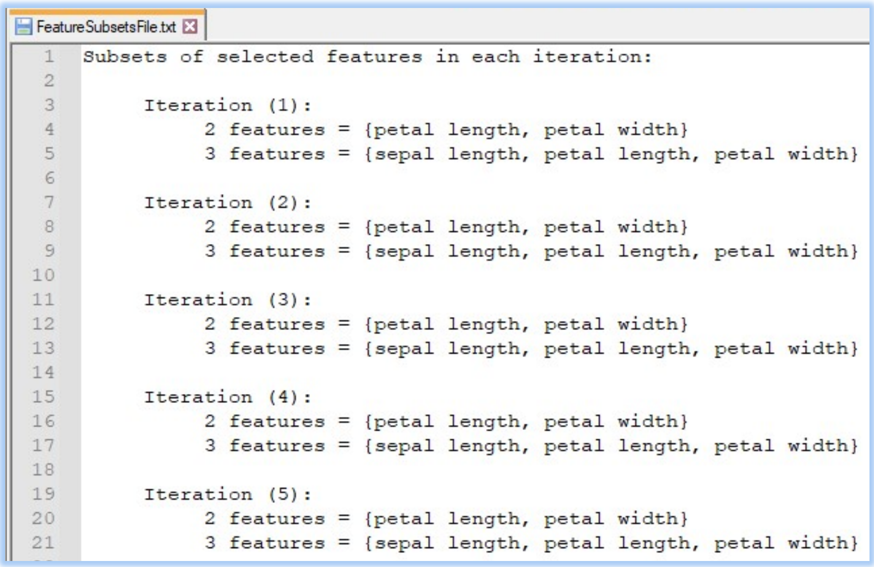


Figure 4.4: An example of the resulting interface.



FeatureSubsetsFile.txt

```
Subsets of selected features in each iteration:

    Iteration (1):
        2 features = {petal length, petal width}
        3 features = {sepal length, petal length, petal width}

    Iteration (2):
        2 features = {petal length, petal width}
        3 features = {sepal length, petal length, petal width}

    Iteration (3):
        2 features = {petal length, petal width}
        3 features = {sepal length, petal length, petal width}

    Iteration (4):
        2 features = {petal length, petal width}
        3 features = {sepal length, petal length, petal width}

    Iteration (5):
        2 features = {petal length, petal width}
        3 features = {sepal length, petal length, petal width}
```

Figure 4.5: An example of feature subsets file.

- format: illustrates the type of file format with two different values: arff and csv.

For example, the "testSet[1-5].arff" file shows the reduced test set file obtained by a given feature selection method from the first iteration based on five selected features with the ARFF format.
These reduced datasets can be easily used to compare fairly feature selection meth-

ods available in the UniFeat and any other feature selection methods implemented in the different software packages.

3. "Save results" button: You can save all information from the resulting interface as a text file by clicking this button.

4.2 Menu items description

Four different menu items correspond to the tasks of the UniFeat tool. Each of these menus is described in the following sections.

4.2.1 Preprocessing of the data

UniFeat supports only a specific dataset format described in Section 4.1.1; thus, a simple preprocessing panel is provided to help users import datasets from different sources and convert them into UniFeat format. Using the "File" → "Preprocess" menu in UniFeat, you can import a dataset and convert it to the correct format. The preprocessing panel is presented in Figure 4.6.

First, you should select a delimiter of the input dataset from the "Delimiter" panel, and then you can perform the two following optional operations over the dataset:

1. "Convert to Comma delimited": if you select this item, the current delimiter of the data is changed to the comma-delimited.
2. "Transpose (rotate) dataset from rows to columns or vice versa": some of the datasets have been presented so that the columns of the data show the samples, and the rows describe a vector of feature values corresponding to the samples. If you select this item, the dataset is rotated from rows to columns or vice versa.

4.2.2 Drawing a diagram

Visualizing the results obtained by UniFeat in the form of diagrams can help users obtain better interpretations. After the feature selection process is done, you can see three diagrams, including execution time, accuracy, and error rate, accessible through the "Diagram" menu. Moreover, the values of the results in each iteration and average values in all iterations can be reported in these diagrams. Figure 4.7 shows an example of the tool's execution time and classification accuracy diagrams. As shown in Figure 4.7, users can save diagrams in a *png* image format to facilitate reporting the results. This option is available in the "File" menu.

4.2.3 Analyzing the results

To show that the experimental results are statistically significant, the Friedman test [10] is currently provided in the UniFeat tool to analyze the results. The Friedman test is a non-parametric test used to measure the statistical differences of methods over multiple datasets. To apply this test, first of all, you should prepare a file as follows:

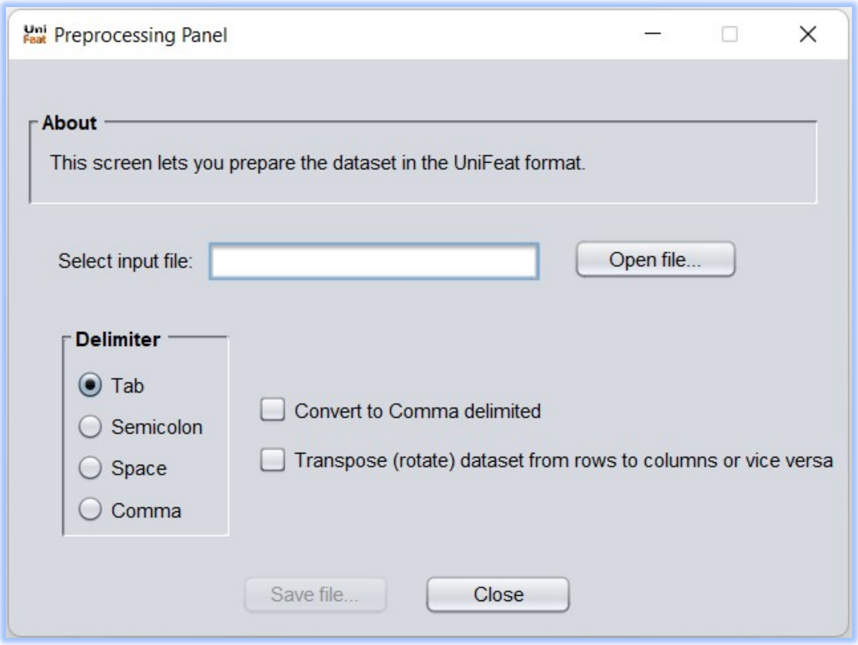
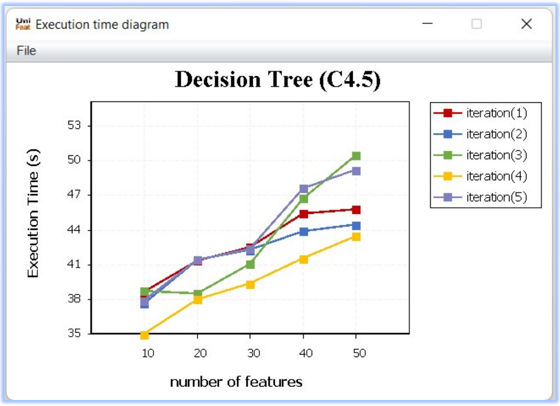
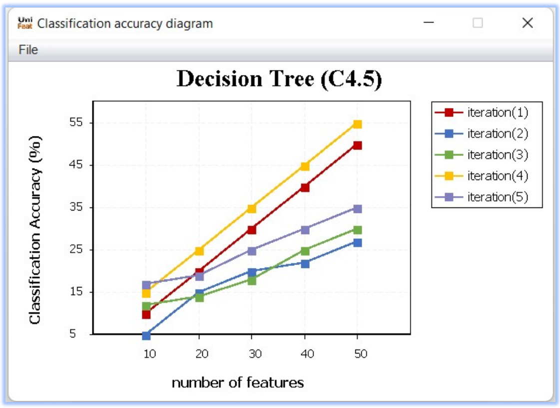


Figure 4.6: The preprocessing panel.



(a) Execution time diagram.



(b) Classification accuracy diagram.

Figure 4.7: Diagrams of UniFeat.

- The first row of the file must have the names of methods.
- The next rows contain all the values, with each row corresponding to the results of the methods on a dataset. Each row starts with the name of a dataset, and then the results of each method are presented, separated by commas.

You can easily import this file into the UniFeat tool as a CSV file. Figure 4.8 shows an example of this file in a spreadsheet. In Figure 4.8, the input file contains the classification error rates obtained by seven methods over five different datasets.

After preparing the results file, you can open the Friedman test panel, import the file,

	A	B	C	D	E	F	G	H
1		Method1	Method2	Method3	Method4	Method5	Method6	Method7
2	Dataset1	21.81	21.81	24.54	38.18	24.54	21.81	33.63
3	Dataset2	25.51	28.27	37.93	45.51	31.72	39.31	36.55
4	Dataset3	17.94	41.02	37.64	38.23	23.52	20.58	35.29
5	Dataset4	26.85	40.57	22.85	34.28	30.85	28	48
6	Dataset5	14.28	17.14	35.71	28.57	19.14	27.71	18

Figure 4.8: An example of the result values in a spreadsheet.

and perform the test on the input file using the "Analyze" → "Friedman test" menu in the UniFeat tool. The Friedman test panel is presented in Figure 4.9.

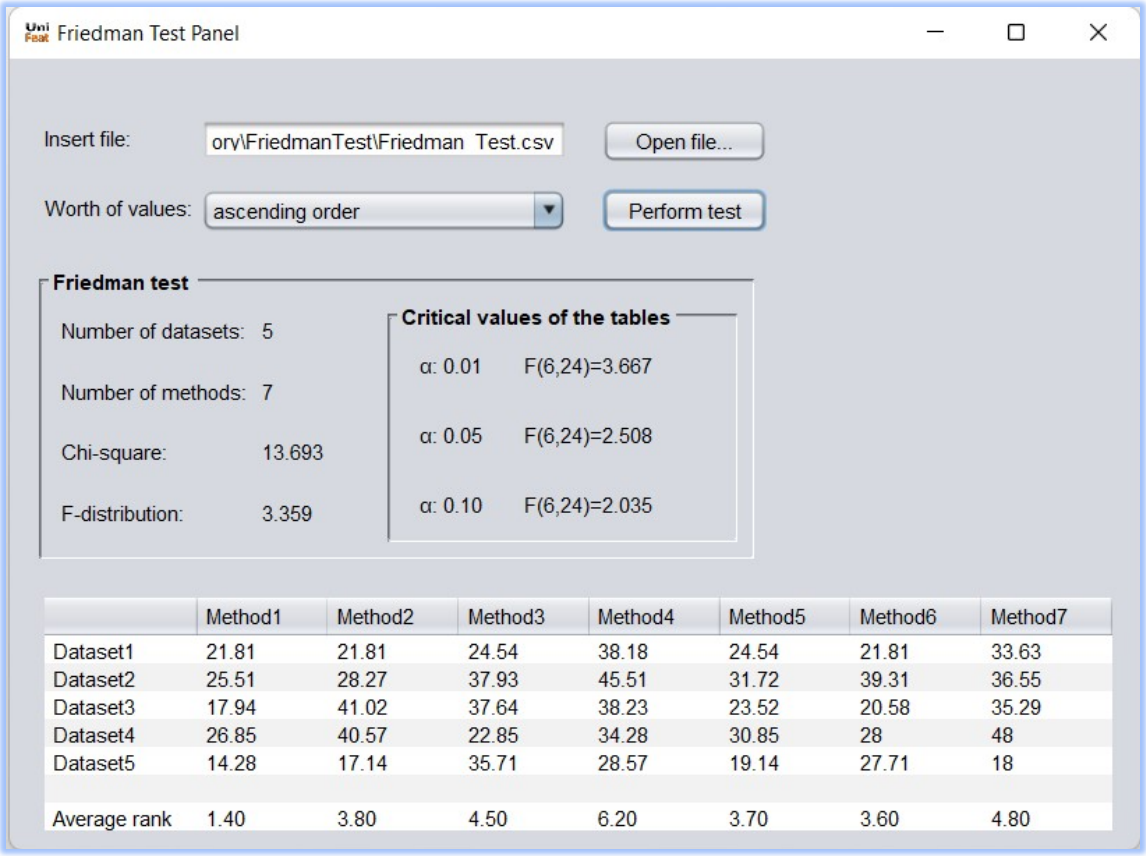


Figure 4.9: The Friedman test panel.

"Worth of values" in the Friedman test panel allows the user to select the worth of values in the file. If "ascending order" is specified, then the tool associates the best rank to the method with the lowest value; otherwise, in the case of "descending order," the tool associates the best rank to the method with the highest value.

The Friedman test panel provides some helpful information, such as the average values of each method over all datasets, Chi-square and F-distribution values, and critical values of the table based on various significant levels (i.e., α parameter).

4.2.4 Help file

By using the "Help" menu in UniFeat, you can access the tool's user manual.

5 Using UniFeat as a library

The easiest way to use UniFeat is through its graphical user interface. Sometimes you use feature selection as a part of your methods, and you prefer to embed the feature selection methods in your Java codes. Therefore, a question about the UniFeat tool has remained: "how to use the UniFeat tool as a jar file in your own Java codes?" In this section, we will describe this issue.

An example is presented in Appendix A to clarify how to read a dataset, call a feature selection method, and obtain the results from your own Java codes. The codes required to use the UniFeat as a jar file are explained in the following sections.

5.1 Reading dataset files

First, you should load all dataset files for future purposes. The input files must be prepared in the UniFeat format for the tool described in Section 4.1.1. If only one dataset file is available, you can easily import the codes in Figure 5.1a in your own Java code. In Figure 5.1a, *path1* is the dataset's path, and *path2* is the path of the class labels file. Both *path1* and *path2* are string values.

On the other hand, if the dataset file is originally divided into training and test sets, you can easily use the codes presented in Figure 5.1b. In Figure 5.1b, *path1* is the training set's path, *path2* is the test set's path, and *path3* is the path of the class labels file. *path1*, *path2*, and *path3* are string values.

```
import unifeat.dataset.DatasetInfo;
...
DatasetInfo data = new DatasetInfo();
data.preProcessing(path1,path2);
```

(a) One file of the dataset.

```
import unifeat.dataset.DatasetInfo;
...
DatasetInfo data = new DatasetInfo();
data.preProcessing(path1,path2,path3);
```

(b) Training and test files.

Figure 5.1: Source codes for reading the dataset.

5.2 Performing feature selection

After reading the dataset file, you can perform a given feature selection method based on the input dataset. The general interface of the feature selection methods currently available in the UniFeat tool can be considered in Figure 5.2.

From Figure 5.2 the following points deserve attention:

1. You can load the dataset for performing the feature selection process in two ways: (a) read the dataset as described in Section 5.1 and then use the "**loadDataSet(DatasetInfo ob)**" code, or (b) you can prepare the dataset as a matrix of double values without the names of features in the first row. Figure 5.3 shows the values of the dataset illustrated in Figure 4.2. In Figure 5.3, the data labels have

```java
public interface featureSelection {
    public void loadDataSet(DatasetInfo ob);
    public void loadDataSet(double[][] data, int numFeat, int numClasses);
    public void evaluateFeatures();
    public int[] getSelectedFeatureSubset();
    public double[] getFeatureValues();
    public String validate();
}
```

Figure 5.2: General interface of the feature selection methods.

been changed (i.e., *Iris-setosa* → 0, *Iris-versicolor* → 1, and *Iris-virginica* → 2). Then the "**loadDataSet(double[][] data, int numFeat, int numClasses)**" code is used where *numFeat* is the number of features and *numClasses* is the number of classes in the dataset.

2. The "**evaluateFeatures()**" function performs a given feature selection method over the input dataset.
3. The "**getSelectedFeatureSubset()**" function returns a subset of features selected by a given feature selection method.
4. The "**getFeatureValues()**" function is used to obtain the weights of features if the method gives weights of features individually and ranks them based on their relevance (i.e., feature weighting methods); otherwise, these values do not exist.
5. The "**validate()**" function is used to verify the validity of user input values. This method returns an empty string if all the input values are correct; otherwise, an error message is demonstrated to the user.

```java
double[][] data = { {4.3, 3, 1.1, 0.1, 0},
                    {4.4, 2.9, 1.4, 0.2, 0},
                    {4.4, 3, 1.3, 0.2, 0},
                    {4.4, 3.2, 1.3, 0.2, 0},
                    {5.5, 2.3, 4, 1.3, 1},
                    {5.5, 2.4, 3.8, 1.1, 1},
                    {5.5, 2.4, 3.7, 1, 1},
                    {5.5, 2.5, 4, 1.3, 1},
                    {6.3, 3.3, 6, 2.5, 2},
                    {6.3, 2.9, 5.6, 1.8, 2},
                    {6.3, 2.7, 4.9, 1.8, 2},
                    {6.3, 2.8, 5.1, 1.5, 2}};
```

Figure 5.3: An example of the data as a matrix.

For example, suppose we want to use information gain as a feature selection method. The required code is presented in Figure 5.4. In Figure 5.4, the *sizeSelectedFeatureSubset* parameter determines the number of features selected by the method, and the *data* is the input dataset obtained from Section 5.1. Also, the *message* keeps the possible error message from user input values, the *subset* supports the subset of features selected by information gain, and *computeValues* holds the information gain values of each feature.

```
import unifeat.featureSelection.filter.supervised.InformationGain;
...
    InformationGain method = new InformationGain(sizeSelectedFeatureSubset);
    method.loadDataSet(data);
    String message = method.validate();
    if (!message.isEmpty()) {
        System.out.print("Error!\n " + message);
    } else {
        method.evaluateFeatures();
        int[] subset = method.getSelectedFeatureSubset();
        double[] computeValues = method.getFeatureValues();
    }
```

Figure 5.4: Source codes for performing feature selection using information gain.

5.3 Creating reduced datasets

When the feature selection process has been done, you can create reduced datasets based on the subset of selected features in the CSV or ARFF formats.

If you want to create training and test sets in CSV or ARFF file formats based on the subset of selected features (i.e., the *subset* in Figure 5.4), you can easily embed the codes in Figure 5.5 in your own Java code. In Figure 5.5, *newPathTrainCSV* is a path for the training set in CSV format, *newPathTestCSV* is a path for the test set in CSV format, *newPathTrainARFF* is a path for the training set in ARFF format, *newPathTestARFF* is a path for the test set in ARFF format, and some temporary files will be created in *tempPath*. Also, *newPathTrainCSV*, *newPathTestCSV*, *newPathTrainARFF*, *newPathTestARFF*, and *tempPath* are string values. Furthermore, the *sizeSelectedFeatureSubset* parameter determines the number of features selected by the method. This code is used when the dataset files are loaded in the way explained in Figure 5.1.

```
import unifeat.dataset.DatasetInfo;
import unifeat.util.FileFunc;
...
    FileFunc.createCSVFile(data.getTrainSet(), subset, newPathTrainCSV,
            data.getNameFeatures(), data.getClassLabel());
    FileFunc.createCSVFile(data.getTestSet(), subset, newPathTestCSV,
            data.getNameFeatures(), data.getClassLabel());
    FileFunc.convertCSVtoARFF(newPathTrainCSV, newPathTrainARFF, tempPath,
            sizeSelectedFeatureSubset, data);
    FileFunc.convertCSVtoARFF(newPathTestCSV, newPathTestARFF, tempPath,
            sizeSelectedFeatureSubset, data);
```

Figure 5.5: Source codes for creating the CSV and ARFF files from the training and test sets based on Figure 5.1.

On the other hand, if the dataset is loaded in the form shown in Figure 5.3, and you want to create the CSV or ARFF files, you can easily embed the codes in Figure 5.6 in your own Java code. In Figure 5.6, *newPathDataCSV* is a path for the dataset in CSV format, *newPathDataARFF* is a path for the dataset in ARFF format, some temporary files will be created in *tempPath*, *FeatureNames* is an array of strings that presents the names of features, and *classNames* is an array of strings that presents the names of classes. Also, *numFeature* is the number of original features, and *numClass* is the number of classes in the dataset.

```
import unifeat.dataset.DatasetInfo;
import unifeat.util.FileFunc;
...
    FileFunc.createCSVFile(data, subset, newPathDataCSV, FeatureNames,
            classNames);
    FileFunc.convertCSVtoARFF(newPathDataCSV, newPathDataARFF, tempPath,
            sizeSelectedFeatureSubset, numFeature, FeatureNames, numClass,
            classNames);
```

Figure 5.6: Source codes for creating the CSV and ARFF files from the input array of the dataset.

It should be noted that you can directly use the Weka software [48] to create the ARFF files based on the created CSV files.

6 Extending UniFeat

The open-source nature and structure of UniFeat can help researchers use and modify the tool to fit their research requirements and facilitate it to share their methods with the scientific community rapidly. Therefore, another question remains about the UniFeat tool: “how can a new feature selection method be added to the tool?” In this section, we will answer this question in sufficient detail.

6.1 Adding a feature selection method to UniFeat

The unified structure of the feature selection package in UniFeat allows researchers to implement their feature selection methods via the UniFeat framework. Figure 6.1 shows the UML class diagram of the UniFeat feature selection approaches. Figure 6.1 reveals that all the feature selection approaches inherit the properties of the *FeatureSelection* abstract class. The functions provided by the *FeatureSelection* class were detailed in Section 5.2.

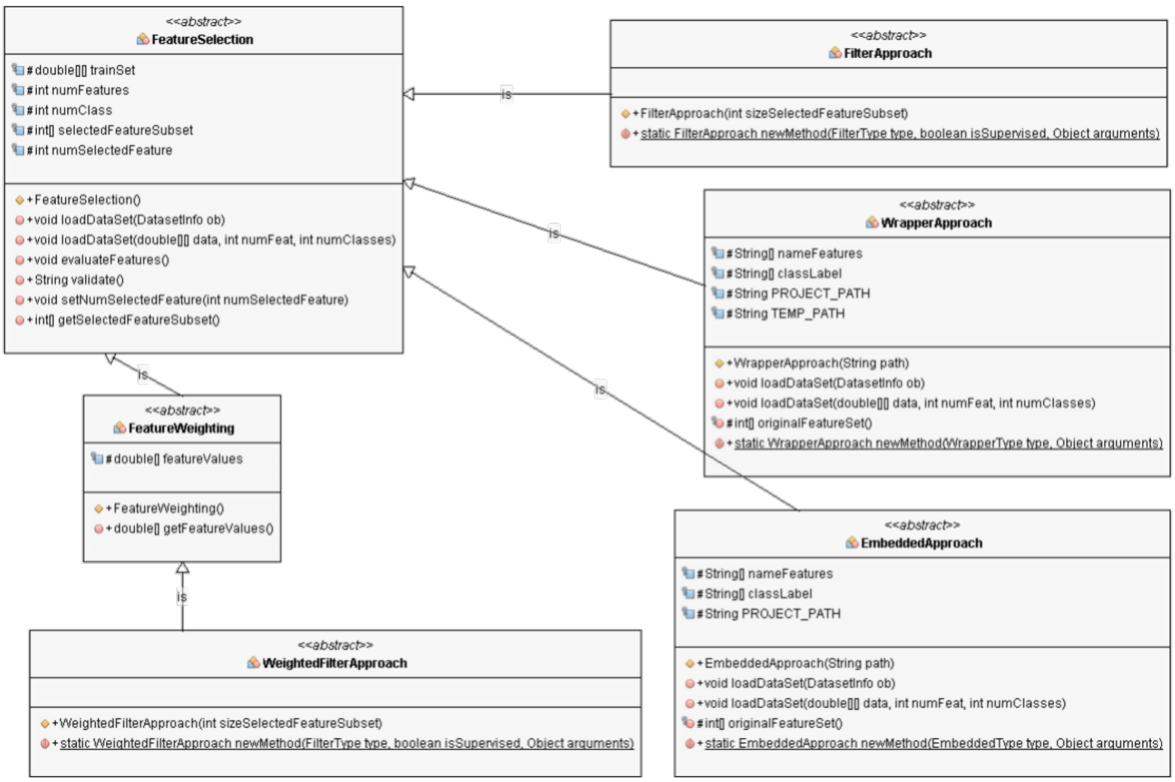


Figure 6.1: UML class diagram of feature selection approaches in UniFeat.

The current feature selection methods in the UniFeat repository are based on the filter, wrapper, and embedded approaches. We provide a separate class for each approach due to its specific requirements. Further details about how to add a new algorithm–considering its feature selection type–are provided in the corresponding sections.

6.1.1 Adding a new filter-based method

Filter-based methods are classified into two classes: the *WeightedFilterApproach* and *FilterApproach* classes, considered for feature weighting and feature subset selection methods, correspondingly. Feature weighting methods assign weights to features individually, rank them based on their relevance, and the top k features are finally returned to form the final feature set. Information gain [29], gain ratio [29], Gini index [39], and symmetrical uncertainty [26] are well-known methods in this category. On the other hand, feature subset selection methods choose a set of features without using any ranking criterion. RRFS [9], mRMR [33], RSM [23], and UFSACO [43] are examples of this category.

The general template class for adding a new filter-based method is presented in Figure 6.2, where the following points deserve attention.

```java
import unifeat.featureSelection.filter.WeightedFilterApproach;
import unifeat.util.ArraysFunc;

public class YourMethodName extends WeightedFilterApproach {

    public YourMethodName(Object... arguments) {
        super((int)arguments[0]);
    }

    public YourMethodName(int sizeSelectedFeatureSubset) {
        super(sizeSelectedFeatureSubset);
    }

    @Override
    public void evaluateFeatures() {
        // TODO feature selection process by your method
        ArraysFunc.sortArray1D(selectedFeatureSubset, false);
    }

    @Override
    public String validate() {
        // TODO validation of user input values
        return "keep this method to return an error message if"
                + " there are any errors in input parameters";
    }
}
```

Figure 6.2: General template class for adding a new filter-based method.

1. The class of your algorithm should extend one of the *WeightedFilterApproach* and *FilterApproach* abstract classes. These two abstract classes have similar functionalities, but *WeightedFilterApproach* returns a set of features associated with weights, while the other abstract class returns only a set of features.
2. Two constructor functions are provided for each filter method. The first function

has a variable argument *Object*. The second function takes the number and types of the arguments. If your method includes several tunable parameters, you should define the parameters as inputs to this function. The first value passed to both of these functions must be an integer that determines the number of features selected by your method.

3. The "**evaluateFeatures**()" function is used to implement the body of your algorithm. Note that this function does not have any input. It takes the required values from the fields provided in the *FeatureSelection* super-class, and these fields are initialized as the "**loadDataSet**()" functions are called. The body of your algorithm should store the final feature subset in the *selectedFeatureSubset* array. This array keeps the indices of the selected features, finally used as a result of your method's implementation. Moreover, the last line of your code is "**ArraysFunc.sortArray1D**()" invoked to sort the features based on their indices. The sorted array of feature indices is required to create the reduced dataset (see Section 5.3 for further details).
4. The "**validate**()" function is used to verify the validity of user input values. If your method does not include any parameter validation, you can remove this method. An implementation of this method is presented in the *FeatureSelection* super-class, where an empty message is returned to demonstrate that there is no error in the input values.

Note that you should add the class of your method into the supervised package if your method is a supervised algorithm; otherwise, you should add it into the unsupervised package.

In the GUI, users can choose feature selection methods. In UniFeat, a specific class for each approach provides a complete list of feature selection methods. All of these classes extend the *EnumType* class. Therefore, you should add the name of your filter-based method to the *FilterType* class, which is used for all filter-based feature selection methods.

6.1.2 Adding a new wrapper-based method

The general template class for adding a new wrapper-based method is similar to filter-based methods. However, the first argument in the constructor functions should be the project's path because some temporary files will be created in this path.

Since wrapper-based methods require a given classifier to evaluate different subsets of features during the feature selection process, four well-known and frequently-used classifiers are currently collected from the Weka software [48]. UniFeat uses the training/test evaluation and k-fold cross-validation [41] techniques to evaluate feature subsets. In training/test evaluation, a reduced dataset is first created based on the selected subset of features, then assessed by applying a classifier to the reduced dataset. Note that the reduced dataset is divided into training and test sets. The training set is used to build the classifier, and the test set is employed to evaluate the performance of the selected features.

Figure 6.3 shows the declarations of the current classifiers used for training/test evaluation. In these declarations, *pathTrainData* is the path of the training set in ARFF format, and *pathTestData* is the path of the test set in ARFF format. Other arguments in each function are needed for a specific classifier. All these functions are static, which means they can be invoked directly from the *TrainTestEvaluation* class.

```
public static Criteria SVM(String pathTrainData, String pathTestData,
                           SVMKernelType svmKernel, double c);
public static Criteria naiveBayes(String pathTrainData, String pathTestData);
public static Criteria dTree(String pathTrainData, String pathTestData,
                           double confidenceValue, int minNumSampleInLeaf);
public static Criteria kNN(String pathTrainData, String pathTestData,
                           int kNNValue);
```

Figure 6.3: Declarations of the classifiers used for training/test evaluation in UniFeat.

In k-fold cross-validation, the dataset is split into k parts. The first k-1 parts are applied in the training process to build a classifier. At the same time, the last one is utilized in the validation process to evaluate the performance of the selected subset. Figure 6.4 shows the declarations of the current classifiers employed in k-fold cross-validation. In Figure 6.4, *pathTrainData* is the path of the training set in ARFF format. Furthermore, *kFold* is an argument for defining the number of folds. All these functions are static, which means they can be invoked directly from the *CrossValidation* class.

```
public static Criteria SVM(String pathTrainData, SVMKernelType svmKernel,
                           double c, int kFold);
public static Criteria naiveBayes(String pathTrainData, int kFold);
public static Criteria dTree(String pathTrainData, double confidenceValue,
                           int minNumSampleInLeaf, int kFold);
public static Criteria kNN(String pathTrainData, int kNNValue, int kFold);
```

Figure 6.4: Declarations of the classifiers used for k-fold cross-validation in UniFeat.

Recently, population-based methods have attracted a lot of attention. Most of them belong to the wrapper approach. These methods consider the interaction between subsets of features, and they show higher performance than filter-based methods. The three most popular methods, including Genetic Algorithm (GA), Ant Colony Optimization (ACO), and Particle Swarm Optimization (PSO), are implemented in UniFeat. The simple implementation of these algorithms in UniFeat helps researchers easily have the structures inherited in their methods to develop. For example, Figure 6.5 provides the basic structure of GA implemented in UniFeat. In this structure, we have provided three abstract classes, including *BasicIndividual*, *BasicPopulation*, and *BasicGA*, which can be used in any GA-based feature selection method. The *BasicIndividual* class is employed to represent an individual, and the *BasicPopulation* class is used to create a population of individuals and apply genetic operations. Finally, *BasicGA* is the main class utilized for iteration of the algorithm an allowed number of times. Moreover, *BasicGA* class inherits the properties of the *WrapperArpproach* class. It is clear from Figure 6.5 that the essential genetic operators, including crossover, mutation, selection, and replacement, have been implemented in UniFeat. As frequently-used operators, they can be invoked from your method.

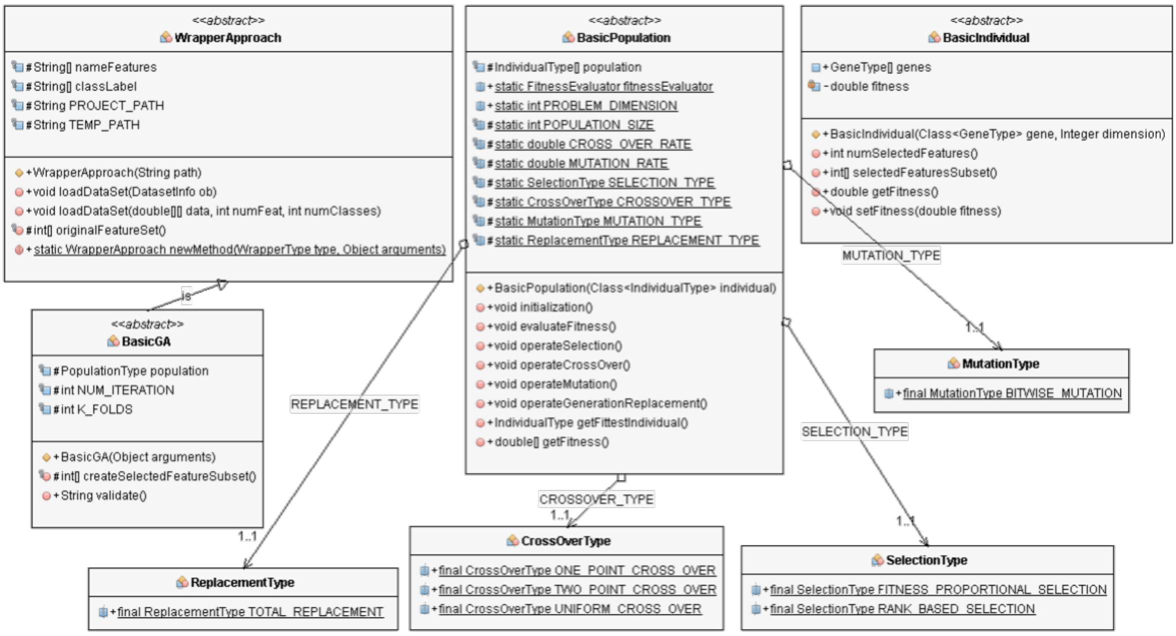


Figure 6.5: UML class diagram of the genetic algorithm in UniFeat.

6.1.3 Adding a new embedded-based method

The general template class for adding a new embedded method is similar to wrapper-based methods. In embedded-based methods, a given classifier is trained by an original feature set, and the obtained results are used to specify the relevance of each feature. Therefore, the functions shown in Figures 6.3 and 6.4 can be used in this approach.

SVM and DT are common classifiers for embedded-based methods implemented in UniFeat. For instance, Figure 6.6 shows the abstract classes of SVM-based methods. In Figure 6.6, there are two main functions "**buildSVM_OneAgainstOne**()" and "**buildSVM_OneAgainstRest**()". In the first function, there is a binary SVM for each pair of sample classes to separate the sample of one class from that of the other. In the second function, however, there is a binary SVM for each sample class to separate the sample of that class from that of the other classes.

6.2 Creating a parameter settings panel in UniFeat

Some feature selection methods have tunable parameters that need to be set. Further details on the parameters of different methods are included in Section 4.1.2. A simple structure has been designed for developers to create a GUI panel for setting these

Figure 6.6: UML class diagram of SVM in UniFeat.

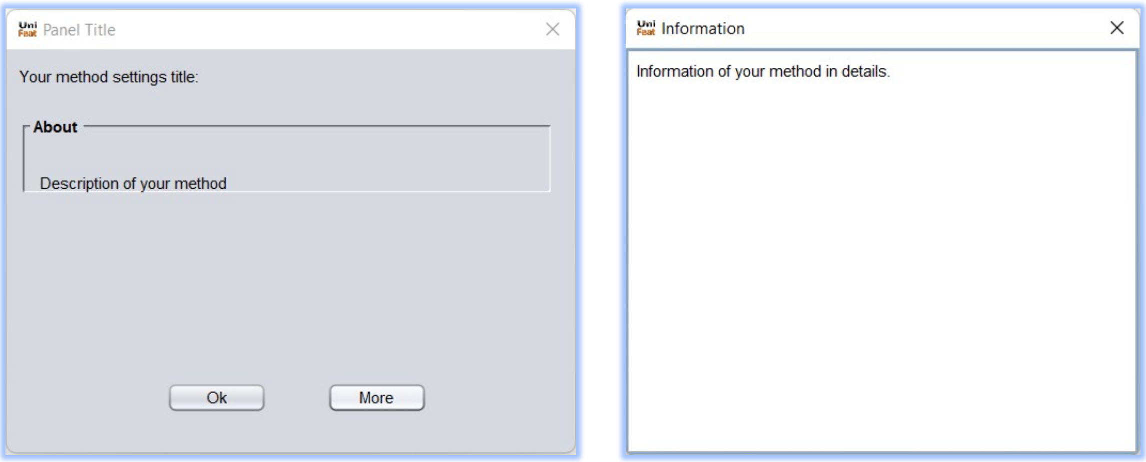
parameters in the UniFeat tool. Figure 6.7 shows a general template for creating a panel in UniFeat.

After running the code provided in Figure 6.7, you will see the general panel of parameter settings, which is illustrated in Figure 6.8a. You can change the "Panel Title," "Your method settings title," and "Description of your method" by calling the functions presented in the *ParameterPanel* super-class. Moreover, adding the desired components to the "**YourPanel**()" constructor function will be observed in the panel illustrated in Figure 6.8a. Furthermore, as seen in Figure 6.8a, if a user clicks on the "More" button, Figure 6.8b will be shown. Further information about the parameters is presented in this panel, and you can change the panel text by calling the relevant function in the *ParameterPanel* class.

After creating the panel, you should add the required code to the *MainPanel* class in the gui package. In this class, four important functions are presented: "**getFilterApproachParameters**()," "**getWeightedFilterApproachParameters**()," "**getWrapperApproachParameters**()," and "**getEmbeddedApproachParameters**()." They pass input parameters, which are obtained through GUI, to a specific method. These four functions are designed for different feature selection approaches.

```java
import unifeat.gui.ParameterPanel;
import java.awt.Dialog;
import java.awt.event.KeyEvent;
import javax.swing.UIManager;
import javax.swing.UnsupportedLookAndFeelException;

public class YourPanel extends ParameterPanel {

    public YourPanel(){
        super();
    }

    @Override
    public void keyReleased(KeyEvent e) {
        //TODO action when a key has been released
    }

    public static void main(String[] arg) {
        try {
            // Check if Nimbus is supported and get its classname
            for (UIManager.LookAndFeelInfo lafInfo :
                        UIManager.getInstalledLookAndFeels()) {
                if ("Nimbus".equals(lafInfo.getName())) {
                    UIManager.setLookAndFeel(lafInfo.getClassName());
                    UIManager.getDefaults().put("TextArea.font",
                                    UIManager.getFont("TextField.font"));
                    break;
                }
            }
        } catch (ClassNotFoundException |
                IllegalAccessException |
                InstantiationException |
                UnsupportedLookAndFeelException eOut) {
            try {
                // If Nimbus is not available, set to the system look and feel
                UIManager.setLookAndFeel(
                                UIManager.getSystemLookAndFeelClassName());
                UIManager.getDefaults().put("TextArea.font",
                                UIManager.getFont("TextField.font"));
            } catch (ClassNotFoundException |
                    InstantiationException |
                    IllegalAccessException |
                    UnsupportedLookAndFeelException eIn) {
                System.out.println("Error setting native LAF: " + eIn);
            }
        }

        YourPanel dtpanel = new YourPanel();
        Dialog dlg = new Dialog(dtpanel);
        dtpanel.setVisible(true);
    }
}
```

Figure 6.7: General template class for creating a GUI panel in UniFeat.



(a) Parameter settings panel. (b) Information about the parameters.

Figure 6.8: GUI panel in UniFeat.

References

[1] M. H. Aghdam, N. Ghasem-Aghaee, and M. E. Basiri. Text feature selection using ant colony optimization. *Expert Systems with Applications*, 36(3):6843–6853, 2009.

[2] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, J. M. Benítez, and F. Herrera. A review of microarray datasets and applied feature selection methods. *Information Sciences*, 282:111–135, 2014.

[3] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[4] G. Brown, A. Pocock, M.-J. Zhao, and M. Lujan. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *Journal of Machine Learning Research*, 13(1):27–66, 2012.

[5] T. Dokeroglu, A. Deniz, and H. E. Kiziloz. A comprehensive survey on recent metaheuristics for feature selection. *Neurocomputing*, 494:269–296, 2022.

[6] X. Du. Mlfeatureselection: General features selection based on certain machine learning algorithm and evaluation methods. Available at `https://github.com/duxuhao/Feature-Selection` (2022/09/24).

[7] K.-B. Duan, J. C. Rajapakse, and M. N. Nguyen. One-versus-one and one-versus-all multiclass svm-rfe for gene selection in cancer classification. In *Evolutionary Computation,Machine Learning and Data Mining in Bioinformatics*, pages 47–56. Springer Berlin Heidelberg, 2007.

[8] A. Dutt. Feature selection for machine learning. Available at `https://github.com/anujdutt9/Feature-Selection-for-Machine-Learning` (2022/09/24).

[9] A. J. Ferreira and M. A. T. Figueiredo. An unsupervised approach to feature discretization and selection. *Pattern Recognition*, 45(9):3048–3060, 2012.

[10] M. Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200): 675–701, 1937.

[11] S. Galli. Feature-engine: A python package for feature engineering for machine learning. *Journal of Open Source Software*, 6(65):3642, 2021.

[12] Q. Gu, Z. Li, and J. Han. Generalized fisher score for feature selection. In *International Conference on Uncertainty in Artificial Intelligence*, 2011.

[13] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[14] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002.

[15] M. Haindl, P. Somol, D. Ververidis, and C. Kotropoulos. *Feature Selection Based on Mutual Correlation*, volume 4225 of *Lecture Notes in Computer Science*, book section 59, pages 569–577. Springer Berlin Heidelberg, 2006.

[16] X. He, D. Cai, and P. Niyogi. Laplacian score for feature selection. *Advances in Neural Information Processing Systems*, 18, 2005.

[17] M. Hofmann and R. Klinkenberg. *RapidMiner: Data Mining Use Cases and Business Analytics Applications*. Chapman & Hall/CRC, 2013.

[18] F. T. Hussein. *Genetic algorithm for feature selection and weighting for off-line character recognition*. Thesis, University of British Columbia, 2002.

[19] M. M. Kabir, M. Shahjahan, and K. Murase. A new local search based hybrid genetic algorithm for feature selection. *Neurocomputing*, 74(17):2914–2928, 2011.

[20] D. Kai-Bo, J. C. Rajapakse, W. Haiying, and F. Azuaje. Multiple svm-rfe for gene selection in cancer classification with expression data. *IEEE Transactions on NanoBioscience*, 4(3):228–234, 2005.

[21] S. Khalid, T. Khalil, and S. Nasreen. A survey of feature selection and feature extraction techniques in machine learning. In *2014 science and information conference*, pages 372–378. IEEE, 2014.

[22] W. Koehrsen. Feature selector: Simple feature selection in python. Available at `https://github.com/WillKoehrsen/feature-selector` (2022/09/24).

[23] C. Lai, M. J. T. Reinders, and L. Wessels. Random subspace method for multivariate feature selection. *Pattern Recognition Letters*, 27(10):1067–1076, 2006.

[24] T. N. Lal, O. Chapelle, J. Weston, and A. Elisseeff. *Embedded Methods*, pages 137–165. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[25] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu. Feature selection: A data perspective. *ACM computing surveys (CSUR)*, 50(6):1–45, 2017.

[26] H. Liu and H. Motoda. *Computational Methods of Feature Selection*. Chapman & Hall/CRC, 2007.

[27] H. Liu and L. Yu. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, 17(4): 491–502, 2005.

[28] Y. Masoudi-Sobhanzadeh, H. Motieghader, and A. Masoudi-Nejad. Featureselect: a software for feature selection based on machine learning approaches. *BMC bioinformatics*, 20(1):1–17, 2019.

[29] T. M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., 1997.

[30] P. Moradi and M. Gholampour. A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy. *Applied Soft Computing*, 43:117–130, 2016.

[31] P. Moradi and M. Rostami. Integration of graph clustering with ant colony optimization for feature selection. *Knowledge-Based Systems*, 84:144–161, 2015.

[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[33] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.

[34] N. Pilnenskiy and I. Smetannikov. Feature selection algorithms as one of the python data analytical tools. *Future Internet*, 12(3):54, 2020.

[35] J. T. Pintas, L. A. F. Fernandes, and A. C. B. Garcia. Feature selection methods for text classification: a systematic literature review. *Artificial Intelligence Review*, 54(8):6149–6200, 2021.

[36] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[37] S. Raschka. Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack. *The Journal of Open Source Software*, 3(24), Apr. 2018.

[38] Y. Saeys, I. Inza, and P. Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.

[39] W. Shang, H. Huang, H. Zhu, Y. Lin, Y. Qu, and Z. Wang. A novel feature selection algorithm for text categorization. *Expert Systems with Applications*, 33(1):1–5, 2007.

[40] S. Solorio-Fernández, J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad. A review of unsupervised feature selection methods. *Artificial Intelligence Review*, 53(2): 907–948, 2020.

[41] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):111–147, 1974.

[42] S. Tabakhi and P. Moradi. Relevance–redundancy feature selection based on ant colony optimization. *Pattern Recognition*, 48(9):2798–2811, 2015.

[43] S. Tabakhi, P. Moradi, and F. Akhlaghian. An unsupervised feature selection algorithm based on ant colony optimization. *Engineering Applications of Artificial Intelligence*, 32(0):112–123, 2014.

[44] S. Tabakhi, A. Najafi, R. Ranjbar, and P. Moradi. Gene selection for microarray data classification using a novel ant colony optimization. *Neurocomputing*, 168: 1024–1036, 2015.

[45] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Elsevier Science, 2008.

[46] J. Too. Jx-wfst : A wrapper feature selection toolbox. Available at https://github.com/JingweiToo/Wrapper-Feature-Selection-Toolbox (2022/09/24).

[47] G. Tsoumakas, E. Spyromitros-Xioufis, J. Vilcek, and I. Vlahavas. Mulan: A java library for multi-label learning. *Journal of Machine Learning Research*, 12:2411–2414, 2011.

[48] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, 4 edition, 2016.

[49] B. Xue. *Particle Swarm Optimisation for Feature Selection in Classification*. Thesis, Victoria University of Wellington, 2014.

[50] B. Xue, M. Zhang, and W. N. Browne. Particle swarm optimisation for feature selection in classification: Novel initialisation and updating mechanisms. *Applied Soft Computing*, 18:261–276, 2014.

[51] K. Yu, W. Ding, and X. Wu. Lofs: A library of online streaming feature selection. *Knowledge-Based Systems*, 113:1–3, 2016.

A An example source code for feature selection

A simple Java program that performs feature selection using the information gain method implemented in the UniFeat tool and displays the results is presented below:

```java
import unifeat.dataset.DatasetInfo;
import unifeat.featureSelection.filter.supervised.InformationGain;
import unifeat.util.FileFunc;

public class Main {
    public static void main(String[] args) {
        //reading the datasets files
        DatasetInfo data = new DatasetInfo();
        data.preProcessing("data/trainSet.csv", "data/testSet.csv", "data/classLabels.txt");

        //printing some information of the dataset
        int sizeSelectedFeatureSubset = 2;
        System.out.println(" no. of all samples : " + data.getNumData()
                + "\n no. of samples in training set :  " + data.getNumTrainSet()
                + "\n no .of samples in test set : " + data.getNumTestSet()
                + "\n no. of features : " + data.getNumFeature()
                + "\n no. of classes : " + data.getNumClass());

        //performing the feature selection by information gain method
        InformationGain method = new InformationGain(sizeSelectedFeatureSubset);
        method.loadDataSet(data);

        String message = method.validate();
        //checking the validity of user input values
        if (!message.isEmpty()) {
            System.out.print("Error!\n  " + message);
        } else {
            method.evaluateFeatures();
            int[] subset = method.getSelectedFeatureSubset();
            double[] infoGainValues = method.getFeatureValues();

            //printing the subset of selected features
            System.out.print("\n subset of selected features: ");
            for (int i = 0; i < subset.length; i++) {
                System.out.print((subset[i] + 1) + "  ");
            }

            //printing the information gain values
            System.out.println("\n\n information gain values: ");
            for (int i = 0; i < infoGainValues.length; i++) {
                System.out.println(" " + (i + 1) + " : " + infoGainValues[i]);
            }

            //creating reduced datasets as the CSV file format
            FileFunc.createCSVFile(data.getTrainSet(), subset, "data/newTrainSet.csv", data.getNameFeatures(), data.getClassLabel());
            FileFunc.createCSVFile(data.getTestSet(), subset, "data/newTestSet.csv", data.getNameFeatures(), data.getClassLabel());

            //creating reduced datasets as the ARFF file format
            FileFunc.convertCSVtoARFF("data/newTrainSet.csv", "data/newTrainSet.arff", "data", sizeSelectedFeatureSubset, data);
            FileFunc.convertCSVtoARFF("data/newTestSet.csv", "data/newTestSet.arff", "data", sizeSelectedFeatureSubset, data);
        }
    }
}
```