



Wprowadzenie
do WEKA

Adam Zagdański,
Artur Suchwałko
(www.suchwalko.pl)

Czym jest WEKA?

Moduły dostępne
w WEKA

Moduł Explorer

Moduł *Knowledge
Flow*

WEKA – informacje
techniczne

Dodatkowe
informacje

Wprowadzenie do WEKA

Adam Zagdański, Artur Suchwałko

5 marca 2011



Plan prezentacji I

Część:

Wprowadzenie
do WEKA

Adam Zagdański,
Artur Suchwałko
(www.suchwalko.pl)

Czym jest WEKA?

Moduły dostępne
w WEKA

Moduł Explorer

Moduł *Knowledge
Flow*

WEKA – informacje
techniczne

Dodatkowe
informacje

- 1 Czym jest WEKA?
 - Główne cechy projektu
 - Wersje oprogramowania
- 2 Moduły dostępne w WEKA
- 3 Moduł Explorer
 - Preprocessing
 - Wizualizacja
 - Klasyfikacja
 - Analiza skupień
- 4 Moduł *Knowledge Flow*
- 5 WEKA – informacje techniczne
 - Wymagane oprogramowanie
 - Format danych
- 6 Dodatkowe informacje
 - Użyteczne linki
 - Wybrane projekty stworzone na bazie WEKA
 - Książka – DM z wykorzystaniem WEKA
 - WEKA – dokumentacja techniczna



Czym jest WEKA?

Weka Machine Learning Project, The University of Waikato, New Zealand

<http://www.cs.waikato.ac.nz/ml/weka/>

Część:

Wprowadzenie
do WEKA

Adam Zagdański,
Artur Suchwałko
(www.suchwalko.pl)

Czym jest WEKA?

Moduły dostępne
w WEKA

Moduł Explorer

Moduł *Knowledge
Flow*

WEKA – informacje
techniczne

Dodatkowe
informacje

Weka: Oprogramowanie z zakresu uczenia maszynowego (*machine learning*) i pozyskiwania wiedzy (*data mining*), stworzone w języku Java,

Weka: Zestaw algorytmów wykorzystywanych do realizacji zadań data miningu,

Weka: Oprogramowanie wykorzystywane w badaniach naukowych, edukacji, a także do zastosowań praktycznych,

Weka: Narzędzia do obróbki wstępnej danych (*pre-processing*), klasyfikacji, regresji, analizy skupień, odkrywania reguł asocjacyjnych i wizualizacji,

Weka: Oprogramowanie towarzyszące książce „*Data Mining: Practical Machine Learning Tools and Techniques*” autorstwa I.H. Witten i E. Franka,



Czym jest WEKA?

Weka Machine Learning Project, The University of Waikato, New Zealand

<http://www.cs.waikato.ac.nz/ml/weka/>

Część:

Wprowadzenie
do WEKA

Adam Zagdański,
Artur Suchwałko
(www.suchwalcko.pl)

Czym jest WEKA?

Moduły dostępne
w WEKA

Moduł Explorer

Moduł *Knowledge
Flow*

WEKA – informacje
techniczne

Dodatkowe
informacje

Weka: Wygodna baza dla rozwijania nowych algorytmów uczenia maszynowego,

Weka: Algorytmy, które mogą być stosowane z wykorzystaniem dostępnych graficznych interfejsów użytkownika lub wywoływane z poziomu własnego kodu/aplikacji napisanej w języku Java, Możliwe jest wykorzystanie klas WEKA w innych programach (np. w środowisku R lub RapidMiner)

Weka: Oprogramowanie typu *open source* udostępnione na licencji GNU General Public License,

Weka: To także... ptak nietop, zagrożony wyginięciem, występujący wyłącznie na terenie Nowej Zelandii.



Czym jest WEKA?

Główne cechy projektu

Część: Główne cechy projektu

Wprowadzenie
do WEKA

Adam Zagdański,
Artur Suchwałko
(www.suchwalko.pl)

Czym jest WEKA?

Moduły dostępne
w WEKA

Moduł Explorer

Moduł *Knowledge
Flow*

WEKA – informacje
techniczne

Dodatkowe
informacje

- Obszerny zestaw narzędzi do przetwarzania wstępnego danych (*pre-processing'u*),
- Algorytmy uczenia maszynowego i metody oceniające ich efektywność,
- Przyjazne graficzne interfejsy użytkownika (w tym, narzędzia do wizualizacji danych),
- Wygodne środowisko do porównania efektywności algorytmów.



Czym jest WEKA?

Wersje oprogramowania I

Część: Wersje oprogramowania

Wprowadzenie
do WEKA

Adam Zagdański,
Artur Suchwałko
(www.suchwalcko.pl)

Czym jest WEKA?

Moduły dostępne
w WEKA

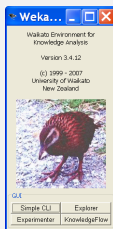
Moduł Explorer

Moduł *Knowledge
Flow*

WEKA – informacje
techniczne

Dodatkowe
informacje

- **book version** – wersja towarzysząca książce „*Data Mining: Practical Machine Learning Tools and Techniques*” autorstwa I.H. Wittena i E. Franka. Wersja została „zamrożona” w 2005 wraz z publikacją książki i nie pojawiają się już dla niej nowe funkcjonalności (np. nowe algorytmy, itd.), a jedynie korygowane są dostrzeżone błędy,



Rysunek: v.3.4.12 (book)



Czym jest WEKA? Wersje oprogramowania II

Część: Wersje oprogramowania

Wprowadzenie
do WEKA

Adam Zagdański,
Artur Suchwałko
(www.suchwalko.pl)

Czym jest WEKA?

Moduły dostępne
w WEKA

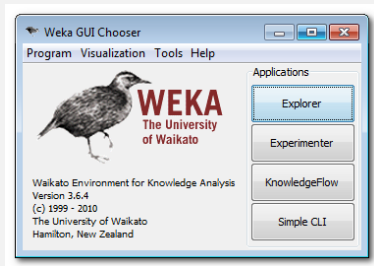
Moduł Explorer

Moduł *Knowledge
Flow*

WEKA – informacje
techniczne

Dodatkowe
informacje

- **stable version** – aktualna wersja stabilna



Rysunek: v.3.6.4 (stable)

- **developer version** – wersja aktualnie rozwijana, uzupełniana o nowe algorytmy, usprawnienia, itp. (v.3.7.3)



Cztery główne moduły dostępne w WEKA

Część:

Wprowadzenie
do WEKA

Adam Zagdański,
Artur Suchwałko
(www.suchwalko.pl)

Czym jest WEKA?

Moduły dostępne
w WEKA

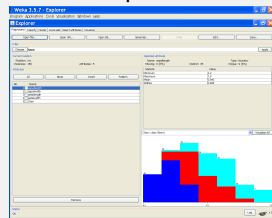
Moduł Explorer

Moduł Knowledge
Flow

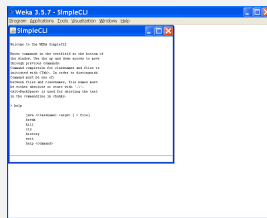
WEKA – informacje
techniczne

Dodatkowe
informacje

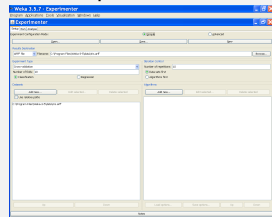
Explorer



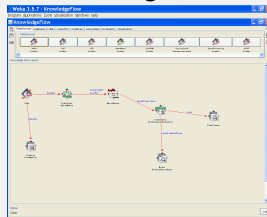
CLI



Experimenter



Knowledge Flow





Cztery główne moduły dostępne w WEKA I

Część:

Wprowadzenie
do WEKA

Adam Zagdański,
Artur Suchwałko
(www.suchwalko.pl)

Czym jest WEKA?

Moduły dostępne
w WEKA

Moduł Explorer

Moduł *Knowledge
Flow*

WEKA – informacje
techniczne

Dodatkowe
informacje

- 1 **Explorer** – główny moduł oferujący dostęp do najważniejszych funkcjonalności. Szereg rozwiązań ułatwiających użytkownikowi przeprowadzenie analiz (łatwa konfiguracja parametrów, kontrola kolejności wykonywania analiz, parametry domyślne, podpowiedzi kontekstowe). Zalecany na początek i dla większości użytkowników w zupełności wystarczający!
- 2 **CLI** – interfejs tekstowy (Command Line Interface). Dostęp do funkcjonalności systemu poprzez wpisywanie komend tekstowych. Zalecany dla doświadczonych użytkowników!



Cztery główne moduły dostępne w WEKA II

Część:

Wprowadzenie
do WEKA

Adam Zagdański,
Artur Suchwałko
(www.suchwalko.pl)

Czym jest WEKA?

Moduły dostępne
w WEKA

Moduł Explorer

Moduł *Knowledge
Flow*

WEKA – informacje
techniczne

Dodatkowe
informacje

- 3 Experimenter** – zaprojektowany aby umożliwić przeanalizowanie, która metoda (np. klasyfikacji lub regresji) i jaki zestaw parametrów jest najlepszy dla naszego problemu. W module zastosowano rozwiązania umożliwiające przeprowadzania złożonych eksperymentów obliczeniowych na wielką skalę. Zaawansowani użytkownicy mają np. możliwość przeprowadzania obliczeń rozproszonych (na wielu komputerach równocześnie), dzięki wykorzystaniu technologii Java RMI (Remote Method Invocation),
- 4 Knowledge Flow** – interfejs graficzny, pozwalający zaprojektować schemat potokowego przetwarzania danych. Wykorzystując technikę „przeciągnij i upuść” możemy łatwo łączyć bloki reprezentujące poszczególne etapy analizy.



Moduł Explorer

Główny panel

Część:

Wprowadzenie
do WEKA

Adam Zagdański,
Artur Suchwałko
(www.suchwalko.pl)

Czym jest WEKA?

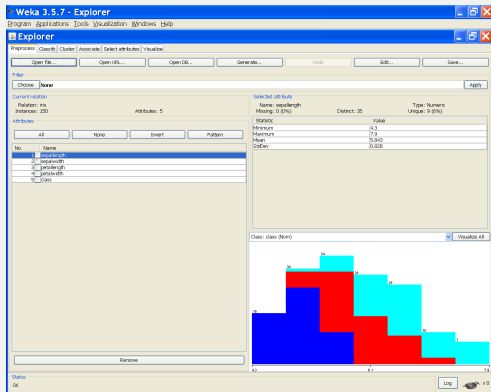
Moduły dostępne
w WEKA

Moduł Explorer

Moduł Knowledge
Flow

WEKA – informacje
techniczne

Dodatkowe
informacje



Rysunek: Moduł Explorer



Moduł Explorer

Zakładki

Część:

Wprowadzenie
do WEKA

Adam Zagdański,
Artur Suchwałko
(www.suchwalcko.pl)

Czym jest WEKA?

Moduły dostępne
w WEKA

Moduł Explorer

Moduł Knowledge
Flow

WEKA – informacje
techniczne

Dodatkowe
informacje

- 1 **Preprocess** – Wczytanie i obróbka wstępna danych,
- 2 **Classify** – konstrukcja prognoz z wykorzystaniem metod klasyfikacji i regresji; uczenie metody i weryfikacja jej efektywności,
- 3 **Cluster** – grupowanie obiektów (analiza skupień),
- 4 **Associate** – odkrywanie reguł asocjacyjnych,
- 5 **Select attributes** – wybór najważniejszych/najbardziej istotnych atrybutów (cech),
- 6 **Visualize** – wizualizacja danych w 2D (z elementami interaktywnymi).



Moduł Explorer Edytor danych

Część: Preprocessing

Wprowadzenie
do WEKA

Adam Zagdański,
Artur Suchwałko
(www.suchwalcko.pl)

Czym jest WEKA?

Moduły dostępne
w WEKA

Moduł Explorer

Moduł Knowledge
Flow

WEKA – informacje
techniczne

Dodatkowe
informacje

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose None Apply

Current relation: Relation: iris Instances: 150 Attributes: 5

Selected attribute: Name: sepal.length Missing: 0 (0%) Distinct: 35 Type: Numeric Unique: 9 (6%)

Viewer

Relation: iris

No.	sepal.length Numeric	sepal.width Numeric	petal.length Numeric	petal.width Numeric	class Nominal
60	5.2	2.7	3.9	1.4	iris-versicolr
135	6.1	2.6	5.6	1.4	iris-virginica
87	6.7	3.1	4.7	1.5	iris-versicolr
67	5.6	3.0	4.5	1.5	iris-versicolr
73	6.3	2.5	4.9	1.5	iris-versicolr
53	6.9	3.1	4.9	1.5	iris-versicolr
55	6.5	2.8	4.6	1.5	iris-versicolr
69	6.2	2.2	4.5	1.5	iris-versicolr
52	6.4	3.2	4.5	1.5	iris-versicolr
79	6.0	2.9	4.5	1.5	iris-versicolr
62	5.9	3.0	4.2	1.5	iris-versicolr
85	5.4	3.0	4.5	1.5	iris-versicolr
120	6.0	2.2	5.0	1.5	iris-virginica
134	6.3	2.8	5.1	1.5	iris-virginica
87	6.3	3.3	4.7	1.6	iris-versicolr
130	7.2	3.0	5.8	1.6	iris-versicolr
84	6.0	2.7	5.1	1.6	iris-versicolr
86	6.0	3.4	4.5	1.6	iris-versicolr
107	4.9	2.5	4.5	1.7	iris-virginica
78	6.7	3.0	5.0	1.7	iris-versicolr
108	7.3	2.9	6.3	1.8	iris-virginica
109	6.7	2.5	5.8	1.8	iris-virginica
117	6.5	3.0	5.5	1.8	iris-virginica
71	5.9	3.2	4.8	1.8	iris-versicolr
124	6.3	2.7	4.9	1.8	iris-virginica
126	7.2	3.2	6.0	1.8	iris-virginica
127	6.2	2.8	4.8	1.8	iris-virginica
128	6.1	3.0	4.9	1.8	iris-virginica
104	6.3	2.9	5.6	1.8	iris-virginica
138	6.4	3.1	5.5	1.8	iris-virginica
139	6.0	3.0	4.8	1.8	iris-virginica
150	5.9	3.0	5.1	1.8	iris-virginica

Undo OK Cancel



Moduł Explorer

Preprocessing – filtry

Część: Preprocessing

Wprowadzenie
do WEKA

Adam Zagdański,
Artur Suchwałko
(www.suchwalcko.pl)

Czym jest WEKA?

Moduły dostępne
w WEKA

Moduł Explorer

Moduł Knowledge
Flow

WEKA – informacje
techniczne

Dodatkowe
informacje

- Narzędzia obróbki wstępnej (pre-processing'u) w programie WEKA są nazywane **filtrami**
- W tej grupie znajdują się m.in. metody pozwalające przeprowadzić:
 - dyskretyzację (przedziałowanie) cech,
 - standaryzację (normalizację) danych,
 - próbkowanie,
 - wybór atrybutów,
 - transformacje i łączenie atrybutów,
 - wyznaczenie składowych głównych (metoda PCA),
- Podział filtrów:
 - **unsupervised** – nienadzorowane,
 - **supervised** – nadzorowane,
- Dla obu kategorii (*unsupervised* i *supervised*) wyróżnia się filtry stosowane dla:
 - atrybutów/cech (*attribute*),
 - przypadków (*instance*).



Moduł Explorer

Preprocessing – przygotowanie danych do analiz

Część: Preprocessing

Wprowadzenie
do WEKA

Adam Zagdański,
Artur Suchwałko
(www.suchwalcko.pl)

Czym jest WEKA?

Moduły dostępne
w WEKA

Moduł Explorer

Moduł Knowledge
Flow

WEKA – informacje
techniczne

Dodatkowe
informacje

The screenshot shows the WEKA Explorer interface. The main window is titled 'Weka 3.5.7 - Explorer'. The 'Preprocessing' tab is active, and the 'weka.filters.unsupervised.attribute.Discretize' filter is applied to the 'sepal.length' attribute. The filter's parameters are shown in a dialog box, including 'binSize' (5), 'ignoreClass' (True), and 'useEqualFrequency' (False). The 'Class: class (Item)' section shows a histogram of the data distribution, with bars colored in blue, red, and cyan. The histogram shows the distribution of sepal lengths across different classes, with the highest frequency in the red class.

Label	Count
1 (red)	32
2 (blue)	41
3 (cyan)	42
4 (red)	24
5 (cyan)	11

Rysunek: Przykład – przekształcenia wstępne
(dyskretyzacja cech ciągłych)



Moduł Explorer

Preprocessing – przygotowanie danych do analiz

Część: Preprocessing

Wprowadzenie
do WEKA

Adam Zagdański,
Artur Suchwałko
(www.suchwalko.pl)

Czym jest WEKA?

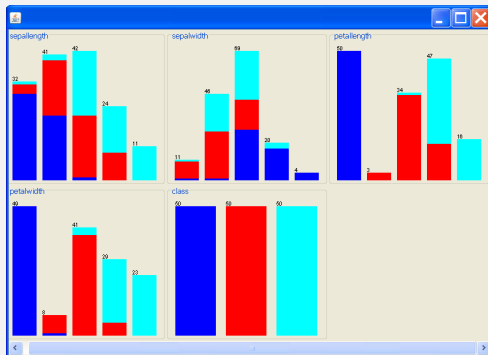
Moduły dostępne
w WEKA

Moduł Explorer

Moduł *Knowledge
Flow*

WEKA – informacje
techniczne

Dodatkowe
informacje



Rysunek: Przykład – skategoryzowane wykresy słupkowe dla wszystkich atrybutów (opcja: *Visualize All*)



Moduł Explorer Wizualizacja

Część: Wizualizacja

Wprowadzenie
do WEKA

Adam Zagdański,
Artur Suchwałko
(www.suchwalcko.pl)

Czym jest WEKA?

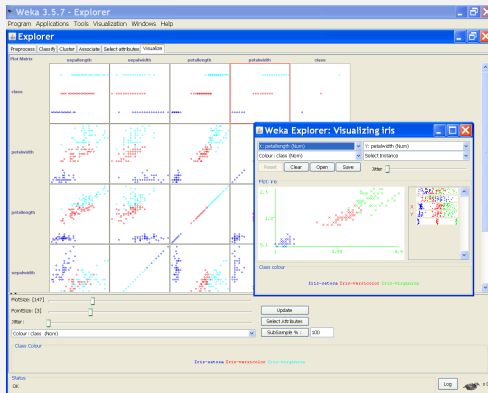
Moduły dostępne
w WEKA

Moduł Explorer

Moduł Knowledge
Flow

WEKA – informacje
techniczne

Dodatkowe
informacje



Rysunek: Przykład – wykresy rozrzutu (scatterplots)



Moduł Explorer

Klasyfikacja

Część: Klasyfikacja

Wprowadzenie
do WEKA

Adam Zagdański,
Artur Suchwałko
(www.suchwalcko.pl)

Czym jest WEKA?

Moduły dostępne
w WEKA

Moduł Explorer

Moduł Knowledge
Flow

WEKA – informacje
techniczne

Dodatkowe
informacje

- Klasyfikatorami w programie WEKA są nazywane modele pozwalające na prognozowanie zmiennych nominalnych (etykiety klas) lub liczbowych (np. modele regresyjne)
- Wybrane, zaimplementowane w WEKA algorytmy klasyfikacji
 - drzewa decyzyjne,
 - lasy losowe (*random forest*),
 - metody najbliższego sąsiada (*instance-based classifiers*),
 - Support Vector Machines (SVM),
 - sieci neuronowe wielowarstwowe,
 - regresja liniowa i logistyczna,
 - naiwny klasyfikator bayesowski,
 - sieci bayesowskie,
 - *Meta-classifiers* – klasyfikatory złożone (zaagregowane),
 - *UserClassifier* – klasyfikator (w formie drzewa decyzyjnego) budowany interaktywnie przez użytkownika,
 - *ZeroR* – klasyfikator referencyjny, prognozowana jest najczęstsza klasa lub wartość średnia (w przypadku prognoz ilościowych),
 - wiele innych...



Moduł Explorer

Klasyfikacja – podział metod

Część: Klasyfikacja

Wprowadzenie
do WEKA

Adam Zagdański,
Artur Suchwałko
(www.suchwalko.pl)

Czym jest WEKA?

Moduły dostępne
w WEKA

Moduł Explorer

Moduł *Knowledge
Flow*

WEKA – informacje
techniczne

Dodatkowe
informacje

W WEKA wyróżnia się następujący podział algorytmów klasyfikacji:

- **bayes**
- **functions**
- **lazy**
- **meta**
- **mi**
- **misc**
- **trees**
- **rules**



Moduł Explorer I

Klasyfikacja – podział metod

Część: Klasyfikacja

Wprowadzenie
do WEKA

Adam Zagdański,
Artur Suchwałko
(www.suchwalko.pl)

Czym jest WEKA?

Moduły dostępne
w WEKA

Moduł Explorer

Moduł Knowledge
Flow

WEKA – informacje
techniczne

Dodatkowe
informacje

- **bayes** – klasyfikatory bayesowskie (m.in.: sieci bayesowskie i naiwny klasyfikator bayesowski),
- **functions** – klasyfikatory, które w naturalny sposób można przedstawić jako równania matematyczne, m.in.: regresja liniowa i logistyczna, sieci neuronowe, SVM. Wyjątkiem jest np. naiwny klasyfikator bayesowski, który należy do osobnej grupy,
- **lazy** – klasyfikatory, które przechowują przypadki ze zbioru uczącego i nie wykonują żadnych obliczeń, aż do momentu klasyfikacji nowych obiektów (m.in.: różne warianty metody najbliższego sąsiada, ale także metoda *LBR – Lazy Bayesian Rules*),



Moduł Explorer II

Klasyfikacja – podział metod

Część: Klasyfikacja

Wprowadzenie
do WEKA

Adam Zagdański,
Artur Suchwałko
(www.suchwalcko.pl)

Czym jest WEKA?

Moduły dostępne
w WEKA

Moduł Explorer

Moduł Knowledge
Flow

WEKA – informacje
techniczne

Dodatkowe
informacje

- **meta** – *meta-classifiers*, klasyfikatory złożone, poprawiające efektywność klasyfikatorów bazowych:
 - różne warianty komitetów/rodzin klasyfikatorów (np. bagging, boosting),
 - **CostSensitiveClassifier** – modyfikacja wag przypadków zgodnie z kryterium kosztu przypisanym każdej z klas lub prognozowanie tej klasy, której odpowiada najmniejszy oczekiwany błąd klasyfikacji (zamiast prognozowania klasy najbardziej prawdopodobnej),
 - **AttributeSelectedClassifier** – klasyfikator z optymalnie wybranymi atrybutami (zastosowanie metod wyboru cech zwanych wrapper'ami),
 - metody oparte na „zamianie typu zadania”, np.:
klasteryzacja \Rightarrow *klasyfikacja*, *predykcja* \Rightarrow *klasyfikacja*,
m.in.: *ClassificationViaRegression*, *RegressionByDiscretization*,
ClassificationViaClustering, *OrdinalClassClassifier*,
MultiClassClassifier,



Moduł Explorer III

Klasyfikacja – podział metod

Część: Klasyfikacja

Wprowadzenie
do WEKA

Adam Zagdański,
Artur Suchwałko
(www.suchwalcko.pl)

Czym jest WEKA?

Moduły dostępne
w WEKA

Moduł Explorer

Moduł *Knowledge
Flow*

WEKA – informacje
techniczne

Dodatkowe
informacje

- **misc** – pozostałe, niestandardowe algorytmy klasyfikacji,
- **trees** – klasyfikatory oparte na drzewach
(m.in.: *DecisionStump*, *Id3*, *J4.8*, *RandomForest*,
UserClassifier),
- **rules** – metody oparte na generowaniu (indukcji) reguł.



Moduł Explorer

Klasyfikacja – przykład wykorzystania drzew decyzyjnych

Część: Klasyfikacja

Wprowadzenie
do WEKA

Adam Zagdański,
Artur Suchwańko
(www.suchwańko.pl)

Czym jest WEKA?

Moduły dostępne
w WEKA

Moduł Explorer

Moduł Knowledge
Flow

WEKA – informacje
techniczne

Dodatkowe
informacje

The screenshot shows the Weka 3.5.7 Explorer window. The 'Classifier' tab is active, displaying the output of a decision tree classifier. The 'Classifier output' section shows the following rules:

```
petalwidth > 0.6  
|  
| petalwidth < 1.1  
| | petalwidth < 4.0 Iris-setosa (40.0/1.0)  
| | petalwidth > 4.0  
| | | petalwidth < 1.0 Iris-versicolour (13.0)  
| | | petalwidth > 1.0 Iris-virginica (19.0/1.1)  
| petalwidth > 1.7 Iris-virginica (46.8/1.3)
```

The 'Classifier output' section also displays the following statistics:

```
Number of Leaves : 5  
Size of the tree : 5  
Time taken to build model: 0.03 seconds  
*** Standardized cross-validation ***  
*** Summary ***  
Correctly Classified Instances 146 96 %  
Incorrectly Classified Instances 5 4 %  
Kappa statistic 0.94  
Mean absolute error 0.036  
Root Mean Squared Error 0.195  
Relative absolute error 7.975 %  
Root relative squared error 30.033 %  
Total Number of Instances 150  
*** Detailed Accuracy By Class ***  
TP Rate FP Rate Precision Recall F-Measure ROC Area Class  
0.98 0 1 0.99 0.99 0.99 Iris-setosa  
0.96 0.03 0.94 0.94 0.94 0.96 Iris-versicolour  
0.96 0.03 0.94 0.96 0.95 0.96 Iris-virginica
```

The 'Weka Classifier Tree Visualiz...' window shows a visualization of the decision tree structure, with nodes and branches representing the classification rules.

Rysunek: Przykład – klasyfikacja z wykorzystaniem drzew decyzyjnych



Moduł Explorer

Klasyfikacja – przykład wykorzystania sieci neuronowych

Część: Klasyfikacja

Wprowadzenie
do WEKA

Adam Zagdański,
Artur Suchwałko
(www.suchwalcko.pl)

Czym jest WEKA?

Moduły dostępne
w WEKA

Moduł Explorer

Moduł Knowledge
Flow

WEKA – informacje
techniczne

Dodatkowe
informacje

The screenshot displays the WEKA Explorer interface. The main window shows the 'Classifier' tab with 'MultilayerPerceptron' selected. The 'Test options' section is set to 'Cross-validation' with 'Folds' set to 20. The 'Classifier output' section shows the 'Pop statistics' for the selected classifier. A 'Neural Network' window is open in the foreground, showing a diagram of a neural network with 4 input nodes (green), 2 hidden nodes (red), and 3 output nodes (yellow). The output nodes are labeled 'Znaczenie', 'Wiek', and 'Zawodnik'. A 'GenericObjectEditor' window is also open, showing the configuration for the 'weka.classifiers.functions.MultilayerPerceptron' class, with various parameters like 'learningRate' and 'momentum' set to 0.3 and 0.2 respectively.

Rysunek: Przykład – klasyfikacja z wykorzystaniem sieci neuronowych



Moduł Explorer

Analiza skupień – zaimplementowane algorytmy

Część: Analiza skupień

Wprowadzenie
do WEKA

Adam Zagdański,
Artur Suchwałko
(www.suchwalko.pl)

Czym jest WEKA?

Moduły dostępne
w WEKA

Moduł Explorer

Moduł *Knowledge
Flow*

WEKA – informacje
techniczne

Dodatkowe
informacje

- **k-Means** – klasyczna metoda k-średnich,
- **EM** – klasteryzacja z wykorzystaniem algorytmu EM (Expectation Maximization),
- **Cobweb** – implementacja algorytmów: *Cobweb* dla zmiennych jakościowych oraz algorytmu *Classit* dla cech numerycznych. Wynikiem działania jest drzewo. Dla każdego przypadku wybierana jest najlepsza z czterech możliwości:
 - Dodanie przypadku do najlepszego hosta,
 - Utworzenie nowego liścia,
 - Połączenie dwóch najlepszych hostów i dodanie przypadku do połączonego węzła,
 - Podział najlepszego hosta i dodanie przypadku do jednego z otrzymanych podzbiorów,



Moduł Explorer

Analiza skupień – zaimplementowane algorytmy

Część: Analiza skupień

Wprowadzenie
do WEKA

Adam Zagdański,
Artur Suchwałko
(www.suchwalko.pl)

Czym jest WEKA?

Moduły dostępne
w WEKA

Moduł Explorer

Moduł *Knowledge
Flow*

WEKA – informacje
techniczne

Dodatkowe
informacje

- **X-means** – rozszerzona wersja algorytmu k-means, uzupełnie algorytmu o etap *Improve-Structure*. W tym etapie próbuje się podzielić centra (środki) w obrębie ich rejonu. Porównanie i wybór pomiędzy strukturą oryginalną i strukturą uzyskaną po podziale centrów, odbywa się na bazie wartości kryteriów BIC (odpowiadających obu strukturom),
- **FarthestFirst** – metoda klasteryzacji oparta na algorytmie przeszukiwania *farthest first*, autorstwa Hochbauma i Shmoys'a (1985); szybka i prosta, metoda poszukiwania k-średnich,
- **DBScan** – *Density-Based Spatial Clustering of Applications with Noise*; algorytm oparty na gęstościach,
- **OPTICS** – uporządkowanie obiektów umożliwiające identyfikację skupisk (interfejs graficzny).



Moduł Explorer

Analiza skupień – dodatkowe informacje

Część: Analiza skupień

Wprowadzenie
do WEKA

Adam Zagdański,
Artur Suchwałko
(www.suchwalcko.pl)

Czym jest WEKA?

Moduły dostępne
w WEKA

Moduł Explorer

Moduł *Knowledge
Flow*

WEKA – informacje
techniczne

Dodatkowe
informacje

- W WEKA nie są w tej chwili zaimplementowane klasyczne metody klasteryzacji hierarchicznej,
- Możliwa jest wizualizacja wyników analizy skupień i ich ewentualne porównanie z prawdziwymi grupami (jeżeli takie są znane),
- Możliwa jest ocena wyników oparta na funkcji wiarygodności, jeżeli algorytm klasteryzacji bazuje na założeniach dotyczących rozkładów,
- Klasa/wrapper *MakeDensityBasedClusterer* umożliwia „opakowanie” dowolnego algorytmu klasteryzacji, tak aby zwracał on rozkład i gęstość. Dopasowany jest rozkład normalny oraz rozkład dyskretny, oszacowane wewnątrz każdego skupiska, „wyprodukowanego” przez wyjściowy algorytm klasteryzacji.



Moduł Explorer

Analiza skupień – przykład zastosowania metody *k*-means

Część: Analiza skupień

Wprowadzenie
do WEKA

Adam Zagdański,
Artur Suchwałko
(www.suchwalko.pl)

Czym jest WEKA?

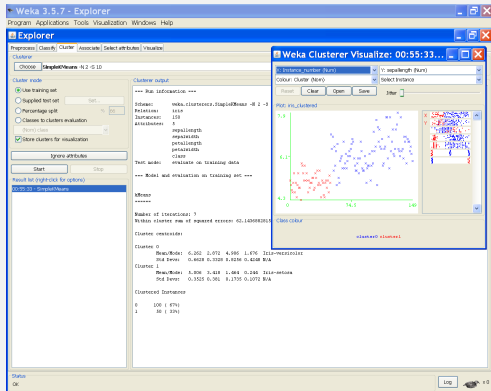
Moduły dostępne
w WEKA

Moduł Explorer

Moduł Knowledge
Flow

WEKA – informacje
techniczne

Dodatkowe
informacje



Rysunek: Przykład – zastosowanie metody *k*-means



WEKA – informacje techniczne

Wymagane oprogramowanie

Część: Wymagane oprogramowanie

Wprowadzenie
do WEKA

Adam Zagdański,
Artur Suchwałko
(www.suchwalko.pl)

Czym jest WEKA?

Moduły dostępne
w WEKA

Moduł Explorer

Moduł *Knowledge
Flow*

WEKA – informacje
techniczne

Dodatkowe
informacje

- WEKA może być uruchamiana praktycznie na dowolnej platformie (Windows, Linux, Mac),
- Do uruchomienia WEKA 3.4.x (i starszych wersji) wymagana jest Java 1.4 (lub nowsza wersja),
- Środowisko uruchomieniowe Java można pobrać za darmo np. ze strony [www Sun Microsystems \(http://www.sun.com/\)](http://www.sun.com/),
- Wersja rozwojowa (*developer version*), począwszy od v.3.5.3, wymaga już Java 5.0.



Wprowadzenie
do WEKA

Adam Zagdański,
Artur Suchwałko
(www.suchwalko.pl)

Czym jest WEKA?

Moduły dostępne
w WEKA

Moduł Explorer

Moduł *Knowledge
Flow*

WEKA – informacje
techniczne

Dodatkowe
informacje

- Dane mogą być importowane z plików w różnych formatach: ARFF, CSV, C4.5, format binarny,
- Można również wczytywać dane podając adres URL lub komunikując się z bazą danych za pomocą języka SQL (wykorzystywany jest JDBC – Java DataBase Connectivity),
- Domyślnym formatem danych wykorzystywanym w WEKA i opracowanym specjalnie na potrzeby tego projektu jest format **ARFF** – Attribute-Relation File Format,
- ARFF jest rodzajem pliku tekstowego ASCII, zawierającym dodatkowo informacje o typach atrybutów.



WEKA – informacje techniczne

Format danych

Część: Format danych

Wprowadzenie
do WEKA

Adam Zagdański,
Artur Suchwałko
(www.suchwalcko.pl)

Czym jest WEKA?

Moduły dostępne
w WEKA

Moduł Explorer

Moduł Knowledge
Flow

WEKA – informacje
techniczne

Dodatkowe
informacje

Przykładowe dane w formacie ARFF

```
% 1. Title: Iris Plants Database
%
% 2. Sources:
%   (a) Creator: R.A. Fisher
%   (b) Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)
%   (c) Date: July, 1988
%
@RELATION iris

@ATTRIBUTE sepallength NUMERIC
@ATTRIBUTE sepalwidth NUMERIC
@ATTRIBUTE petallength NUMERIC
@ATTRIBUTE petalwidth NUMERIC
@ATTRIBUTE class {Iris-setosa,Iris-versicolor,Iris-virginica}

@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
```



WEKA – użyteczne linki

Część: Użyteczne linki

Wprowadzenie
do WEKA

Adam Zagdański,
Artur Suchwałko
(www.suchwalko.pl)

Czym jest WEKA?

Moduły dostępne
w WEKA

Moduł Explorer

Moduł *Knowledge
Flow*

WEKA – informacje
techniczne

Dodatkowe
informacje

- **WEKA Homepage**
<http://www.cs.waikato.ac.nz/~ml/weka/>
- **WEKA Mailing list**
<https://list.scms.waikato.ac.nz/mailman/listinfo/wekalist>
- **WekaWiki**
<http://weka.wikispaces.com/>
- **Frequently Asked Questions (FAQ)**
<http://weka.wikispaces.com/Frequently+Asked+Questions>
- **Weka-related Projects**
http://www.cs.waikato.ac.nz/~ml/weka/index_related.html
- **Javadoc**
<http://weka.sourceforge.net/doc/>



Wybrane projekty stworzone na bazie WEKA

Źródło: Weka-related Projects,

http://www.cs.waikato.ac.nz/ml/weka/index_related.html

Część: Wybrane projekty stworzone na bazie WEKA

Wprowadzenie
do WEKA

Adam Zagdański,
Artur Suchwałko
(www.suchwalko.pl)

Czym jest WEKA?

Moduły dostępne
w WEKA

Moduł Explorer

Moduł *Knowledge
Flow*

WEKA – informacje
techniczne

Dodatkowe
informacje

- **YALE** - Yet Another Learning Environment,
- **Weka-Parallel** - parallel processing for Weka,
- **Automatic Knowledge Miner** - online data mining reports,
- **Weka Visualization tools** - using PMML, VisWiz, and ROCOn,
- **Weka on Text** - software for text mining,
- **Judge** - software for document classification and clustering,
- **Grid Weka** - grid computing with Weka,
- **FAEHIM** - Data Mining Web services,
- **Fuzzy algorithms** - for clustering and classification.



Wybrane projekty stworzone na bazie WEKA

Źródło: Weka-related Projects,

http://www.cs.waikato.ac.nz/ml/weka/index_related.html

Część: Wybrane projekty stworzone na bazie WEKA

Wprowadzenie
do WEKA

Adam Zagdański,
Artur Suchwałko
(www.suchwalko.pl)

Czym jest WEKA?

Moduły dostępne
w WEKA

Moduł Explorer

Moduł *Knowledge
Flow*

WEKA – informacje
techniczne

Dodatkowe
informacje

- **BioWeka** - knowledge discovery and analysis for biologists,
- **Mathematica interface for Weka**,
- **weka4WS** - distributed data mining,
- **RWeka** - an R interface to Weka,
- **Mayday** - Machine Learning for Microarrays - plugin for the WEKA machine Learning Library,
- **PROMPT** - Statistical comparison and mapping of protein sets. Import/Export of WEKA arff data files,
- **GeneticProgramming** - Genetic Programming Classifier for Weka,
- **Weka-GDPM** - extended version of Weka 3.4 to support automatic geographic data preprocessing for spatial data mining.



Książka – DM z wykorzystaniem WEKA

Część: Książka – DM z wykorzystaniem WEKA

Wprowadzenie
do WEKA

Adam Zagdański,
Artur Suchwałko
(www.suchwalcko.pl)

Czym jest WEKA?

Moduły dostępne
w WEKA

Moduł Explorer

Moduł *Knowledge
Flow*

WEKA – informacje
techniczne

Dodatkowe
informacje



- Ian H. Witten, Eibe Frank, *Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)*, Morgan Kaufmann, 2005
- Adres [www](http://www.cs.waikato.ac.nz/~ml/weka/book.html):
<http://www.cs.waikato.ac.nz/~ml/weka/book.html>
- Wyjaśnienie idei działania algorytmów Data mining
- Pomoc w wyborze odpowiedniego algorytmu dla określonego problemu oraz odpowiedniej metod oceny efektywności



WEKA – dokumentacja techniczna w formacie javadoc

Część: WEKA – dokumentacja techniczna

Wprowadzenie do WEKA

Adam Zagdański,
Artur Suchwałko
(www.suchwalko.pl)

Czym jest WEKA?

Moduły dostępne w WEKA

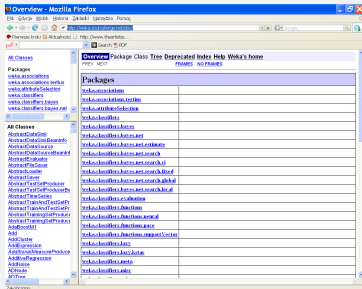
Moduł Explorer

Moduł Knowledge Flow

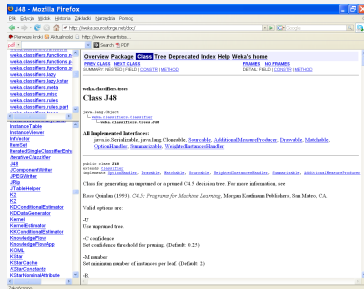
WEKA – informacje techniczne

Dodatkowe informacje

- Metody zaimplementowane w WEKA mogą być wykorzystywane bez konieczności uruchomienia graficznych interfejsów użytkownika,
- Kod źródłowy (klasy) zaimplementowane w ramach projektu WEKA mogą być wykorzystywane np. we własnych aplikacjach napisanych w języku Java lub w innych programach, np. w systemie statystycznym **R** (<http://cran.r-project.org/web/packages/RWeka/RWeka.pdf>) lub w programie **RapidMiner** (<http://www.rapidminer.com/>),
- Korzystanie z klas WEKA ułatwia szczegółowa dokumentacja techniczna w formacie javadoc.



Rysunek: Struktura klas



Rysunek: Klasa J48 (drzewo klasyfikacyjne)



Wprowadzenie
do WEKA

Adam Zagdański,
Artur Suchwałko
(www.suchwalcko.pl)

Czym jest WEKA?

Moduły dostępne
w WEKA

Moduł Explorer

Moduł *Knowledge
Flow*

WEKA – informacje
techniczne

Dodatkowe
informacje



Weka home page.

Internet.

<http://www.cs.waikato.ac.nz/ml/weka/>.



Weka wiki.

Internet.

<http://weka.wikispaces.com/>.



I.H. Witten and E. Frank.

*Data Mining: Practical Machine Learning Tools and
Techniques.*

Morgan Kaufmann, 2005.