WekaBioSimilarity — extending Weka with resemblance measures[*]

C. Domínguez, J. Heras, E. Mata, and V. Pascual

Department of Mathematics and Computer Science, University of La Rioja

15 September 2016
TAMIDA'16

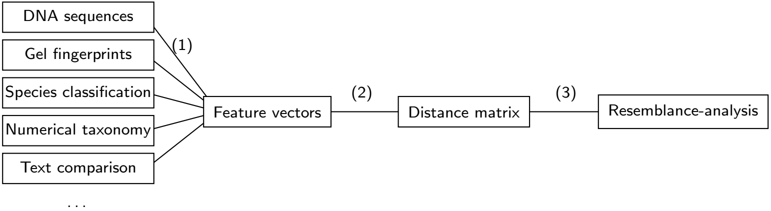
[*] Partially supported by the Agencia de Desarrollo Económico de La Rioja (ADER) [2013-I-IDD-00123] and Ministerio de Economía y Competitividad [MTM2014-54151-P]

Introduction

Resemblance-analysis is an important concern in several fields

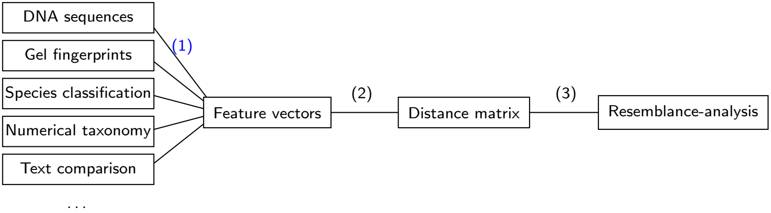
Introduction

Resemblance-analysis is an important concern in several fields



Introduction

Resemblance-analysis is an important concern in several fields

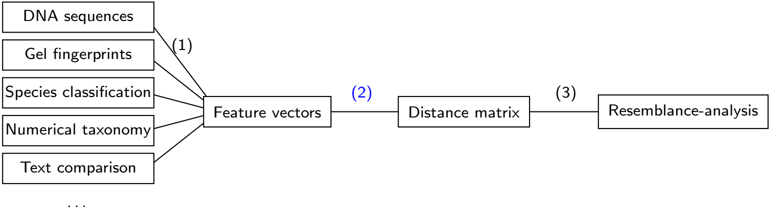


Depend on the concrete problem

Different types of features: binary, multi-value (nominal), string, or numerical

Introduction

Resemblance-analysis is an important concern in several fields

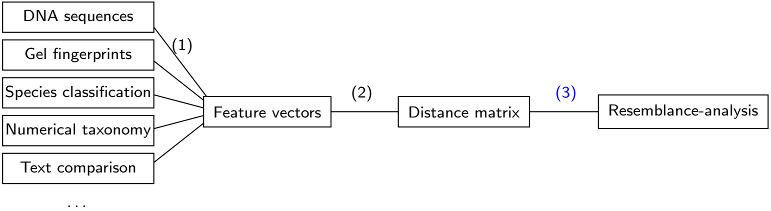


Several resemblance (similarity and distance) measures have been proposed

S. S. Choi et al. *A survey of binary similarity and distance measures*. Journal of Systemics, Cybernetics and Informatics 8(1),43–48. 2010.

Introduction

Resemblance-analysis is an important concern in several fields



Clustering algorithms

Hierarchical clustering can be visualised using a tree representation

Problem

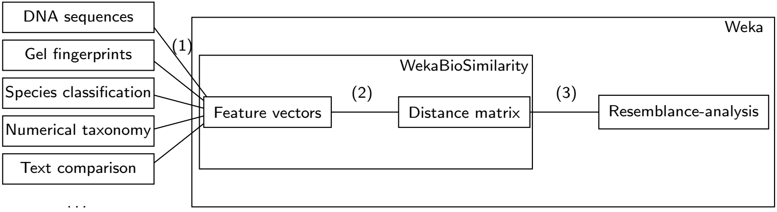
Several statistical packages (e.g. R, Matlab, Octave, Weka, or SPSS) provide the functionality for resemblance analysis

Problem

Several statistical packages (e.g. R, Matlab, Octave, Weka, or SPSS) provide the functionality for resemblance analysis

- Problem: they only support binary/or numerical features
- They cannot handle:
  - Comparison of DNA sequences (multi-value/string descriptors)
  - DNA fingerprints (numerical feature vectors)
  - Phylogenetic or data mining (heterogeneous descriptors)
  - . . .

Solution



WekaBioSimilarity:

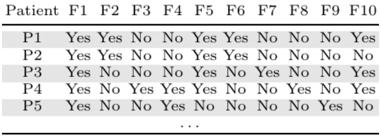
- Extends Weka with resemblance measures and comparison modes
- Works with binary, numerical, nominal, and heterogeneous data
- Open, easily extensible and integrable with other systems

Outline

1. Introduction
2. Resemblance measures in WekaBioSimilarity
3. Integration of WekaBioSimilarity in Weka
4. Conclusions

Binary data

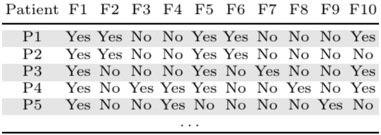
SPECT heart dataset:



Patient	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
P1	Yes	Yes	No	No	Yes	Yes	No	No	No	Yes
P2	Yes	Yes	No	No	Yes	Yes	No	No	No	No
P3	Yes	No	No	No	Yes	No	Yes	No	No	Yes
P4	Yes	No	Yes	Yes	Yes	No	No	Yes	No	Yes
P5	Yes	No	No	Yes	No	No	No	No	Yes	No
					...					

Binary data

SPECT heart dataset:



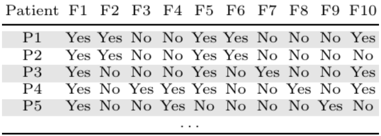
Patient	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
P1	Yes	Yes	No	No	Yes	Yes	No	No	No	Yes
P2	Yes	Yes	No	No	Yes	Yes	No	No	No	No
P3	Yes	No	No	No	Yes	No	Yes	No	No	Yes
P4	Yes	No	Yes	Yes	Yes	No	No	Yes	No	Yes
P5	Yes	No	No	Yes	No	No	No	No	Yes	No
...										

Given two objects A and B, four values are computed:

- $M_{11} = \sharp$`attributes present both in A and B`
- $M_{10} = \sharp$`attributes present in A but not in B`
- $M_{01} = \sharp$`attributes present in B but not in A`
- $M_{00} = \sharp$`attributes present neither in A nor in B`

Binary data

SPECT heart dataset:



Patient	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
P1	Yes	Yes	No	No	Yes	Yes	No	No	No	Yes
P2	Yes	Yes	No	No	Yes	Yes	No	No	No	No
P3	Yes	No	No	No	Yes	No	Yes	No	No	Yes
P4	Yes	No	Yes	Yes	Yes	No	No	Yes	No	Yes
P5	Yes	No	No	Yes	No	No	No	No	Yes	No
					...					

Given two objects A and B, four values are computed:

- $M_{11} = \sharp\texttt{attributes present both in A and B}$
- $M_{10} = \sharp\texttt{attributes present in A but not in B}$
- $M_{01} = \sharp\texttt{attributes present in B but not in A}$
- $M_{00} = \sharp\texttt{attributes present neither in A nor in B}$

Resemblance measures are computed from those values:

$$S(A,B) = \frac{M_{11} + M_{00}}{M_{11} + M_{10} + M_{01} + M_{00}}$$

Binary data

SPECT heart dataset:

Patient	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
P1	Yes	Yes	No	No	Yes	Yes	No	No	No	Yes
P2	Yes	Yes	No	No	Yes	Yes	No	No	No	No
P3	Yes	No	No	No	Yes	No	Yes	No	No	Yes
P4	Yes	No	Yes	Yes	Yes	No	No	Yes	No	Yes
P5	Yes	No	No	Yes	No	No	No	No	Yes	No
					...					

Given two objects A and B, four values are computed:

- $M_{11} = \sharp$`attributes present both in A and B`
- $M_{10} = \sharp$`attributes present in A but not in B`
- $M_{01} = \sharp$`attributes present in B but not in A`
- $M_{00} = \sharp$`attributes present neither in A nor in B`

Resemblance measures are computed from those values:

$$S(A,B) = \frac{M_{11} + M_{00}}{M_{11} + M_{10} + M_{01} + M_{00}}$$

WekaBioSimilarity features 76 binary resemblance measures

Multi-value/String data — pairwise comparison

HIV-1 protease cleavage dataset:

```
I1: AAAKFERQ
I2: AAAMKRHG
I3: AAAMSSAI
```

Multi-value/String data — pairwise comparison

HIV-1 protease cleavage dataset:

```
I1: AAAKFERQ
I2: AAAMKRHG
I3: AAAMSSAI
```

- Length of feature vectors is the same
- Position of attributes in feature vectors is important

Multi-value/String data — pairwise comparison

HIV-1 protease cleavage dataset:

```
I1: AAAKFERQ
I2: AAAMKRHG
I3: AAAMSSAI
```

- Length of feature vectors is the same
- Position of attributes in feature vectors is important
- Given two objects A and B, agreements and disagreements are computed

Multi-value/String data — pairwise comparison

HIV-1 protease cleavage dataset:

```
I1: AAAKFERQ
I2: AAAMKRHG
I3: AAAMSSAI
```

- Length of feature vectors is the same
- Position of attributes in feature vectors is important
- Given two objects A and B, agreements and disagreements are computed
- Resemblance measures are computed from those values:

$$S(A,B) = \frac{agreements}{agreements + disagreements}$$

Multi-value/String data — pairwise comparison

HIV-1 protease cleavage dataset:

```
I1: AAAKFERQ
I2: AAAMKRHG
I3: AAAMSSAI
```

- Length of feature vectors is the same
- Position of attributes in feature vectors is important
- Given two objects A and B, agreements and disagreements are computed
- Resemblance measures are computed from those values:

$$S(A,B) = \frac{agreements}{agreements + disagreements}$$

- WekaBioSimilarity features 25 resemblance measures (generalised from the binary case)

Multi-value/String data — set occurrence

USDA plants database:

```
abelia: fl, nc
abelia x grandiflora: fl, nc
abel.: ct, dc, fl, hi, il, ky, la, md, mi, ms
abel. esc.: ct, dc, fl, il, ky, la, md, mi, ms
```

Multi-value/String data — set occurrence

USDA plants database:

```
abelia: fl, nc
abelia x grandiflora: fl, nc
abel.: ct, dc, fl, hi, il, ky, la, md, mi, ms
abel. esc.: ct, dc, fl, il, ky, la, md, mi, ms
```

- Size of feature vectors might be different
- Position of attributes is not relevant

Multi-value/String data — set occurrence

USDA plants database:

```
abelia: fl, nc
abelia x grandiflora: fl, nc
abel.: ct, dc, fl, hi, il, ky, la, md, mi, ms
abel. esc.: ct, dc, fl, il, ky, la, md, mi, ms
```

- Size of feature vectors might be different
- Position of attributes is not relevant
- Given A and B, $|S_A \cap S_B|$, $|S_A \setminus S_B|$ and $|S_B \setminus S_A|$ are used to compute resemblance measures:

$$S(A,B) = \frac{|S_A \cap S_B|}{|S_A \cap S_B| + |S_A \setminus S_B| + |S_B \setminus S_A|}$$

Multi-value/String data — set occurrence

USDA plants database:

```
abelia: fl, nc
abelia x grandiflora: fl, nc
abel.: ct, dc, fl, hi, il, ky, la, md, mi, ms
abel. esc.: ct, dc, fl, il, ky, la, md, mi, ms
```

- Size of feature vectors might be different
- Position of attributes is not relevant
- Given A and B, $|S_A \cap S_B|$, $|S_A \setminus S_B|$ and $|S_B \setminus S_A|$ are used to compute resemblance measures:

$$S(A,B) = \frac{|S_A \cap S_B|}{|S_A \cap S_B| + |S_A \setminus S_B| + |S_B \setminus S_A|}$$

- WekaBioSimilarity features 76 resemblance measures (the same as in the binary case)

Numerical data

Three cases:

- Typically: Euclidean distance, Pearson correlation, and so on

Numerical data

Three cases:

- Typically: Euclidean distance, Pearson correlation, and so on
- Pairwise comparison (e.g. compare regions based on age demographics)
- Set occurrence (e.g. classify DNA fingerprints)

Numerical data

Three cases:

- Typically: Euclidean distance, Pearson correlation, and so on
- Pairwise comparison (e.g. compare regions based on age demographics)
- Set occurrence (e.g. classify DNA fingerprints)

First case available in Weka and most statistical packages

Numerical data

Three cases:

- Typically: Euclidean distance, Pearson correlation, and so on
- Pairwise comparison (e.g. compare regions based on age demographics)
- Set occurrence (e.g. classify DNA fingerprints)

First case available in Weka and most statistical packages
Last two cases only available on WekaBioSimilarity

Numerical data

Three cases:

- Typically: Euclidean distance, Pearson correlation, and so on
- Pairwise comparison (e.g. compare regions based on age demographics)
- Set occurrence (e.g. classify DNA fingerprints)

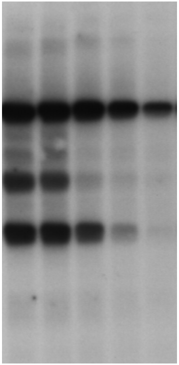
First case available in Weka and most statistical packages
Last two cases only available on WekaBioSimilarity
Notion of closeness: tolerance value

Example: DNA fingerprinting

Comparison of DNA patterns

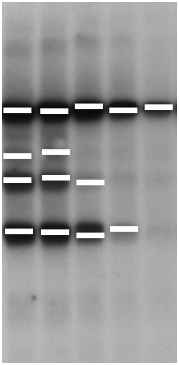
Example: DNA fingerprinting

Comparison of DNA patterns



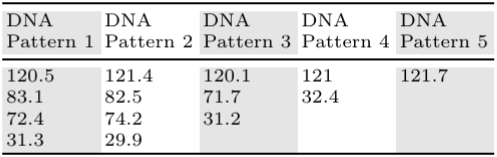
Example: DNA fingerprinting

Comparison of DNA patterns



Example: DNA fingerprinting

Comparison of DNA patterns



DNA Pattern 1	DNA Pattern 2	DNA Pattern 3	DNA Pattern 4	DNA Pattern 5
120.5	121.4	120.1	121	121.7
83.1	82.5	71.7	32.4	
72.4	74.2	31.2		
31.3	29.9			

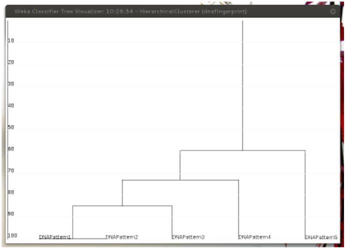
Values are different, but they are close enough (tolerance)

Example: DNA fingerprinting

Comparison of DNA patterns

DNA Pattern 1	DNA Pattern 2	DNA Pattern 3	DNA Pattern 4	DNA Pattern 5
120.5	121.4	120.1	121	121.7
83.1	82.5	71.7	32.4	
72.4	74.2	31.2		
31.3	29.9			

Values are different, but they are close enough (tolerance)


Weka Classifier Tree Visualizer: 10:26:34 - HierarchicalClusterer (dnafingerprint)
10
20
30
40
50
60
70
80
90
100
DNAPattern1
DNAPattern2
DNAPattern3
DNAPattern4
DNAPattern5

Heterogeneous data

Attributes that describe an object may have different types:

- E.g. Numerical taxonomy (attributes like presence of hair, habitat, number of limbs, . . .)

Heterogeneous data

Attributes that describe an object may have different types:

- E.g. Numerical taxonomy (attributes like presence of hair, habitat, number of limbs, . . .)
- Only pairwise comparison can be applied

Heterogeneous data

Attributes that describe an object may have different types:

- E.g. Numerical taxonomy (attributes like presence of hair, habitat, number of limbs, . . .)
- Only pairwise comparison can be applied
- WekaBioSimilarity implements 25 measures

Outline

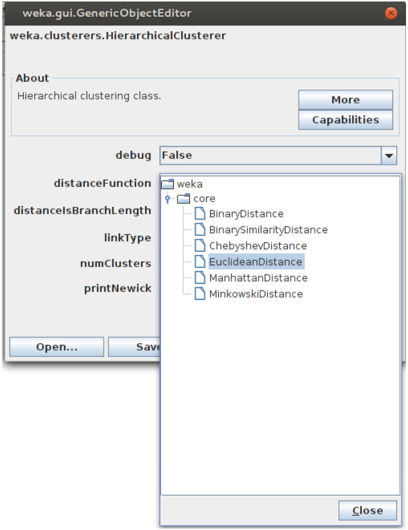
1 Introduction

2 Resemblance measures in WekaBioSimilarity

3 Integration of WekaBioSimilarity in Weka

4 Conclusions

Integration of WekaBioSimilarity in Weka



weka.gui.GenericObjectEditor

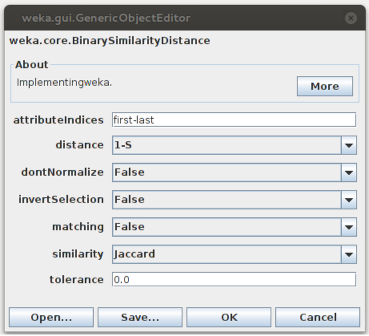
weka.clusterers.HierarchicalClusterer

About

Hierarchical clustering class.

debug: False

Integration of WekaBioSimilarity in Weka



weka.core.BinarySimilarityDistance

About

Implementingweka.

attributeIndices first-last

distance 1-S

dontNormalize False

invertSelection False

matching False

similarity Jaccard

tolerance 0.0

Outline

1 Introduction

2 Resemblance measures in WekaBioSimilarity

3 Integration of WekaBioSimilarity in Weka

4 Conclusions

Conclusions

Resemblance analysis:

- common problem in several contexts
- highly dependent on computing resemblance among objects
- usually special-purpose packages are necessary

Conclusions

Resemblance analysis:

- common problem in several contexts
- highly dependent on computing resemblance among objects
- usually special-purpose packages are necessary

WekaBioSimilarity:

- features the same binary measures that other packages
- generalises binary measures to other types of descriptors
- provides tolerance for numerical data
- supports two comparison modes

Conclusions

Resemblance analysis:

- common problem in several contexts
- highly dependent on computing resemblance among objects
- usually special-purpose packages are necessary

WekaBioSimilarity:

- features the same binary measures that other packages
- generalises binary measures to other types of descriptors
- provides tolerance for numerical data
- supports two comparison modes

Result:

- a tool applicable in a wide-variety of problems
- used as a standalone application or integrated in other software

WekaBioSimilarity — extending Weka with resemblance measures

C. Domínguez, J. Heras, E. Mata, and V. Pascual

Department of Mathematics and Computer Science, University of La Rioja

15 September 2016
TAMIDA'16