WekaBioSimilarity — extending Weka with resemblance measures

César Domínguez, Jónathan Heras, Eloy Mata, and Vico Pascual

Department of Mathematics and Computer Science, University of La Rioja, Spain
{cesar.dominguez, jonathan.heras, eloy.mata, vico.pascual}@unirioja.es

Abstract. The classification of organisms is a daily-basis task in biology as well as other contexts. This process is usually carried out by comparing a set of descriptors associated with each object. However, general-purpose statistical packages offer a limited number of methods to perform such a comparison, and specific tools are required for each concrete problem. Weka is a freely-available framework that supports both supervised and unsupervised machine-learning algorithms. Here, we present WekaBioSimilarity, an extension of Weka implementing several resemblance measures to compare different kinds of descriptors. Namely, WekaBioSimilarity works with binary, multi-value, string, numerical, and heterogeneous data. WekaBioSimilarity, together with Weka, offers the functionality to classify objects using different resemblance measures, and clustering and classification algorithms. The combination of these two systems can be used as a standalone application or can be incorporated in the workflow of other software systems that require a classification process. WekaBioSimilarity is available at http://wekabiosimilarity.sourceforge.net.

1 Introduction

The classification and resemblance-analysis of objects is one of the most important concerns in several areas such as biology [23], text comparison [14], chemistry [30], geology [15], biometrics [29], complex networks [22], and web data-mining [20] among others. The similarity among objects is obtained through the comparison of a set of descriptors (encoded by means of feature vectors); and such a similarity is the basis to classify the objects into groups. The descriptors depend on the concrete problem, and their possible types are binary, multi-value (also known as nominal), string or numerical.

The most common procedure to classify objects consists of two steps: the computation of a *distance matrix*, and the construction of *clusters*. In the former step, the resemblance information is gathered into a square matrix called the distance matrix: given a list of n objects L, the distance matrix of L is an $n \times n$ matrix where the element of row i and column j encodes the distance between the ith and jth object of L. In the latter step, the distance matrix is used to group objects using clustering algorithms [31]. In several contexts, it is common to use a special kind of clustering algorithms called hierarchical; the clusters

produced by these methods can be visualised using a tree representation (e.g. dendrograms, cladograms and evolutionary trees). The classification process is summarised in Figure 1.

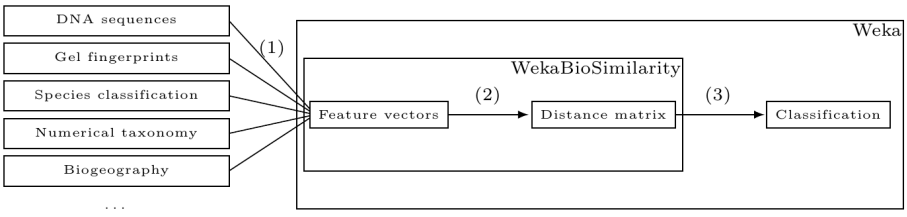


Fig. 1. *Workflow of the classification process proposed by the authors.* In our case, Step (1) is a preprocessing stage that depends on the concrete problem, Step (2) is provided by WekaBioSimilarity using several resemblance measures, and Step (3) is carried out using Weka algorithms.

Estimating the distance between objects is the crucial aspect of the classification process, and several *distance* and *similarity* measures (in general, *resemblance* measures) have been proposed in the literature [4] — similarity measures usually range from 0 to 1; and given a similarity measure S, its associated distance D can be computed as $D = 1 - S$, $D = \sqrt{1 - S}$, or $D = \sqrt{1 - S^2}$ [11].

General-purpose statistical packages (e.g. R, Matlab, Octave, Weka, or SPSS) provide the functionality to classify data using clustering algorithms, but they only support a few resemblance measures for either binary or numerical feature vectors — a summary of the resemblance measures included in these systems is provided in Table 1. This means that problems like the comparison of DNA sequences [16] (uses multi-value/string vectors), DNA fingerprints [7] (uses numerical vectors), or, in general, phylogenetics [17] or data-mining [1] (might work with heterogeneous descriptors) cannot be directly handled in these systems, and special-purpose packages are required.

Weka [5] is an open-source interface that serves to run a wide variety of machine-learning algorithms. It features several variants of clustering; however, it only supports 4 distance measures to compare objects (all of them related to numerical descriptors, see Table 1). In this paper, we present *WekaBioSimilarity*, a Weka extension implemented in Java that enhances this system with several resemblance measures and comparison modes for different types of descriptors. WekaBioSimilarity includes resemblance measures that have not been included in any other package; it is open, free, easily extensible and integrable in other systems.

The rest of this paper is organised as follows. In Section 2, we explain how WekaBioSimilarity computes the similarity among feature vectors of different type — for the sake of explanation, several simple examples are provided in this section. The integration of WekaBioSimilarity in Weka is presented in Sec-

System	Measures
Matlab	**Numerical measures:** Euclidean, Standarized Euclidean, CityBlock, Minkowski, Chebychev, Mahalanobis, Cosine, Correlation, Spearman, Hamming, Jaccard.
Octave	**Numerical measures:** Euclidean, Squared Euclidean, Chi-Squared, Cosine, Earth Mover's, L1.
R	**Numerical measures:** Euclidean, Maximum, Manhattan, Canberra, Binary, Minkowski. **Binary measures:** soerensen, jaccard, ochiai, mountford, whittaker, lande, wilsonshmida, cocogaston, magurran, harrison, cody, williams, williams2, harte, simpson, lennon, weiher, ruggiero, lennon2, rout1ledge, rout2ledge, rout3ledge, sokal1, dice, kulcz1insky, kulcz2insky, mcconnagh, manhattan, simplematching, margaleff, pearson, roger, baroni, dennis, fossum, gower, legendre, sokal2, sokal3, sokal4, stiles, yule, michael, hamann, forbes, chisquare, peirce, eyraud, simpson2, legendre2, fager, maarel, lamont, johnson, sorgenfrei, johnson2.
SPSS	**Numerical measures:** Euclidean, Squared Euclidean, Pearson correlation, Cosine, Chebychev, Block, Minkowski, Customized. **Binary measures:** Euclidean, Squared Euclidean, Size difference, Pattern difference, Variance, Dispersion, Shape, Simple Matching, Phi 4-point correlation, Lambda, Anderberg's D, Dice, Hamann, Jaccard, Kulczynski 1, Kulczynski 2, Lance and Williams, Ochiai, Rogers and Tanimoto, Sokal and Sneath 1, Sokal and Sneath 2, Sokal and Sneath 3, Sokal and Sneath 4, Sokal and Sneath 5, Yule's Y, Yule's Q
Weka	**Numerical measures:** Chebyshev, Euclidean, Manhattan, Minkowski
WekaBioSimilarity	**Numerical measures:** Chebyshev, Euclidean, Manhattan, Minkowski. **Numerical, binary, multi-value/string, and heterogeneous measures:** Jaccard, Dice, Czekanowski, Jaccard3W, NeiLi, SokalSneathI, SokalMichener, SokalSneathII, RogerTanimoto, Faith, GowerLegendre, Intersection, InnerProduct, GilbertWells, OchiaiI, Forbesi, Fossum, SorgenFrei, Mountford, Otsuka, McConnaughey, Tarwid, KulczynskiII, DriverKroeber, Johnson, Dennis, Simpson, BraunBanquet, Ample, Tarantula, Eyraud, Peirce, BaroniUrbaniBuserII, BaroniUrbaniBuserI, GoodManKruskal, Anderberg, Michael, Hamann, Disperson, Tanimoto, KulczynskiI, Yulew, Yuleq, OchiaiII, Stiles, Cole, SokalSneathIII, PearsonHeronII, PearsonHeronI, PearsonIII, PearsonI, PearsonII, Gower, SokalSneathIV, ForbesII, FagerMcGowan, Hamming, Euclidean, SquareEuclidean, Manhattan, MeanManhattan, Vari, SizeDifference, ShapeDifference, PatternDifference, LanceWilliams, BrayCurtis, Hellinger, Chord

Table 1. *Statistical packages and their resemblance measures.*

tion 3. Section 4 is devoted to present a case study where WekaBioSimilarity is applied to compare DNA fingerprints, and Section 5 introduces how resemblance measures can be compared using supervised-learning algorithms in Weka. The paper ends with the Conclusions and the Bibliography. WekaBioSimilarity is available at http://wekabiosimilarity.sourceforge.net — installation instructions, examples and videos can also be downloaded from this webpage.

2 Resemblance Measures in WekaBioSimilarity

WekaBioSimilarity works with data of different type: binary, multi-value, string, numerical and heterogeneous. Each type of data has its particularities, and, therefore, the comparison of feature vectors is different in each case. The comparison of binary data is tackled by SPSS and the Simba [8] package of R, but Weka together with WekaBioSimilarity is the only general-purpose tool that works with the other kind of data (see Table 1).

In this section, we explain how WekaBioSimilarity carries out the comparison of feature vectors of different types. We will illustrate the functionality for each case considering real-world examples obtained from the UC Irvine Machine Learning Repository [12] — for the sake of explanation, we will use fragments of those datasets.

2.1 Binary data

In the simplest case, objects are represented by means of binary feature vectors that encode the presence/absence of a set of attributes (or properties). Given two objects, A and B, represented by means of binary vectors, four values are

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
					...					

$$\begin{array}{c} \\ P1 \\ P2 \\ P3 \\ P4 \\ P5 \end{array} \begin{array}{c} \begin{array}{ccccc} P1 & P2 & P3 & P4 & P5 \end{array} \\ \begin{pmatrix} 1 & 0.9 & 0.7 & 0.5 & 0.4 \\ 0.9 & 1 & 0.6 & 0.4 & 0.5 \\ 0.7 & 0.6 & 1 & 0.6 & 0.5 \\ 0.5 & 0.4 & 0.6 & 1 & 0.5 \\ 0.4 & 0.5 & 0.5 & 0.5 & 1 \end{pmatrix} \end{array}$$

Fig. 2. *Fragment of SPECT heart dataset.* **Left.** Table wit the first 10 binary features of the first 5 patients in the dataset. **Right.** Similarity matrix obtained using the simple-matching measure.

computed: M_{11} (the number of attributes present both in A and B), M_{10} (the number of attributes present in A but not in B), M_{01} (the number of attributes present in B but not in A), and M_{00} (the number of attributes present neither in A nor in B). From these values, several resemblance measures can be defined; for instance, $S(A,B) = \frac{M_{11}+M_{00}}{M_{11}+M_{10}+M_{01}+M_{00}}$ is the simple-matching similarity measure.

The WekaBioSimilarity package features the 76 binary resemblance measures surveyed in [4] including widely employed measures like Dice or Jaccard (see Table 1 for the complete list of measures included in WekaBioSimilarity).

Example 1. The SPECT heart dataset [9] is a database that describes diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images. This dataset contains 267 instances that are described by means of 23 binary features indicating partial diagnosis at different stages. A fragment of this database is shown in Figure 2. Using such information and the simple-matching measure, the similarity matrix obtained by WekaBioSimilarity is given in Figure 2.

2.2 Multi-value/String data

In the binary case, the resemblance of two objects is computed through a pairwise comparison of the feature vectors associated with the objects. Therefore, the length of the vectors must be the same and the position of the elements is relevant. This approach can also be applied to multi-value and string feature vectors (e.g. in the comparison of DNA sequences, each descriptor is one of the four nucleobases [16]). To compare multi-value/string vectors, we must consider *agreements* and *disagreements* (descriptors for which the two objects have, respectively, the same and different values), and extend, when possible, the binary measures. For instance, the generalisation of the simple-matching measure is $S(A,B) = \frac{agreements}{agreements+disagreements}$.

Example 2. The HIV-1 protease cleavage dataset [21] contains lists of octamers (8 amino acids) — each one of the 8 attributes is a letter denoting an amino acid (e.g. G is Glycine and P is Proline) — and a flag (-1 or 1) depending on whether HIV-1 protease will cleave in the central position (between amino acids 4 and 5). If we compare the three first instances of this dataset (`I1: AAAKFERQ,-1`,

`I2: AAAMKRHG,-1`, and `I3: AAAMSSAI,-1`) using the simple-matching measure, WekaBioSimilarity can compute the similarity between them obtaining the following results: $S(\texttt{I1}, \texttt{I2}) = 0.44$, $S(\texttt{I1}, \texttt{I3}) = 0.44$, and $S(\texttt{I2}, \texttt{I3}) = 0.55$.

A different situation occurs when a string feature vector represents the set of components of an object (e.g. in the study of the distribution of species [13], or for tracking the different areas of a web site that are visited by users [3]). Hence, the size of the vectors associated with two objects might be different, and the position of the attributes in those vectors is no longer relevant. In this situation, the similarity of two objects A and B (which associated sets are S_A and S_B, respectively) is obtained using three values: $|S_A \cap S_B|$, $|S_A \setminus S_B|$ and $|S_B \setminus S_A|$; and generalising the binary measures. For instance, the simple-matching measure is given by $S(A, B) = \frac{|S_A \cap S_B|}{|S_A \cap S_B| + |S_A \setminus S_B| + |S_B \setminus S_A|}$.

Example 3. The USDA plants database [25] contains the plants of the USA and the states where they occur (the original dataset consists of almost 35000 instances, and each plant inhabits from 1 to 69 regions). Some instances of such a database are: *abelia*: fl, nc; *abelia x grandiflora*: fl, nc; *abel.*: ct, dc, fl, hi, il, ky, la, md, mi, ms; and, *abel. esc.*: ct, dc, fl, il, ky, la, md, mi, ms. Using the simple-matching measure, the similarity between the species abelmoschus (abel.) and abelmoschus esculentus (abel. esc.) is given by

$$S(abel., abel.\ esc.) = \frac{|\{ct, dc, fl, il, ky, la, md, mi, ms\}|}{|\{ct, dc, fl, hi, il, ky, la, md, mi, ms\}| + |\{hi\}| + |\emptyset|} = \frac{9}{10}.$$

When working with multi-value/string feature vectors, the user of WekaBioSimilarity must select the kind of comparison that is performed: pairwise (former scenario) or set-occurrence (latter scenario) — see Figure 3. In the pairwise case, the 25 binary measures that can be generalised are supported; and, in the set-occurrence case, all the measures supported for binary data are available.

2.3 Numerical data

The comparison of numerical data is usually performed using measures like the Euclidean distance or the Pearson correlation coefficient [11] — 4 of these measures were already implemented in Weka (see Table 1). In addition, the two situations presented in the previous subsection also make sense when working with numerical feature vectors (e.g. to compare regions based on age demographics, or to classify DNA fingerprints [7]). Hence, the pairwise-comparison and set-occurrence modes have been implemented for numerical feature vectors in WekaBioSimilarity.

Working with numerical descriptors has a particularity: there exists a notion of "closeness": values that, in spite of not being equal, are close enough to be considered the same. To deal with this issue, WekaBioSimilarity provides a configurable parameter called *tolerance* — see Figure 3. This parameter allows the

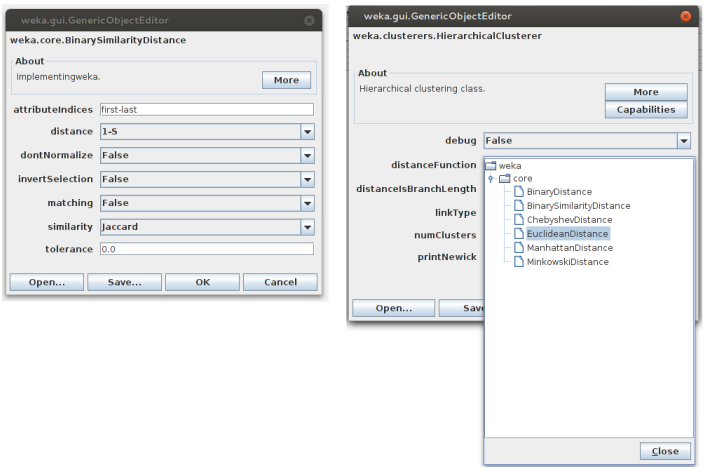


Fig. 3. Left: WekaBioSimilarity interface. **Right:** Distances available in Weka.

user to fix the "closeness" depending on the concrete problem, and it is essential, for instance, when precision errors must be taken into account. A detailed example will be introduced in Section 4.

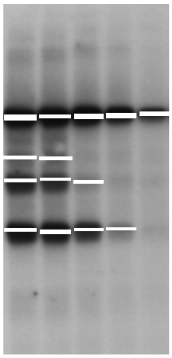
2.4 Heterogeneous data

In general, the attributes that describe an object may have different types — e.g. when comparing animals, descriptors like the presence of hair (binary), habitat (multi-value), or number of limbs (numerical) might be considered. In this context, only a pairwise comparison can be applied, and WekaBioSimilarity implements 25 measures for heterogeneous data (analogously to the multi-value/string and numerical situations). In the case of numerical descriptors included in heterogeneous data, the user can also fix a tolerance value as explained in the previous subsection. We will introduce an example of heterogeneous data in Section 5.

3 Integration of WekaBioSimilarity in Weka

In this section, we explain how Weka integrates WekaBioSimilarity. Weka implements several supervised and unsupervised algorithms that require the selection of a distance measure; for instance, k-means or hierarchical clustering in the former case, and k-nearest neighbour or locally-weighted learning in the latter. Once WekaBioSimilarity is installed in Weka, the resemblance measures implemented in this plugin become available for those algorithms, see Figure 3.

WekaBioSimilarity automatically recognises the kind of data that the user is processing and acts accordingly as explained in the previous section; additionally, there are some parameters that might be configured by the user using the interface presented in Figure 3: *resemblance measure* (the user can select



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			

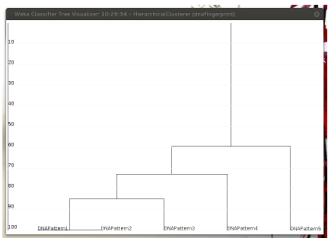


Fig. 4. Left: Image of DNA fingerprints. The image contains 5 DNA patterns (the 5 vertical lanes). In white, the bands of each pattern. **Centre:** Molecular weights of the bands of the image. **Right:** Dendrogram associated with the DNA patterns.

among the different resemblance measures available for each kind of data), *distance computation* (if the user selects a similarity measure S, she must fix how the distance measure D is computed from S using one of the following formulas $D = 1 - S$, $D = \sqrt{1 - S}$, or $D = \sqrt{1 - S^2}$), *matching* (if this option, only available for string and numerical data, is fixed as true, the set-occurrence mode is used; otherwise, the pairwise mode is applied), and *tolerance* (this parameter, only available for numerical data, and heterogeneous data that contain numerical attributes, allows the user to fix the closeness value).

4 Case Study: DNA fingerprinting

In this section, we present an application that illustrates some of the features supported by WekaBioSimilarity. The problem presented here cannot be handled using general-purpose software packages, and specific tools are required.

DNA fingerprinting [7] is a genetic typing technique that allows the analysis of the genomic relatedness between samples, and the comparison of DNA patterns. This technique has multiple applications in different fields (medical diagnosis, forensic science, parentage testing, food industry, agriculture and many others) [19]. The comparison of DNA fingerprints follows the workflow presented in Figure 1: (1) construction of feature vectors, (2) computation of similarity matrix, and (3) classification (using hierarchical clustering).

In the first stage, a feature vector for each DNA pattern is constructed. Each DNA pattern consists of a set of bands (see Figure 4), and in turn, each band has associated a numerical value called "molecular weight" (see Figure 4 for the molecular weights of the bands of Figure 4); the molecular weights of a DNA pattern correspond to its feature vector. In the second step, DNA patterns are compared considering the matchings of the molecular weights of their bands (i.e. a set-occurrence comparison is necessary). In this task, two bands are matched even if their molecular weights are not exactly the same, but they are close

enough; hence, a tolerance value is required — in our concrete example, we take a tolerance value of 3. Finally, the DNA patterns are grouped together using hierarchical clustering, and the result is visualised using a dendrogram.

Given a file containing the feature vectors of several DNA patterns, Weka combined with WekaBioSimilarity can generate the dendrogram of such patterns (see Figure 4) using the most common resemblance measures (Dice, Jaccard, and simple-matching [26]) and hierarchical clustering algorithms (UPGMA, single linkage, neighbour joining [26]) applied in this context.

Several software systems have been developed for DNA fingerprint analysis. Some of them internally implement the whole classification process generating a dendrogram as a final result. However, several systems (e.g. GelAnalyzer [10] or Dolphin 1D [28]) only generate the molecular weights of bands (i.e. Step (1)). In the latter packages, their output might be fed as input to WekaBioSimilarity to compare DNA patterns using the most common measures applied in this subject — as in the former packages, the final result will be a dendrogram.

5 Comparing Resemblance Measures

In the previous section, we have presented the application of WekaBioSimilarity in a context where some resemblance measures are considered as standards (works related to DNA fingerprinting mainly employ Dice, Jaccard, and simple-matching measures). In other situations, it is useful to explore different alternatives as shown in [27,2,24,6]; WekaBioSimilarity can also be used to this aim as we will illustrate in this section.

The statlog dataset [18] is a heart disease database that consists of 270 instances having 13 attributes (7 numeric, 3 binary, and 3 multi-value) — i.e. it is a heterogeneous dataset. From such a dataset, several classifiers can be trained in Weka to make predictions using supervised machine-learning algorithms. Some of those algorithms (e.g. k-nearest neighbour or locally-weighted learning) can work with the resemblance measures implemented in WekaBioSimilarity.

We consider the *ClassificationViaClustering* algorithm of Weka — a meta-classifier that uses clustering for classification. For this classifier, we can pick hierarchical clustering and try different distance measures. In particular, we have selected 3 similarity measures (Dice, Ochiai, and Sokal & Sneath, using the 3 different formulas available to compute the distance value) and 3 distance measures (Hamming, Vari, and Lance & Williams) of WekaBioSimilarity, and 3 measures (Chebyshev, Euclidean, and Manhattan) included by default in Weka. As can be seen in Table 2, different success rates are obtained when the resemblance measure is changed.

In this example, we have not pretended to be exhaustive in the analysis of the different measures for predicting heart diseases, but just show that several alternatives can be easily explored thanks to WekaBioSimilarity.

Measure	Success rate
Dice	81.11% $(1-S)$, 83.33% $(\sqrt{1-S})$, 55.19% $(\sqrt{1-S^2})$
Ochiai	81.11% $(1-S)$,81.11% $(\sqrt{1-S})$,82.22% $(\sqrt{1-S^2})$
Sokal & Sneath	54.82%$(1-S)$, 51.49% $(\sqrt{1-S})$, 59.26% $(\sqrt{1-S^2})$
Hamming	81.85%
Vari	79.63%
Lance & Williams	55.93%
Chebyshev	55.93%
Euclidean	80.37%
Manhattan	81.11%

Table 2. *Performance of several measures in the statlog dataset.*

6 Conclusions

The classification of objects is a common problem in several contexts, and it is highly dependent on computing the resemblance among the feature vectors of the objects. Usually, this task is carried out by special-purpose packages developed for each concrete problem. Weka is a general-purpose tool that offers several supervised and unsupervised machine-learning algorithms, but it only supported a few methods to obtain the similarity between numerical feature vectors. This drawback has been overcome with the WekaBioSimilarity package: a Weka extension that enhances this system with several resemblance measures and comparison modes. Namely, WekaBioSimilarity features: (1) the same (and even more) binary measures than other general-purpose statistical packages; (2) a generalisation of the binary measures to multi-value, string, numerical, and heterogeneous data; (3) a configurable tolerance parameter for numerical data; and, (4) two comparison modes: pairwise and set-occurrence. As far as we are aware, functionalities (2)–(4) have not been previously implemented in other general-purpose systems. As a result, we have a tool that can be applied in a wide variety of contexts either used as a standalone application, or integrated into other software packages.

References

1. Arif, M., Basalama, S.: Similarity-Dissimilarity Plot for High Dimensional Data of Different Attribute Types in Biomedical Datasets. International Journal of Innovative Computing, Information and Control 8(2), 1173–1181 (2012)
2. Boriah, S., Chandola, V., Kumar, V.: Similarity measures for categorical data: A comparative evaluation. In: Proceedings of the 8th SIAM International Conference on Data Mining. pp. 243–254 (2008)
3. Breese, J., Heckerman, D., Kadie, D.: Empirical analysis of predictive algorithms for collaborative filtering. In: Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (1998)
4. Choi, S.S., et al.: A survey of binary similarity and distance measures. Journal of Systemics, Cybernetics and Informatics 8(1), 43–48 (2010)
5. Hall, M., et al.: The weka data mining software: An update. SIGKDD Explorations 11(1), 10–18 (2009)

6. Hubálek, Z.: Coefficients of Association and Similarity, based on Binary (Presence-Absence) Data: An Evaluation. Biological Reviews 57(4), 669–689 (2008)
7. Jeffreys, A.J., Wilson, V., Thein, S.L.: Hypervariable 'minisatellite' regions in human DNA. Nature 314, 67–73 (1985)
8. Jurasinski, G., Retzer, V.: simba: A collection of functions for similarity analysis of vegetation data (2012)
9. Kurgan, L.A., et al.: Knowledge Discovery Approach to Automated Cardiac SPECT Diagnosis. Artificial Intelligence in Medicine 23(2), 149–169 (2001)
10. Lazar, I.: Gelanalyzer 2010a (2010), http://www.gelanalyzer.com/
11. Legendre, P., Legendre, L.: Numerical Ecology. Elsevier (1999)
12. Lichman, M.: UCI machine learning repository (2013), http://archive.ics.uci.edu/ml
13. MacArthur, R.: Geographical Ecology: Patterns in the Distribution of Species. Princeton University Press, New Jersey, USA (1984)
14. Manning, C., Schütze, H.: Foundations of statistical natural language processing. Cambridge, MA: The MIT Press (2001)
15. Michael, H.: Binary coefficients: A theoretical and empirical study. Mathematical Geology 8(2), 137–150 (1976)
16. Miyamoto, M., Cacraft, J.: Phylogenetic Analysis of DNA Sequences. Oxford University Press, Oxford, UK (1991)
17. Nei, M., Kumar, S.: Molecular Evolution and Phylogenetics. Oxford University Press, Oxford, UK (2000)
18. Nutt, C.L., et al.: Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. Cancer Research 63(7), 1602–1607 (2003)
19. Read, M.M. (ed.): Trends in DNA Fingerprint Research. Nova Science Publishers, Inc., New York, USA (2005)
20. Rettinger, A., et al.: Mining the Semantic Web. Data Mining and Knowledge Discovery 24, 613–662 (2012)
21. Rögnvaldsson, T., You, L., Garwicz, D.: State of the art prediction of HIV-1 protease cleavage sites. BioInformatics 31(8), 1204–1210 (2015)
22. Silva, T.C., Zhao, L.: Machine learning in complex networks. Springer (2016)
23. Sneath, P., Sokal, R.: Numerical Taxonomy: The Principles and Practice of Numerical Classification. W H Freeman & Co, San Francisco, USA (1973)
24. Spertus, E., Sahami, M., Buyukkokten, O.: Evaluating similarity measures: a large-scale study in the orkut social network. In: Proceedings of the 11th ACM SIGKDD conference on Knowledge discovery in data mining. pp. 678–684 (2005)
25. USDA, NRCS: The plants database (2008), http://plants.usda.gov
26. Vauterin, L., Vauterin, P.: Molecular Identification, Systematics, and Population Structure of Prokaryotes, chap. Integrated Databasing and Analysis. Springer-Verlag (2006)
27. Wang, X., et al.: Experimental comparison of representation methods and distance measures for time series data. Data Mining and Knowledge Discovery 26, 275–309 (2013)
28. Wealtec: Dolphin-1D software version 2.4 (2006), http://www.wealtec.com/products/imaging/software/dolphin-1d-software.htm
29. Willett, P.: Similarity-based approaches to virtual screening. Biochemical Society Transactions 31, 603–606 (2003)
30. Willett, P., Barnard, J.M., Downs, G.M.: Chemical Similarity Searching. Journal of Chemical Information and Computer Sciences 38, 983–996 (1998)
31. Xu, R., Wunsch, D.C.: Clustering. IEEE Computer Society Press (2008)