Subject Section

graphkernels: R and Python packages for graph comparison

Mahito Sugiyama 1,2,*, M. Elisabetta Ghisu 3,4, Felipe Linares-López 3,4, and Karsten Borgwardt 3,4*

1 National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan,
2 JST PRESTO, 4-1-8 Honcho, Kawaguchi, Saitama, 332-0012, Japan,
3 D-BSSE, ETH Zürich, Mattenstrasse 26, Basel, 4058, Switzerland,
4 Swiss Institute of Bioinformatics, Basel, 4058, Switzerland.

* To whom correspondence should be addressed.

Associate Editor: Dr. Jonathan Wren

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Measuring the similarity of graphs is a fundamental step in the analysis of graph-structured data, which is omnipresent in computational biology. Graph kernels have been proposed as a powerful and efficient approach to this problem of graph comparison. Here we provide graphkernels, the first R and Python graph kernel libraries including baseline kernels such as label histogram based kernels, classic graph kernels such as random walk based kernels, and the state-of-the-art Weisfeiler–Lehman graph kernel. The core of all graph kernels is implemented in C++ for efficiency. Using the kernel matrices computed by the package, we can easily perform tasks such as classification, regression and clustering on graph-structured samples.

Availability and Implementation: The R and Python packages including source code are available at https://CRAN.R-project.org/package=graphkernels and https://pypi.python.org/pypi/graphkernels

Contact: mahito@nii.ac.jp (R and C++); elisabetta.ghisu@bsse.ethz.ch (Python);

Supplementary information: Supplementary Material is available online at Bioinformatics.

1 Introduction

Graph-structured data is steadily growing and extensively being analyzed in computational biology. For example, chemical compounds are modeled as graphs in drug discovery (Takigawa and Mamitsuka, 2013), and proteins are represented as graphs in protein function prediction (Dhifli and Nguif, 2015). Finding efficient solutions for measuring the similarity between a pair of graphs, known as the graph comparison problem, is a fundamental step in graph analysis, in order to perform classification or regression on graph data. There are two approaches to graph comparison: alignment-based methods (Faisal et al., 2015) that compare graphs via finding node mappings and alignment-free methods (Yaveroğlu et al., 2015) that measure the similarity between graphs using features such as degree distributions or subgraph counts without identifying correspondences between nodes.

To date, among the alignment-free approaches, graph kernels have become a popular approach to quantify the similarity between graphs (Borgwardt and Kriegel, 2005; Costa and Grave, 2010; Gärtner et al., 2003; Kashima et al., 2003; Shervashidze et al., 2009, 2011; Vishwanathan et al., 2010; Sugiyama and Borgwardt, 2015), and are at the heart of many machine learning approaches in computational biology. A number of key applications of graph kernels exist such as biological function prediction from graph-based representations of chemical compounds. However, there is no convenient R or Python implementation that can simply and efficiently compute graph kernels, although R is a popular programming environment in Bioinformatics and Python in Machine Learning.

Here we present graphkernels, the first package in R and Python with efficient C++ implementations of various graph kernels including the following prominent kernel families: (1) simple kernels between vertex and/or edge label histograms, (2) graphlet kernels, (3) random walk kernels (popular baselines), and (4) the Weisfeiler-Lehman graph kernel (state-of-the-art kernels). The packages can be easily used to perform graph classification and regression by machine learning algorithms (Figure 1), such as support vector machines (SVMs) or the k-nearest neighbors algorithm (k-NN).

© The Author(s) 2017. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
Collection of graphs Kernel matrix

Fig. 1. Overview. The kernel value $K_{ij}$ represents the similarity between graphs $i$ and $j$.

2 Methods

Each function implemented in the graphkernels packages receives a collection of graphs $G_1, G_2, \ldots, G_n$ and returns the kernel (Gram) matrix $(K_{ij}) \in \mathbb{R}^{n \times n}$ with the respective graph kernel, where each kernel value $K_{ij}$ shows the similarity between graphs $G_i$ and $G_j$. The packages support the following 13 graph kernels:

- Linear kernels on label histograms: $\text{VertexHist}$, $\text{EdgeHist}$, $\text{VertexVertexEdgeHist}$, $\text{VertexEdgeHist}$, $\text{EdgeHistGauss}$, $\text{VertexEdgeHistGauss}$.
- Gaussian RBF kernels on label histograms: $\text{VertexHistGauss}$, $\text{EdgeHistGauss}$, $\text{VertexEdgeHistGauss}$.
- Graphlet kernels: $\text{Graphlet}$, $\text{ConnectedGraphlet}$.
- Random walk based kernels: $\text{KStepRandomWalk}$, $\text{ConnectedGraphlet}$, $\text{GeoRandomWalk}$, $\text{ExponentialRandomWalk}$, $\text{EdgeHist}$, $\text{VertexHist}$, $\text{VertexHistGauss}$.
- The Vapnik-Chernovenkis entropic graph kernel: $\text{VertexVertexEdgeHist}$, $\text{VertexHist}$, $\text{VertexHistGauss}$.

All kernels are implemented in C++ and compiled through the packages Rcpp (Bates, 2013) and RcppEigen (Bates and Eddelbuettel, 2013) in the R package. We use SWIG (Beazley, 1996) interfaces to wrap C++ code in Python. See the Supplementary Material for detailed mathematical definitions of these graph kernels.

In our packages, each graph is treated as an igraph object (Csardi and Nepusz, 2006) and a collection of graphs is kept as a list of igraph graphs. An example usage in R is in the following, where we use the dataset MUTAG (Debnath et al., 1991), a typical benchmark dataset that is also provided in our package.

```r
> library(graphkernels) # load the package
> data(mutag) # load a sample dataset,
> which is a list of (igraph) graphs
> mutag[1]

```

**Fig. 2.** Accuracy (left) and running time (in seconds, right) on the MUTAG dataset.

3 Application

As a representative application, we demonstrate graph classification using the MUTAG dataset. In the dataset, there are 188 graphs, and the objective is to predict labels of graphs, indicating whether or not they are mutagenic. We used 10-fold cross validation for graph classification. Randomly divided the entire dataset into ten folds then, in each iteration, one of the ten folds was used for testing and the rest for training. We computed the kernel matrix of the training data using one of our functions implemented in the graphkernels, and use these data to train an SVM using the kernlab package (Karatzoglou et al., 2005). We then predicted labels on the test data, and obtained the accuracy by comparison with the ground-truth labels. The detailed experimental methodology and the R code to reproduce these results are provided in the Supplementary Material.

Figure 2 shows the prediction accuracy for graph kernels in our package and the CPU running time needed to compute each kernel matrix. This examples demonstrates that our package allows for an easy comparison of the effectiveness and the efficiency of various popular graph kernels and will serve as a baseline when designing graph kernels for specialized applications in computational biology.

Funding

This work has been supported by JSPS KAKENHI Grant Numbers JP16K16115 and JP16H02870 (MS).

References


