© 2013 FIZ Karlsruhe

io-port 05930869

Hinselmann, Georg; Jahn, Andreas; Fechner, Nikolas; Rosenbaum, Lars; Zell, Andreas

Approximation of graph kernel similarities for chemical graphs by kernel principal component analysis.

Pizzuti, Clara (ed.) et al., Evolutionary computation, machine learning and data mining in bioinformatics. 9th European conference, EvoBIO 2011, Torino, Italy, April 27–29, 2011. Proceedings. Berlin: Springer (ISBN 978-3-642-20388-6/pbk). Lecture Notes in Computer Science 6623, 123-134 (2011).

Summary: Graph kernels have been successfully applied on chemical graphs on small to medium sized machine learning problems. However, graph kernels often require a graph transformation before the computation can be applied. Furthermore, the kernel calculation can have a polynomial complexity of degree three and higher. Therefore, they cannot be applied in large instance-based machine learning problems. By using kernel principal component analysis, we mapped the compounds to the principal components, obtaining q-dimensional real-valued vectors. The goal of this study is to investigate the correlation between the graph kernel similarities and the similarities between the vectors. In the experiments we compared the similarities on various data sets, covering a wide range of typical chemical data mining problems. The similarity matrix between the vectorial projection was computed with the Jaccard and Cosine similarity coefficient and was correlated with the similarity matrix of the original graph kernel. The main result is that there is a strong correlation between the similarities of the vectors and the original graph kernel regarding rank correlation and linear correlation. The method seems to be robust and independent of the choice of the reference subset with observed standard deviations below 5%. An important application of the approach are instance-based data mining and machine learning tasks where the computation of the original graph kernel would be prohibitive.

doi:10.1007/978-3-642-20389-3_12