## Whole-neighbourhood similarity measures permit big-data analysis with controllable bias

Tom Lorimer[1], Ruedi Stoop[1,2]

[1] Institute of Neuroinformatics, University of Zurich and ETH Zurich, Switzerland
[2] Institute for Computational Science, University of Zurich, Switzerland

A first step in the analysis of data typically is to look for structure. This often begins with the search for ‘clusters', sets of data points made up by ‘similar’ data. Recent work has revealed, however, that many popular clustering algorithms identify clusters that are inconsistent with real-world data structures. This is because cluster tags do not contain the full relevant data structure, and because even advanced present day data clustering approaches modify the underlying data structures in a largely unpredictable manner. We present a novel framework based on graphs, for the identification and representation of arbitrary high dimensional data structure. In the approach, we apply whole-neighbourhood non-iterative measures of similarity between nodes on a k-nearest-neighbour graph over the data. In this way, we can prune non-representative edges in a tuneable manner, taking into account specific interest that an observer has in the data. From the resulting graphs, standard network characteristics can be extracted. In conjunction with a novel visualisation scheme, this approach is fully general, transparent, controllable and interpretable. We validate our approach on synthetic and natural data sets, exhibiting different strengths of the proposed method.



## K-means clustering using an improved firefly algorithm applied to real world data sets

Masaki Takeuchi[1], Thomas Ott[2], Haruna Matsushita[1], Yoko Uwate[1], Yoshifumi Nishio[1]

[1] Tokushima University, Japan
[2] Institute of Applied Simulation, ZHAW Zurich University of Applied Sciences, Switzerland

K-means clustering is a type of unsupervised learning, which is well used when the number of clusters is known and the clusters tend to be spherical. The goal of the K-means algorithm is to find K cluster centers and assign each object to the closest cluster center such that the sum of the squared distances between the objects and the corresponding cluster centers is minimal. This means that the K-means clustering problem is a minimum optimization problem.
The firefly algorithm (FA) has been proposed by Yang and is idealized the flashing characteristics of fireflies. FA is an efficient optimization algorithm because it has a deterministic component and a random component. In 2011, Senthilinath et al. proposed a new algorithm combined the firefly algorithm and the K-means algorithm (KMFA). Numerical experiments have indicated that this algorithm is more efficient than the K-means algorithm or other improved algorithms.
In this study, we propose a new algorithm that used improved firefly algorithm for the K-means clustering (KMIFA). In this algorithm, each firefly has its own new parameter. The new parameter has a following effects on the movement of firefly. At the beginning of the search, all fireflies move with a relatively strong random influence. Hence, they can more easily escape from local optima. As number of iterations increases, the firefly does not move randomly. Therefore, the firefly tends to converge. We compare the K-means algorithm, KMFA and KMIFA using several data sets. These experiments indicate that our algorithm is more efficient than the other algorithms.