Overcoming Key Weaknesses of Distance-based Neighbourhood Methods using **a Data Dependent Dissimilarity Measure**

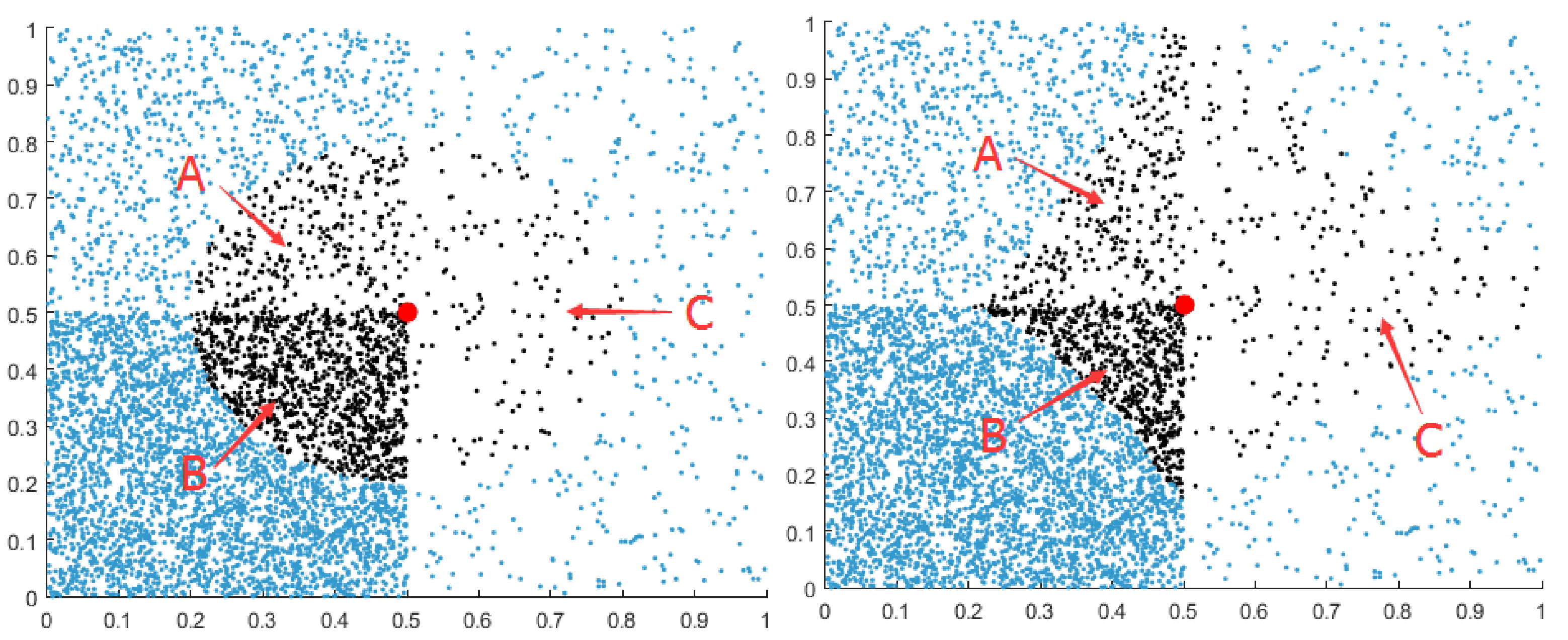
Kai Ming Ting[1]; Ye Zhu[2]; Mark Carman[2]; Yue Zhu[3]; Zhi-Hua Zhou[3]
[1]Federation University, [2]Monash University, [3]Nanjing University

KDD2016
22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining
August 13 - 17, 2016 | San Francisco, California

Introduction

The distance calculation is the core process that has been applied to all aspects of data mining tasks, including density estimation, clustering, anomaly detection and classification. Despite its widespread applications, research in psychology has pointed out since 1970's that distance measures do not possess the key property of dissimilarity as judged by humans, i.e., the characteristic where **two instances in a dense region are less similar to each other than two instances of the same interpoint distance in a sparse region**.

This project introduces the first generic version of data dependent dissimilarity and shows that it provides a better closest match than distance measures for three existing algorithms in clustering, anomaly detection and multi-label classification. For each algorithm, we show that **simply replacing the distance measure with a data-dependent dissimilarity measure, overcomes a key weakness of the otherwise unchanged algorithm**.



(a) ε-neighbourhood region (b) μ-neighbourhood region

Figure 1: ε-neighbourhood density vs μ-neighbourhood mass

$\boldsymbol{\varepsilon}$-neighbourhood density function:
$N_\varepsilon = \#\{y \in D \mid l_p(x,y) \le \varepsilon\}$, volume is constant and region is symmetrical

$\boldsymbol{\mu}$-neighbourhood mass function:
$M_\mu = \#\{y \in D \mid m_e(x,y) \le \mu\}$, volume is unfixed and region is asymmetrical

Mass-based Dissimilarity

Let D be a data sample from pdf (probability density function) F; and $\mathrm{H} \in \mathcal{H}(D)$ be a hierarchical partitioning model of the space into non-overlapping and non-empty regions. The definitions for the domain of $\mathbb{R}^d$ are given as follows.

Definition 1. $R(x,y|H;D)$ is the smallest local region covering x and y wrt H and D is defined as:

$$R(x,y|H;D) = arg \min_{r \subset H|\, s.t. \{x,y\} \in r} \sum_{z \in D} \mathbf{1}(z \in r)$$

where $\mathbf{1}(.)$ is an indicator function.

Definition 2. Mass-based dissimilarity of x and y wrt D and F is defined as the expected probability of $R(x,y|H;D)$:

$$m(x,y|D,F) \ = \ E_{\mathcal{H}(D)}[P_F\ (R(x,y|H;D))]$$

where $P_F(.)$ is the probability wrt F; and the expectation is taken over all models in $\mathcal{H}(D)$.
In practice, the mass-based dissimilarity would be estimated from a finite number of models $H_i \in \mathcal{H}(D), i \ = 1, \ldots, t$ as follows:

$$m_e(x,y|D) = \frac{1}{t} \sum_{i=1}^{t} \tilde{P}(R(x,y|H_i;D))$$

where $\tilde{P}(R) = \frac{1}{|D|} \sum_{z \in D} \mathbf{1}(z \in R)$.

Note that $R(x,y|H_i;D)$ is the smallest local region covering x and y, it is analogous to the shortest distance between x and y used in the geometric model.

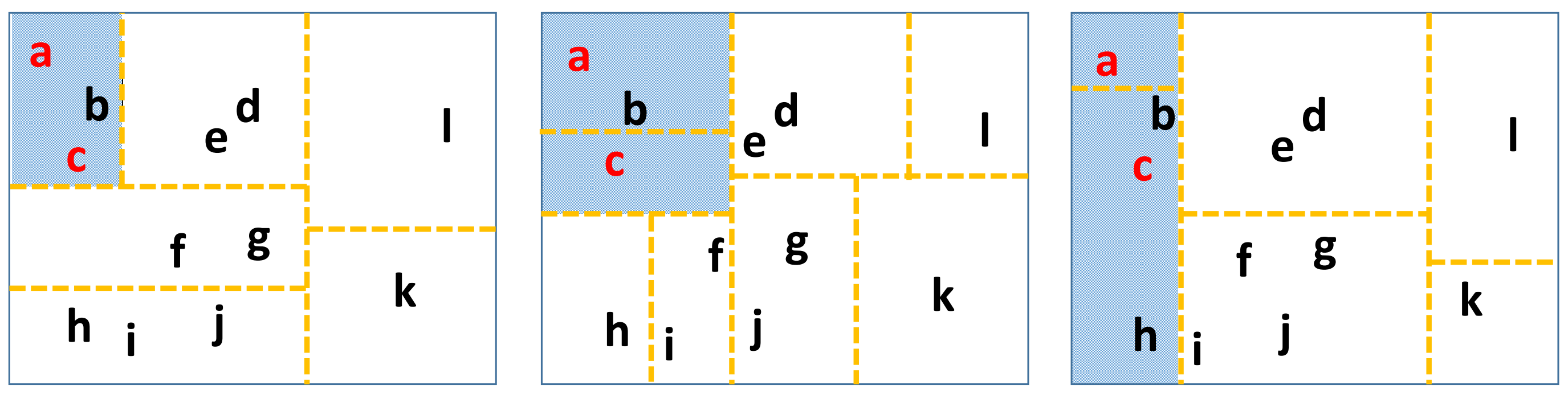


Figure 2: A 2 dimensional data is partitioned by a 3-level **Isolation Forest (iForest)** with 3 independent iTrees. $R(a,c)$ is shaded for each partitioning model. $m_e(a,c) = (3+3+4)/12/3$=0.28 in this case.

Fundamental change in perspective in finding closest match neighbourhood

- From **nearest neighbour** to **lowest probability mass neighbour**
- Lowest probability mass represents the most similar

Density-based/distance-based	Mass-based
k-nearest neighbour	k-lowest probability mass neighbour
DBSCAN (density-based method)	MBSCAN (mass-based method)

Implication of this work

- Dissimilarity measures are **assumed to be a metric as a necessary criterion** for all data mining tasks.
- This work shows for the first time that this assumption can **impede the development of better performing models**.

Empirical Evaluation

Application 1: Density-based Clustering

Enable a clustering algorithm detect varying density clusters with a global threshold.

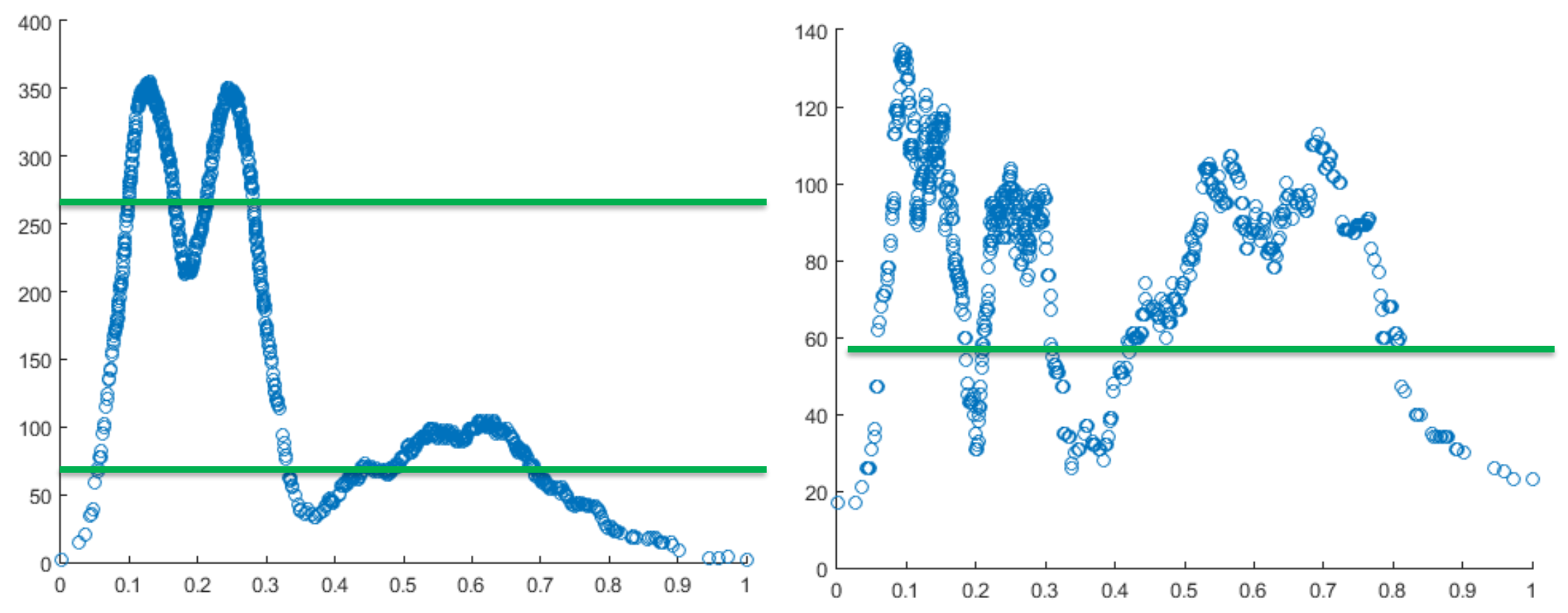


Figure 3: ε- neighbourhood density vs μ- neighbourhood mass on a "hard distribution" data.

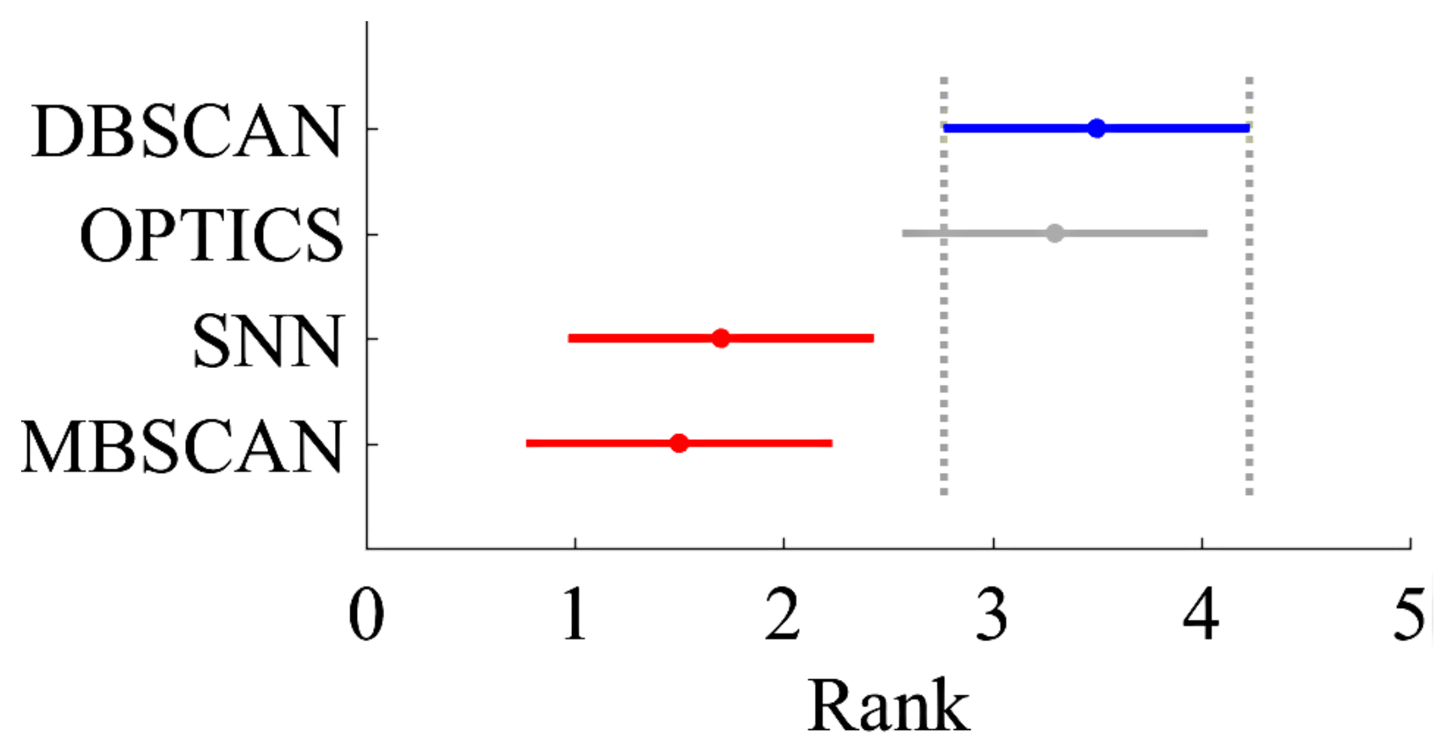


Figure 4: Rank of clustering algorithms on 10 datasets.

Application 2: Anomaly Detection

kNN is unable to detect local anomalies. Yet, Mass-based kNN is able to detect all fringe instances of the dense cluster as anomalies.

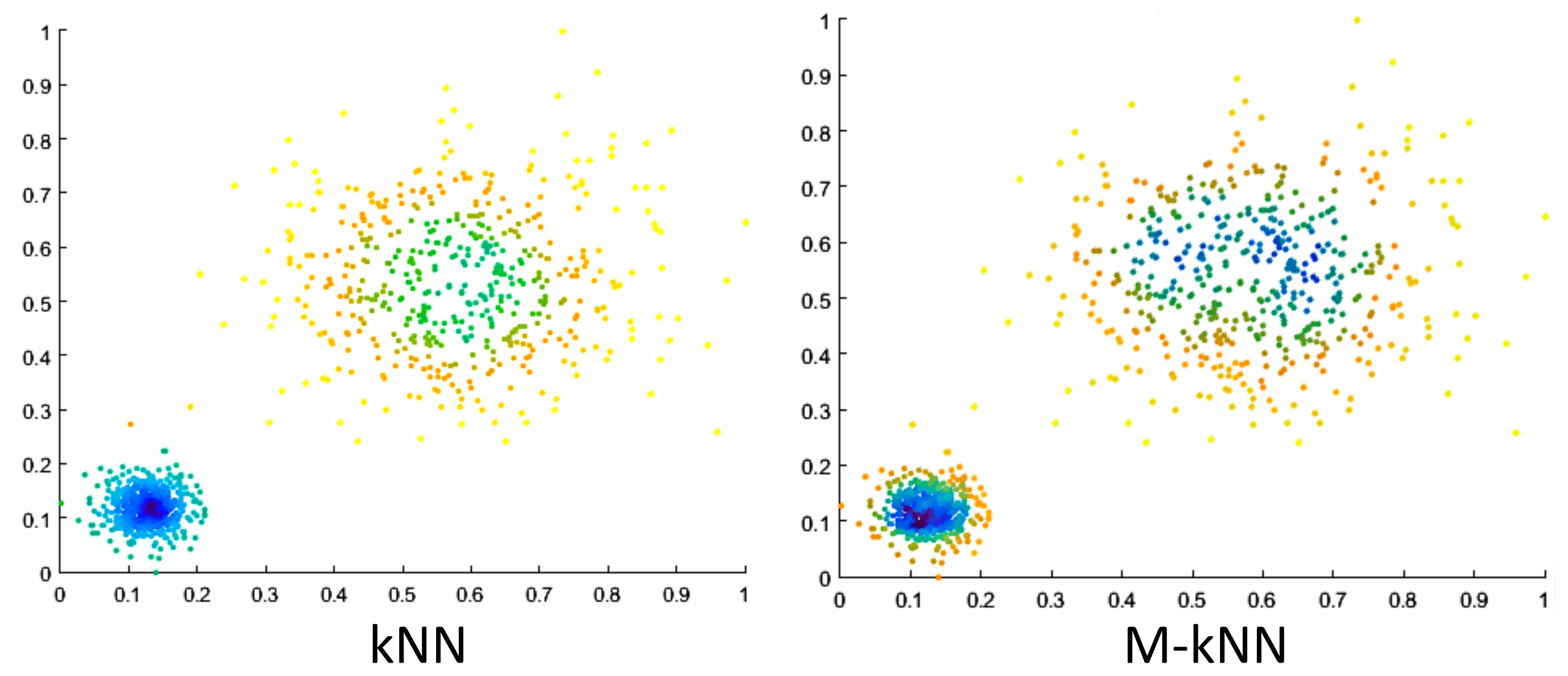


Figure 5:The ability to detect local anomalies in the dense cluster.

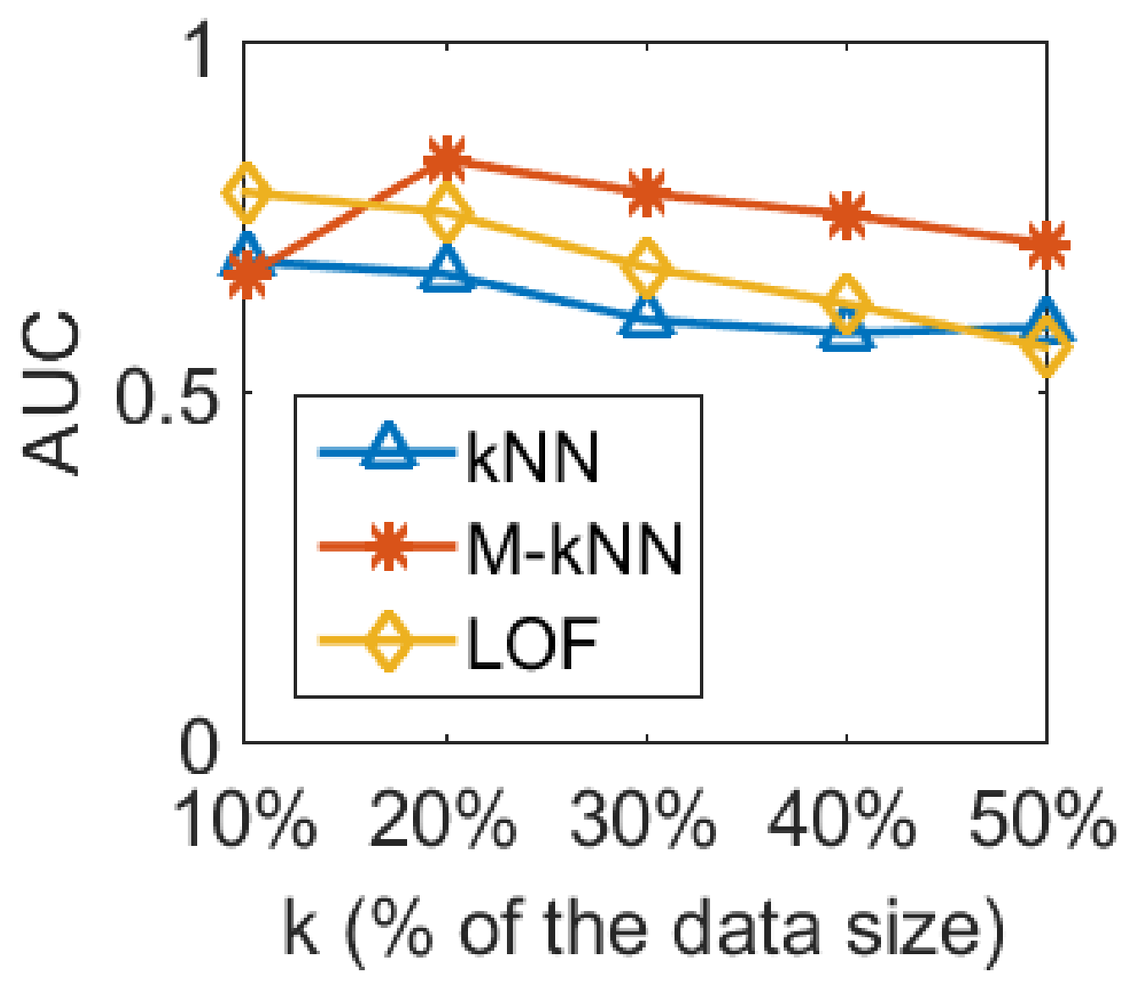


Figure 6: AUC on BloodDonation

Application 3: Multi-label classification

MLkNN vs M-MLkNN: instances of different labels are easier to separate using m_e than distance.

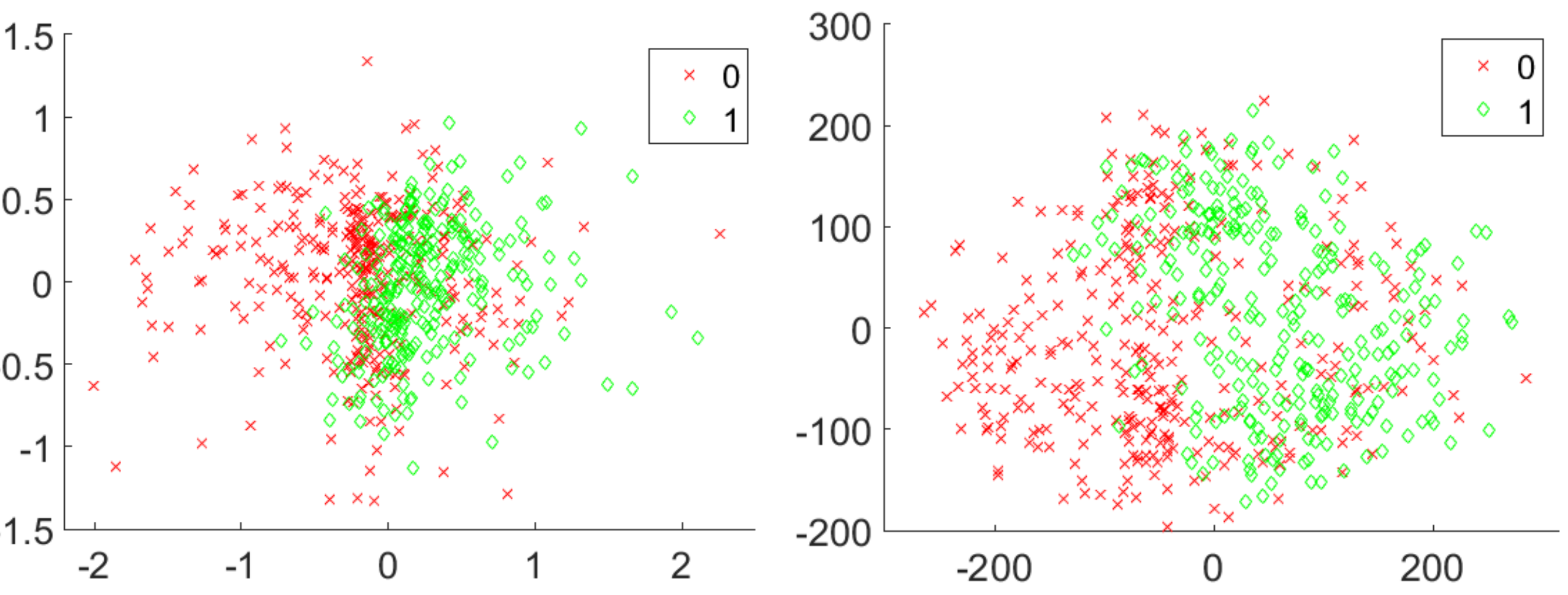


Figure 7: MDS plots using distance matrix and mass-based dissimilarity matrix on the Emotions dataset. Green and red points represent the positive and negative instances of the majority label, respectively.

Conclusions

- The data dependent dissimilarity overcomes key weaknesses of three existing algorithms that rely on distance, and **effectively improves their task-specific performance on density-based clustering, kNN anomaly detection and multi-label classification**.
- These existing algorithms are transformed by **simply replacing the distance measure with the mass-based dissimilarity**, leaving the rest of the procedure unchanged.
- As the transformation heralds a fundamental change of perspective in finding the closest match neighbourhood, **the converted algorithms are more aptly called lowest probability mass neighbour algorithms than nearest neighbour algorithms**.

The source code of mass-based dissimilarity can be obtained at https://sourceforge.net/projects/mass-based-dissimilarity/.

Contact: Kai Ming Ting, Federation University, Australia
kaiming.ting@federation.edu.au

Federation
UNIVERSITY · AUSTRALIA
MONASH University