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 Australia, [2]Monash University, [3]Nanjing University

Contents

1. Introduction

   Many weaknesses of data mining algorithms are due to a root problem, i.e., the use of distance measure.
2. Data-dependent dissimilarity is one solution to the root problem.
3. Evidence in three tasks: density-based clustering, anomaly detection and multi-label classification
4. A change in perspective and its implications
5. Conclusions

Known weaknesses of existing algorithms

- Density-based clustering algorithms have difficulty in detecting all clusters of varying densities
- K nearest neighbour anomaly detectors cannot detect local anomalies
- K nearest neighbour multi-label classifier has poor likelihood estimation

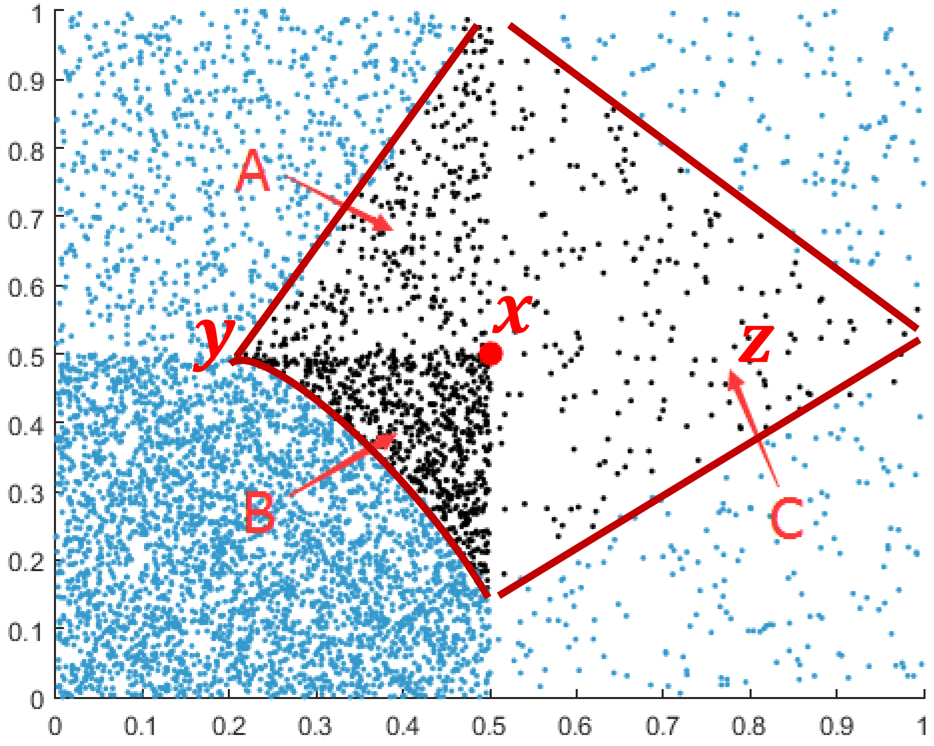
What is common to these algorithms?

They all use distance measure

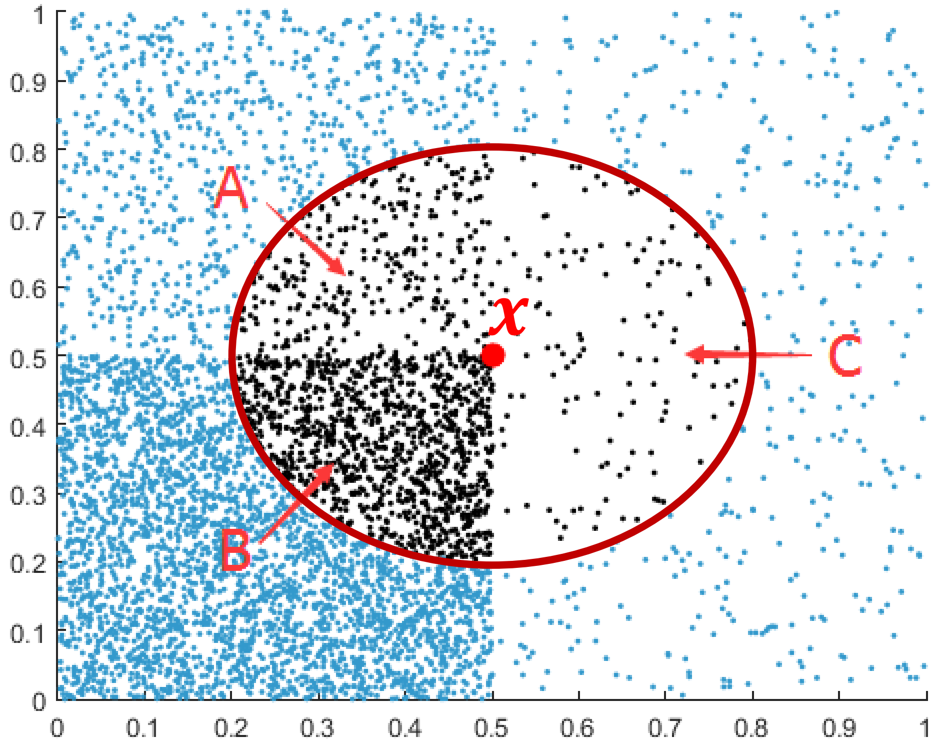
- Compute the dissimilarity of two points solely based on their geometric positions.
- A data independent measure, i.e., it produces the same dissimilarity for any two points of equal interpoint distance regardless of the data distribution.
- We identify that **distance measure is the root cause of the weaknesses of the three algorithms**.

Solution to the root problem

Use **data dependent dissimilarity** rather than **data independent distance measure**



$$M_\mu(x) = \#\{y \in D \mid \underline{m_e(x,y)} \le \mu\}$$



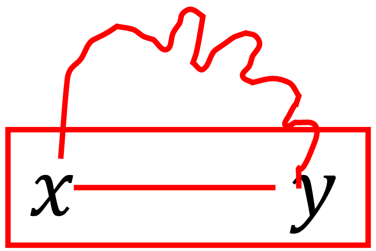
$$N_\epsilon(x) = \#\{y \in D \mid \underline{\ell_p(x,y)} \le \epsilon\}$$

Data-dependent dissimilarity

- Compute the dissimilarity between two points based primarily on the data distribution around and between them.
- Two points in the sparse region is more similar to each other than two points of the same inter-point distance in the dense region.
- Simply replacing the distance measure with the data-dependent dissimilarity overcomes the key weaknesses of density-based clustering, kNN anomaly detector and kNN multi-label classifier, particularly in data with varying densities.

Data-dependent dissimilarity : Generic definition

- An extension of **mass estimation** (Ting et al, KDD2010, Chen et al, MLJ2015) of one point to a dissimilarity of two points.
- A general definition of data dependent dissimilarity in which m_p-dissimilarity (Aryal et al, ICDM2014) is a special case.
- Analogous to the **shortest distance** between x and y used in the distance measure, data-dependent dissimilarity uses the **smallest local region** covering x and y in model H generated from sample D, i.e., $R(x, y|H; D)$



Definitions of Data-dependent dissimilarity (1)

Let D be a data sample from pdf (probability density function) F; and $H \in \mathcal{H}(D)$ be a hierarchical partitioning model of the space into non-overlapping and non-empty regions.

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

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

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 a random data point would lie in region $R(x, y|H; D)$:

$$m(x, y|D, F) = E_{\mathcal{H}(D)}[\underline{P_F\,(R(x, y|H; D))}] \qquad (2)$$

where $P_F(.)$ is the probability wrt F.

Definitions of Data-dependent dissimilarity (2)

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:

$$m_e(x, y|D) = \frac{1}{t} \sum_{i=1}^{t} \tilde{P}(R(x, y|H_i; D)) \qquad (3)$$

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

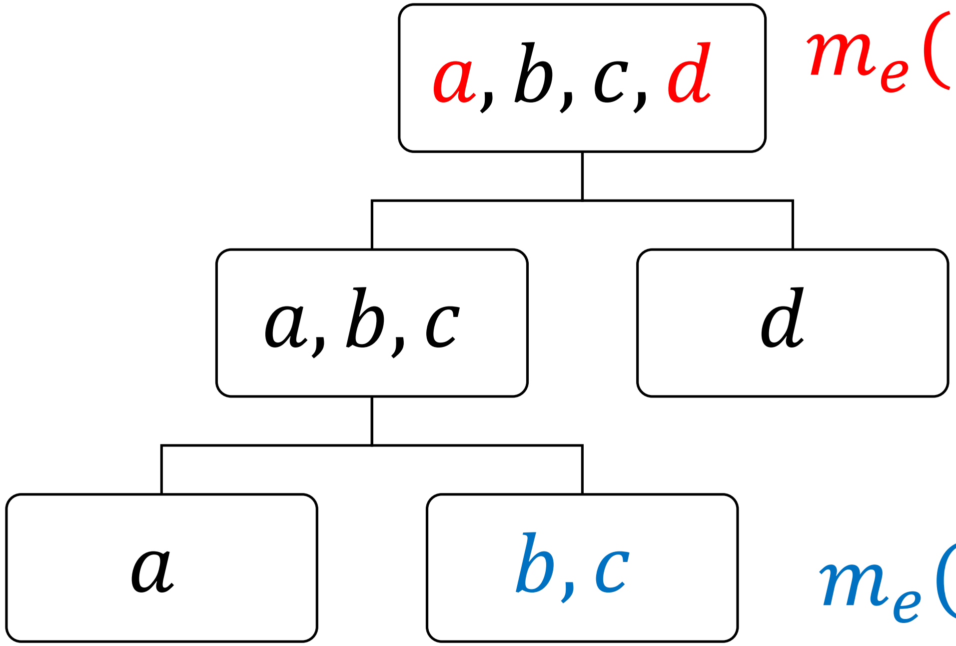
Implementation using Isolation Forest (iForest)

We use a recursive partitioning scheme called iForest (Liu et al, 2008), consisting of t iTrees as the partitioning structure R, to define regions.

Test points x and y are parsed through each iTree to calculate the mass of the lowest node containing both x and y, i.e., $|R(x, y|H)|$. Finally, $m_e(x, y)$ is the mean of these mass values over t iTrees as defined below:

$$m_e(x, y) = \frac{1}{t} \sum_{i=1}^{t} \frac{|R(x, y|H_i)|}{|D|} \tag{4}$$

Implementation : An Example of iTree



$m_e(a, d) = 4$

$m_e(b, c) = 2$

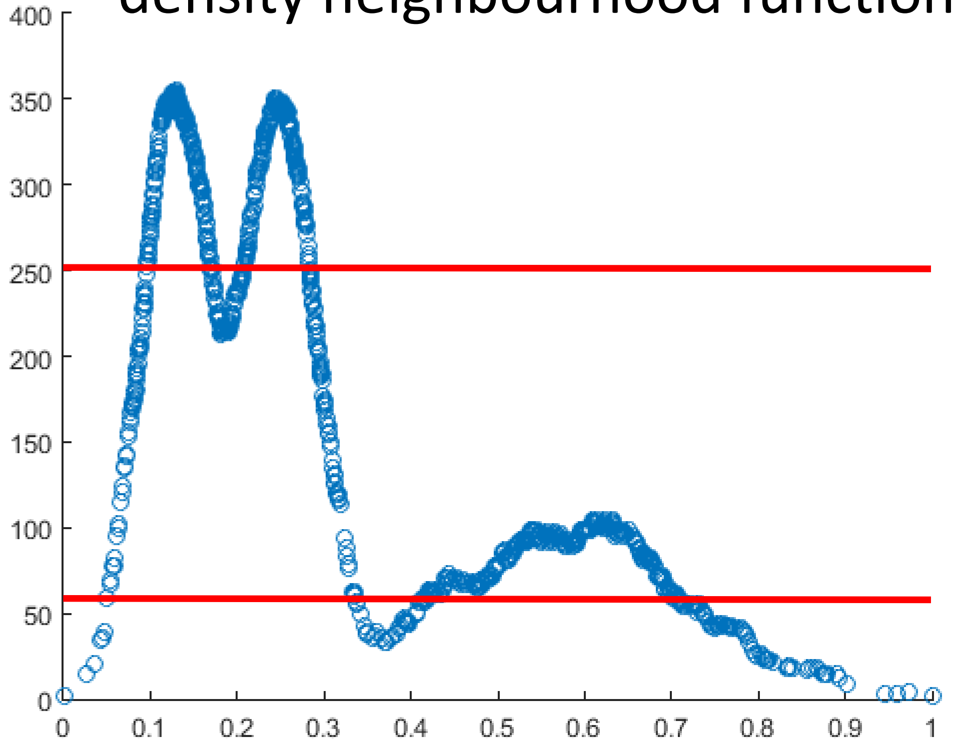
Self-dissimilarity is not constant:

$$m_e(a, a) = m_e(d, d) = 1,$$
$$m_e(b, b) = m_e(c, c) = 2.$$

Four instances partitioned by a 2-level iTree.

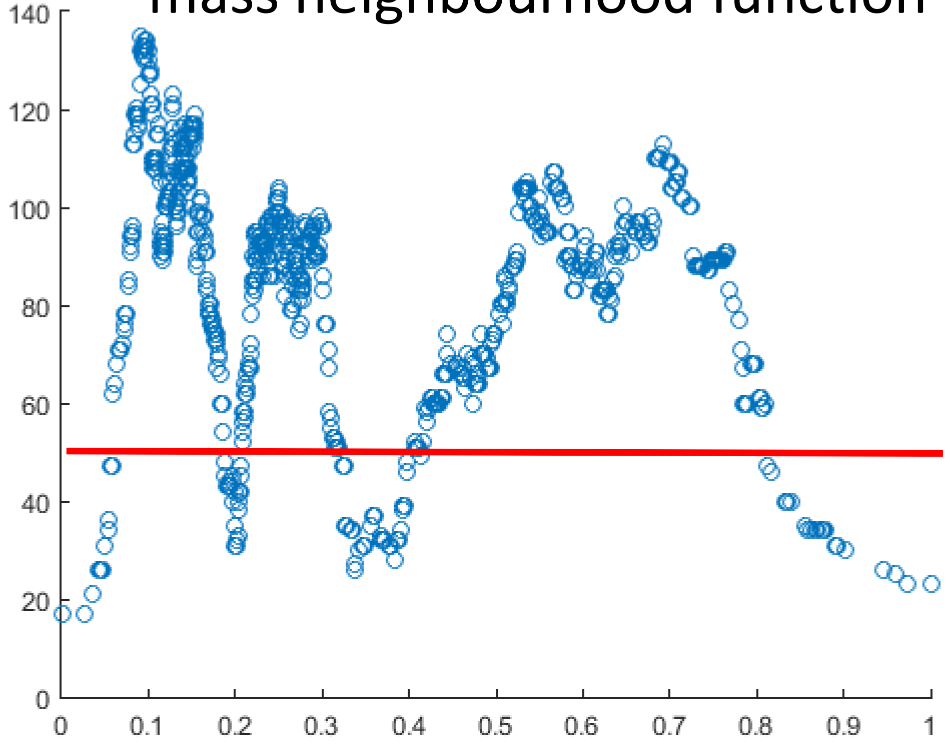
Application 1: Density-based clustering (a)
DBSCAN is unable to find all clusters of varying densities

Density distribution due to density neighbourhood function



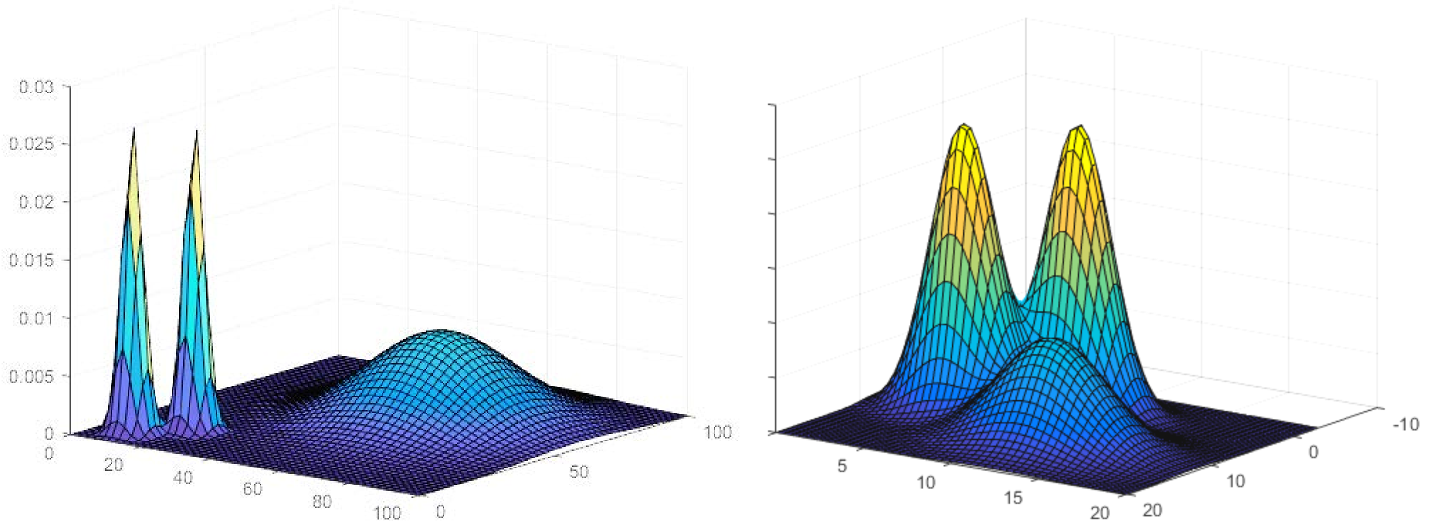
$$N_\epsilon(x) = \#\{y \in D \mid \ell_p(x,y) \le \epsilon\}$$

Mass distribution due to mass neighbourhood function



$$M_\mu(x) = \#\{y \in D \mid m_e(x,y) \le \mu\}$$

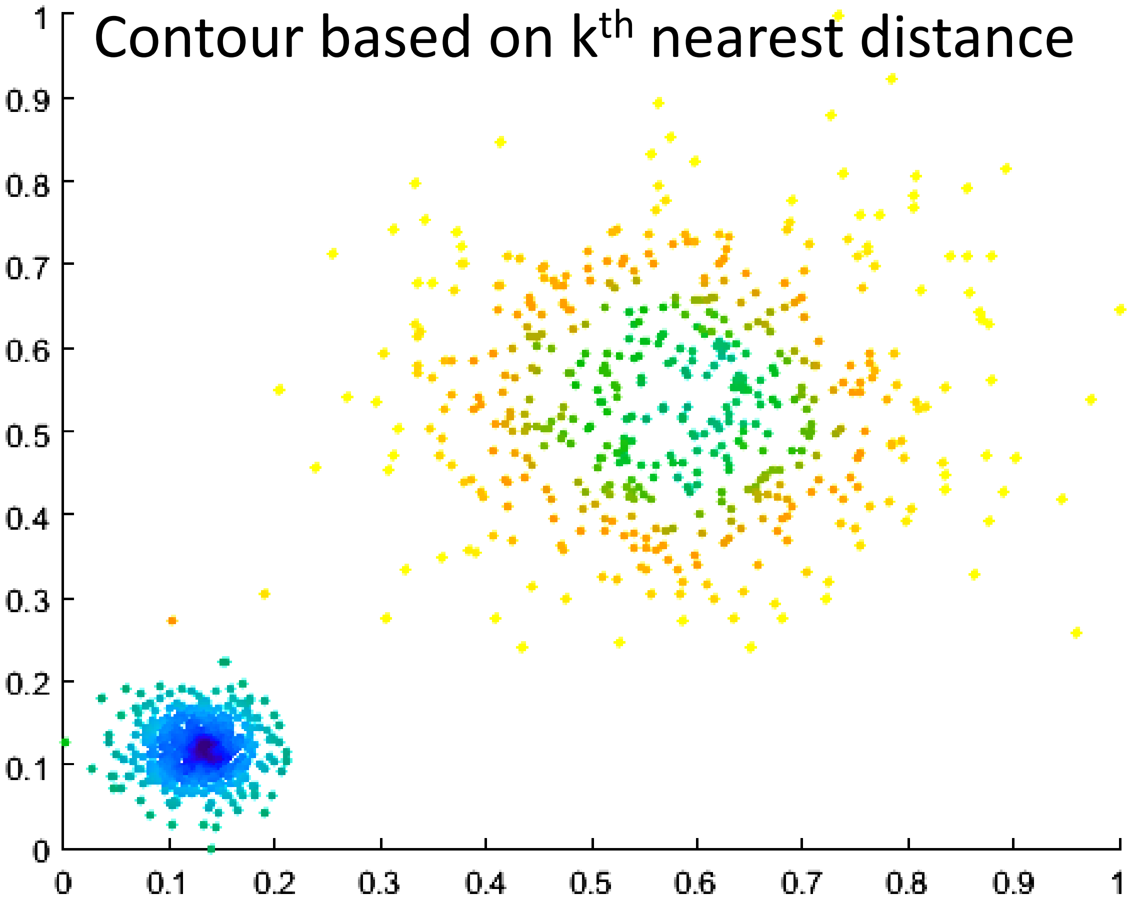
Application 1: Density-based clustering (b)
DBSCAN is unable to find all clusters of varying densities



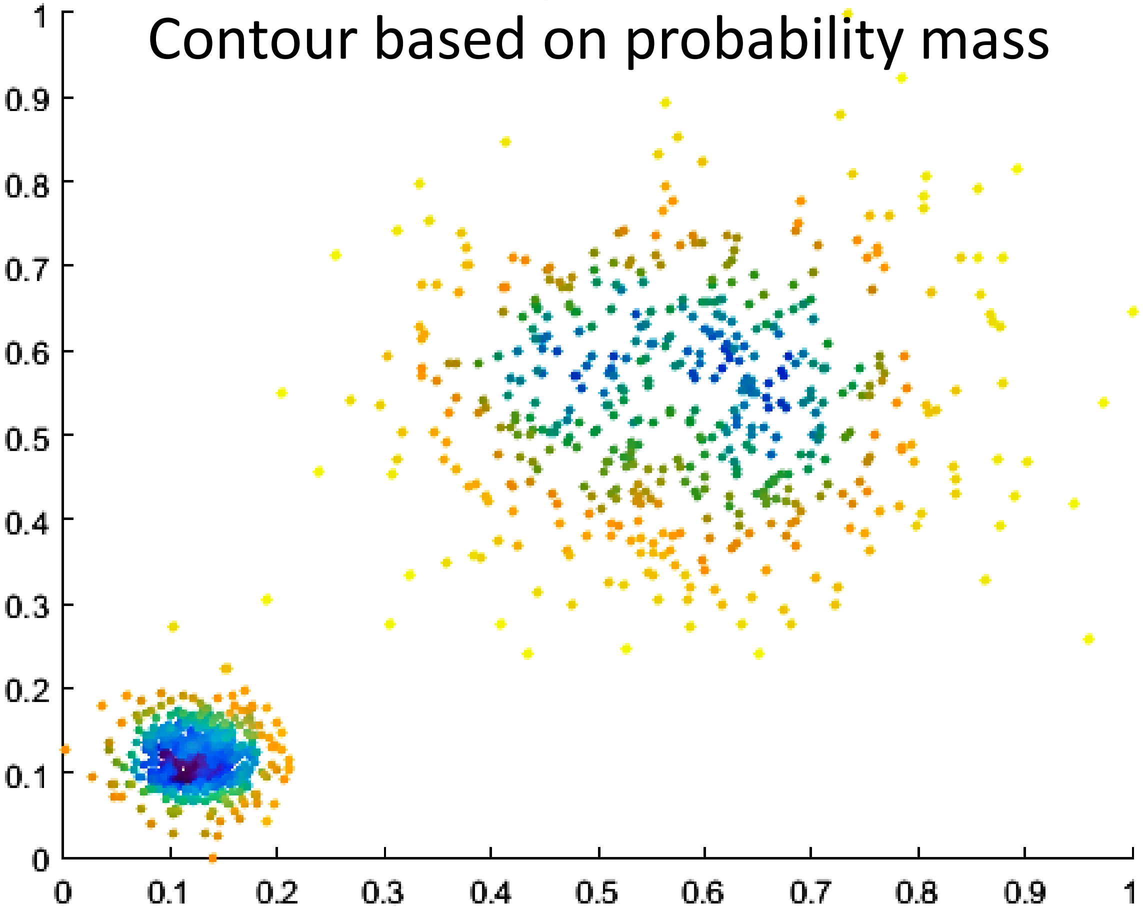
	Easy Distribution	Hard Distribution
DBSCAN (using distance measure)	0.94	0.34
DBSCAN (using mass-based dissimilarity)	0.993	0.62

Clustering results in terms of F1-measure

Application 2: kNN anomaly detector - unable to detect local anomalies



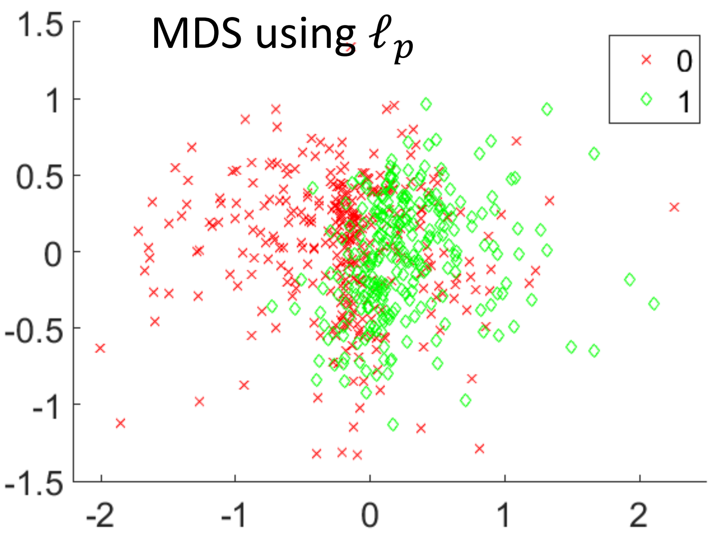
kNN using distance measure

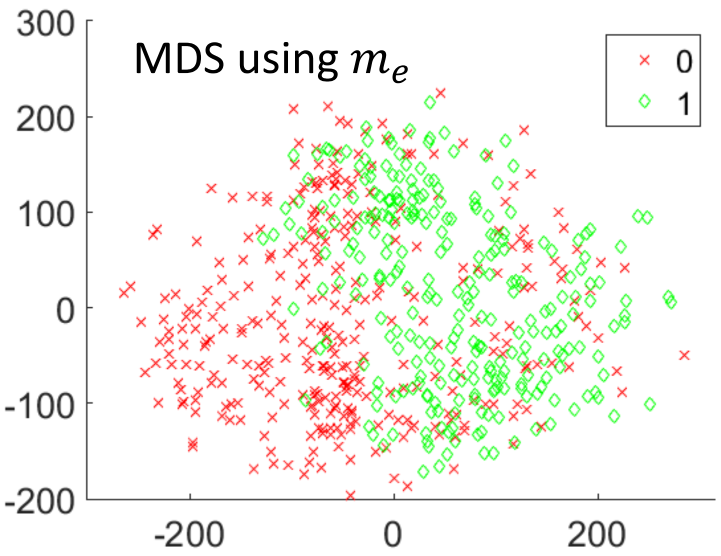


kNN using mass-based dissimilarity

Application 3: Multi-Label Classification
MLkNN - poor likelihood estimation in varying densities

An example using Multi-Dimensional Scaling (MDS) plot on the Emotion data set. Green and red points represent the positive and negative instances of the majority label, respectively.





	Birds	CAL500	Emotions	Enron	Scene
MLkNN (ℓ_p)	0.392	0.489	0.692	0.604	0.774
MLkNN (m_e)	0.600	0.489	0.776	0.640	0.794

Classification result in terms of Average Precision

Runtime comparison: Dissimilarity matrix calculations

Data set (Data size) (#Dimenisons)	Segment (2310) (19)	Pendigit (10992) (16)	P53Mutant (10387) (5408)	Time complexity
Euclidean distance	5	110	8182	$O(n^2d)$
Mass-based dissimilarity	31	600	548	$O(n^2C)$
SNN-similarity	26	573	9141	$O(n^2k^2+n^2d)$

Time in seconds (n:data size, d:#dimensions, C:constant, k: parameter in kNN)

A fundamental change in perspective

- Finding closest match neighbourhood:
  Change from **nearest neighbour**
  to **lowest probability mass neighbour**
- Lowest probability mass neighbours represent the most similar neighbours

Distance-based or Density-based	Mass-based
k-nearest neighbour →	k-lowest probability mass neighbour
DBSCAN (density-based method) →	MBSCAN (mass-based method)

Implications of this work

Dissimilarity measures are assumed to be a metric as a necessary criterion for all data mining tasks.

This work shows that this assumption can be an impediment to producing good performing models in three tasks: clustering, anomaly detection and multi-label classification.

Conclusions

- **The proposed 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 each 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**, since the lowest mass neighbours represent the most similar neighbours.

Acknowledgments

Funding supports:

- US Air Force Office of Scientific Research, Asian Office of Aerospace Research & Development under award number: #FA2386-13-1-4043 and 111 Project (B14020) (Kai Ming Ting);
- NSFC (61333014) and the Collaborative Innovation Center of Novel Software Technology and Industrialization (Zhi-Hua Zhou).

Thank you