MONASH University
Information Technology

Mass Estimation and its applications

Kai Ming Ting, Guang- Tong Zhou, Tony Liu, James Tan
Gippsland School of Information Technology
Monash University
Australia

Background

- Density = Mass / Area
- Given equal- size grid, density = mass
- Use density to solve various tasks in data mining
- Unlike physics, we do not use mass
- This work is the first attempt to use mass to solve data mining problems
- We show that mass can solve three data mining tasks more effectively and efficiently than density

Contents

- **Density Estimation**
- **Mass & Mass Estimation**
- **Density Estimation vs Mass Estimation**
- **How mass can be applied**
  - Direct Approach & Indirect Approach
- **Results in Regression, Content- based Image Retrieval and Anomaly Detection**
- **Characteristics of Mass Estimation**
- **Contributions**
- **Final Remark and Future Work**

Density Estimation

'..estimation of densities is a universal problem of statistics (knowing the densities one can solve various problems.)'
[Vapnik, 2000]

Examples:

- **Class- conditional density function p(x|c)**
- **Posterior probability p(c|x)**
- **Density- based, distance- based methods**

Kernel density estimation, kNN, Maximum Likelihood procedures or Bayesian methods

What Mass Estimation is not

- Probability mass function
- Probability

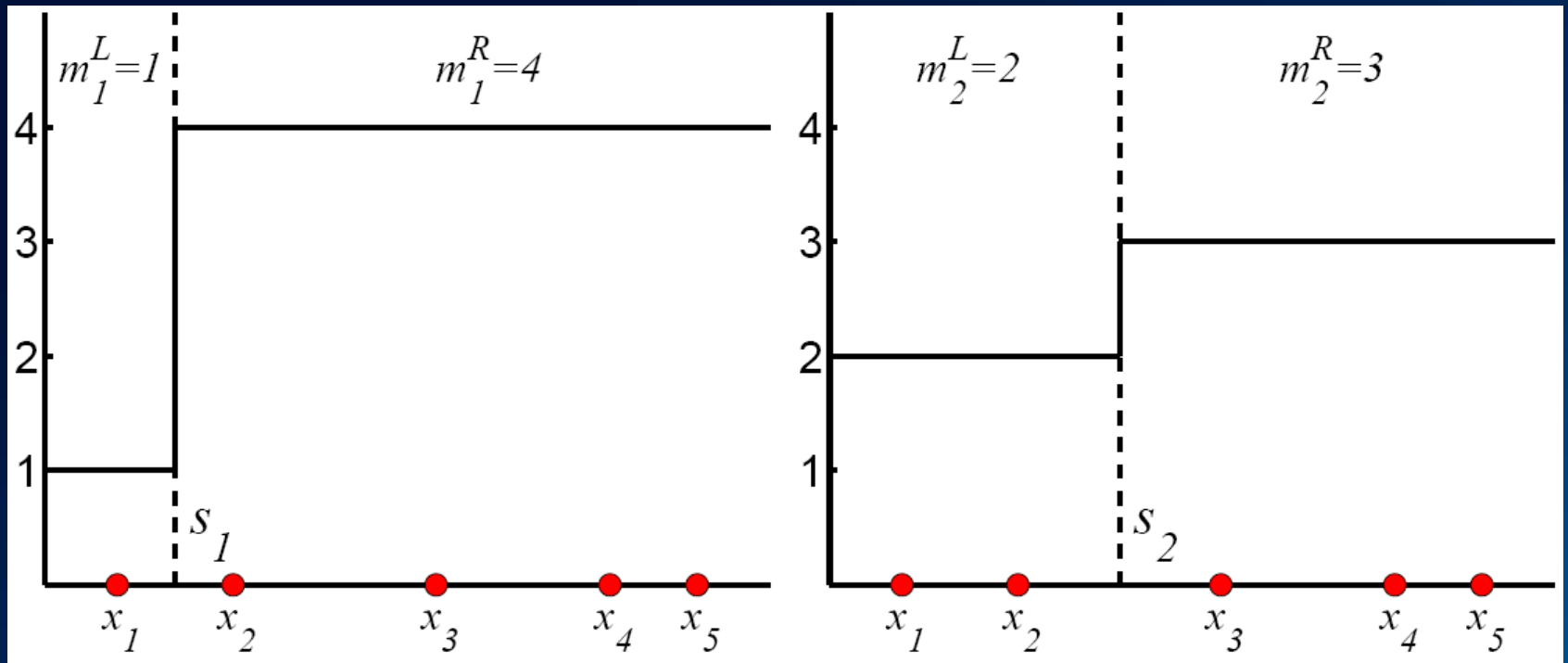
Data Mass or Mass

- Mass is defined as the number of points in a region.
- Two groups of data can have the same mass regardless of the characteristics of the regions (e.g., density, shape or volume.)
- Mass in a given region is defined by a rectangular function which has the same value for the entire region in which the mass is measured.

Definition 1. *Mass base function: $m_i(x)$ as a result of s_i, is defined as*

$$m_i(x) = \begin{cases} m_i^L & \text{if } x \text{ is on the left of } s_i \\ m_i^R & \text{if } x \text{ is on the right of } s_i \end{cases}$$

Note that $m_i^L = n - m_i^R = i$.



(a) $m_i(x)$ due to s_1 (b) $m_i(x)$ due to s_2

$$= \sum_{i=a} ip(s_i) + \sum_{j=1} (n-j)p(s_j) \qquad (1)$$

Theorems

Theorem 1. *$mass(x_a)$ is the maximum at $a = n/2$ for any density distribution of $\{x_1, \cdots, x_n\}$; and the points x_a, where $x_1 < x_2 < \cdots < x_{n-1} < x_n$ on the real line, can be ordered based on mass as follows.*

$$mass(x_a) < mass(x_{a+1}),\ a < n/2$$
$$mass(x_a) > mass(x_{a+1}),\ a > n/2$$

Theorem 2. *$mass(x_a)$ is a concave function defined w.r.t. $\{x_1, x_2, \ldots, x_n\}$, when $p(s_i) = (x_{i+1} - x_i)/(x_n - x_1)$.*

Corollaries

Corollary 1. *A mass distribution estimated using binary splits stipulates an ordering, based on mass, of the points in a data cloud from $x_{n/2}$ (with the maximum mass) to the fringe points (with the minimum mass at either side of $x_{n/2}$), irrespective of the density distribution including uniform density distribution.*

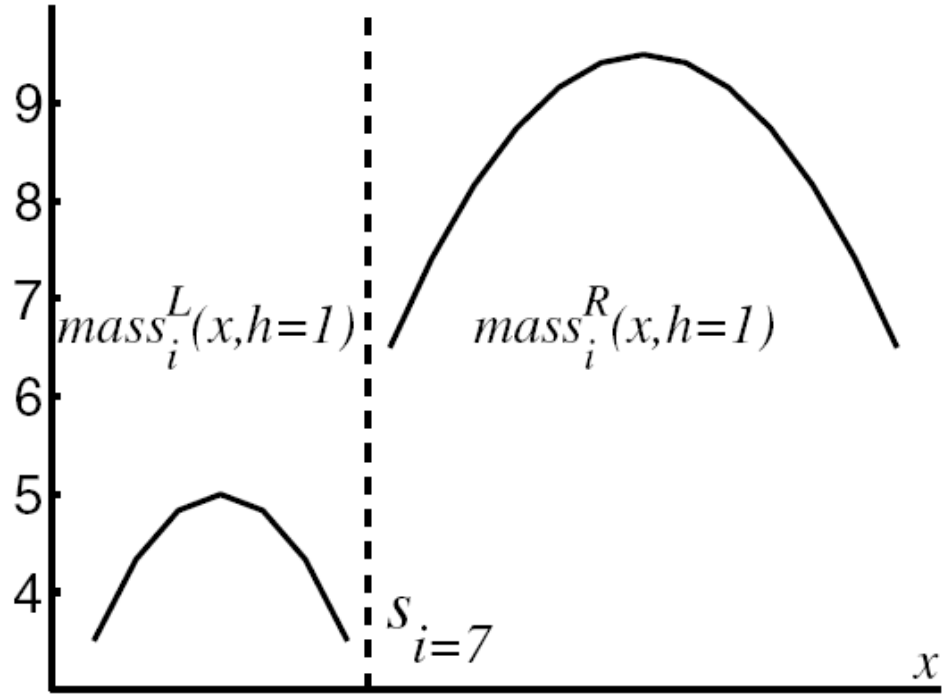
Corollary 2. *The concavity of mass distribution stipulates that fringe points have markedly smaller mass than points close to $x_{n/2}$.*

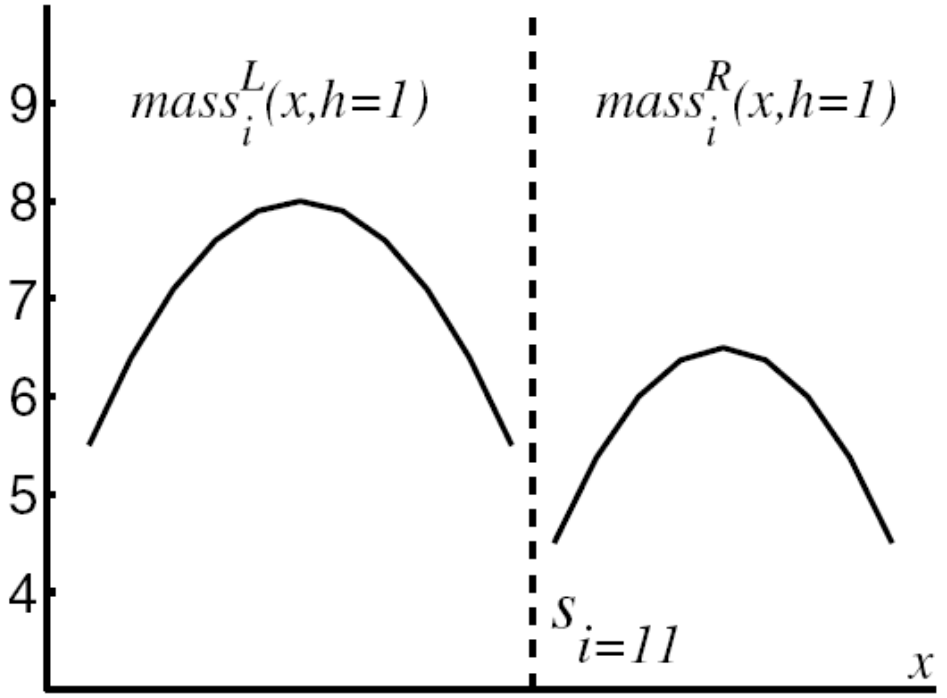
- **Mass is a measure of relevance with respect to the concept underlying the data from which the mass distribution is generated.**
- **Points having high mass are highly relevance to the concept.**

Level- *h* Mass Estimation

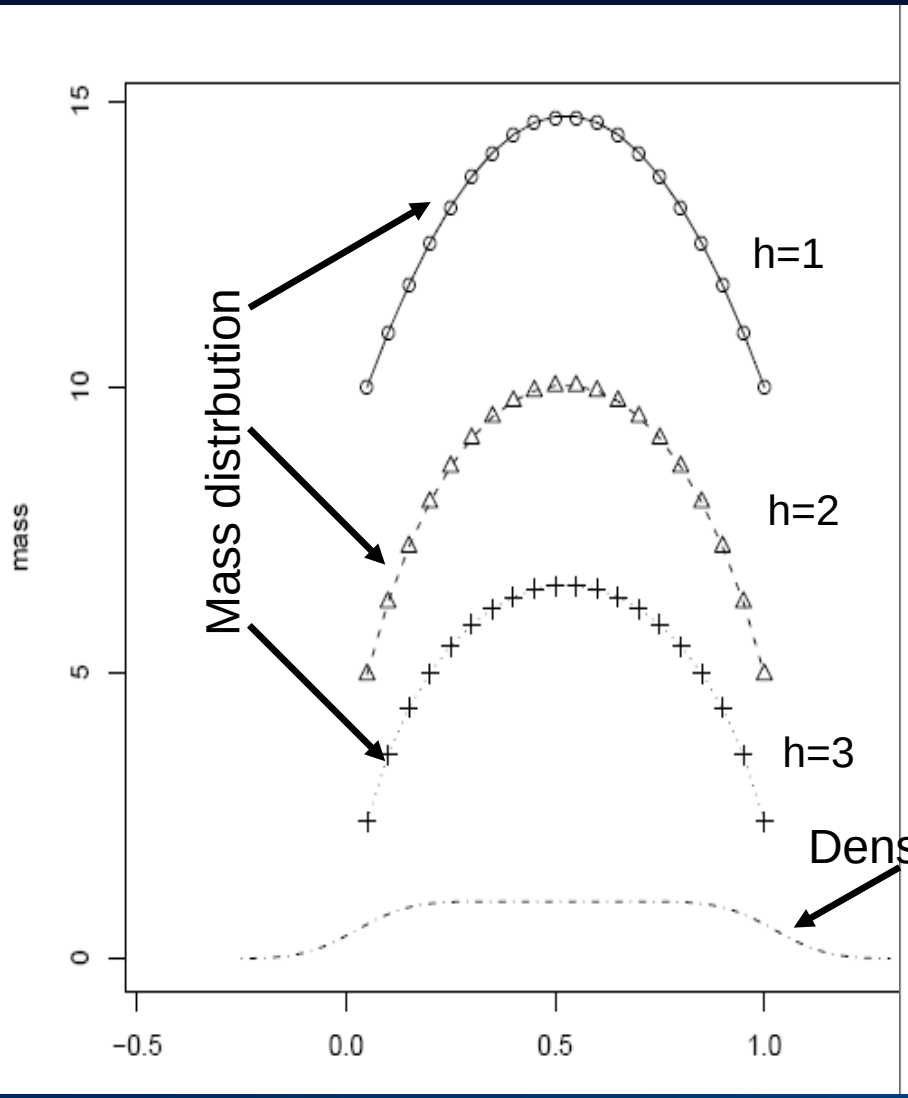
Definition 3. *Level-h mass distribution for a point* $x_a \in \{x_1, \ldots, x_n\}$, *where* $h < n$, *is expressed as*

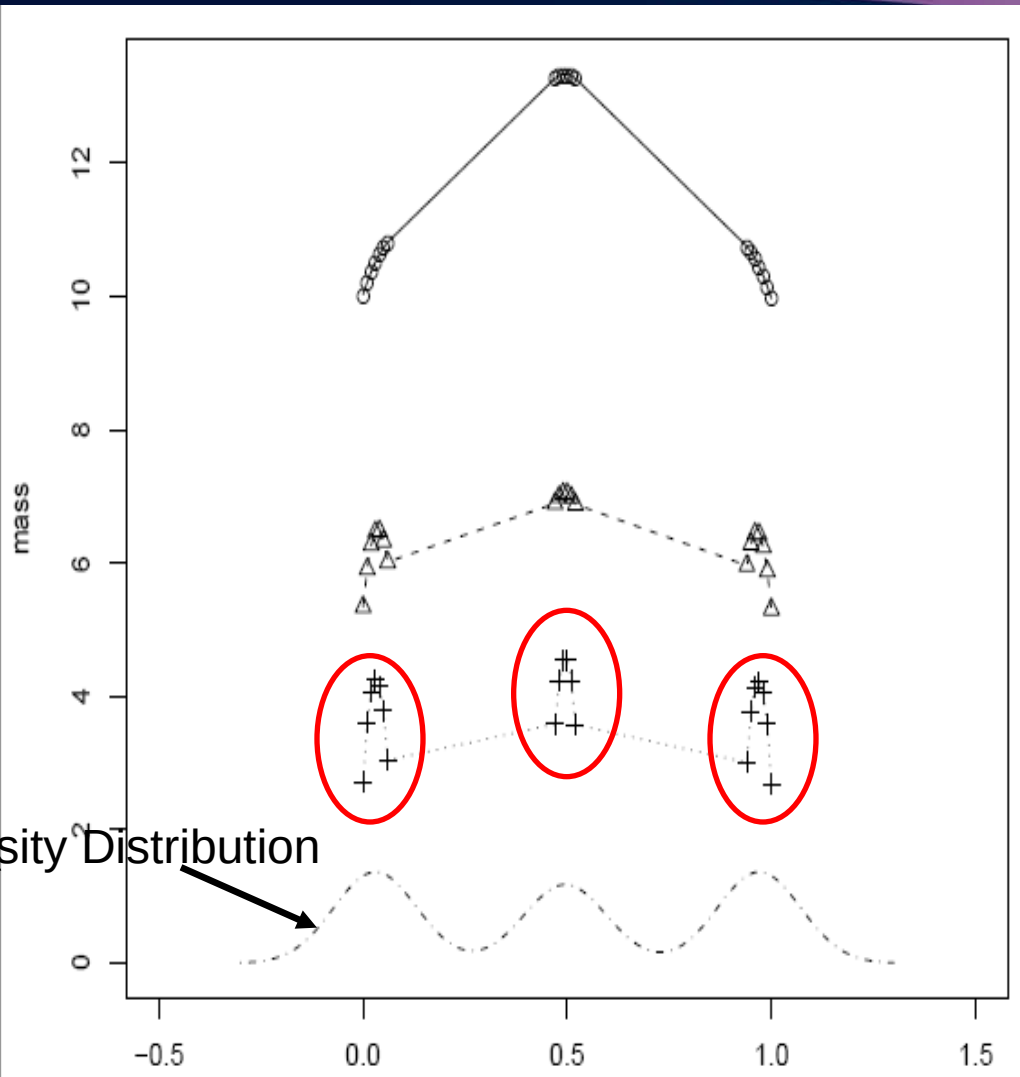
$$mass(x_a, h) = \sum_{i=1}^{n-1} mass_i(x_a, h\text{-}1)p(s_i)$$



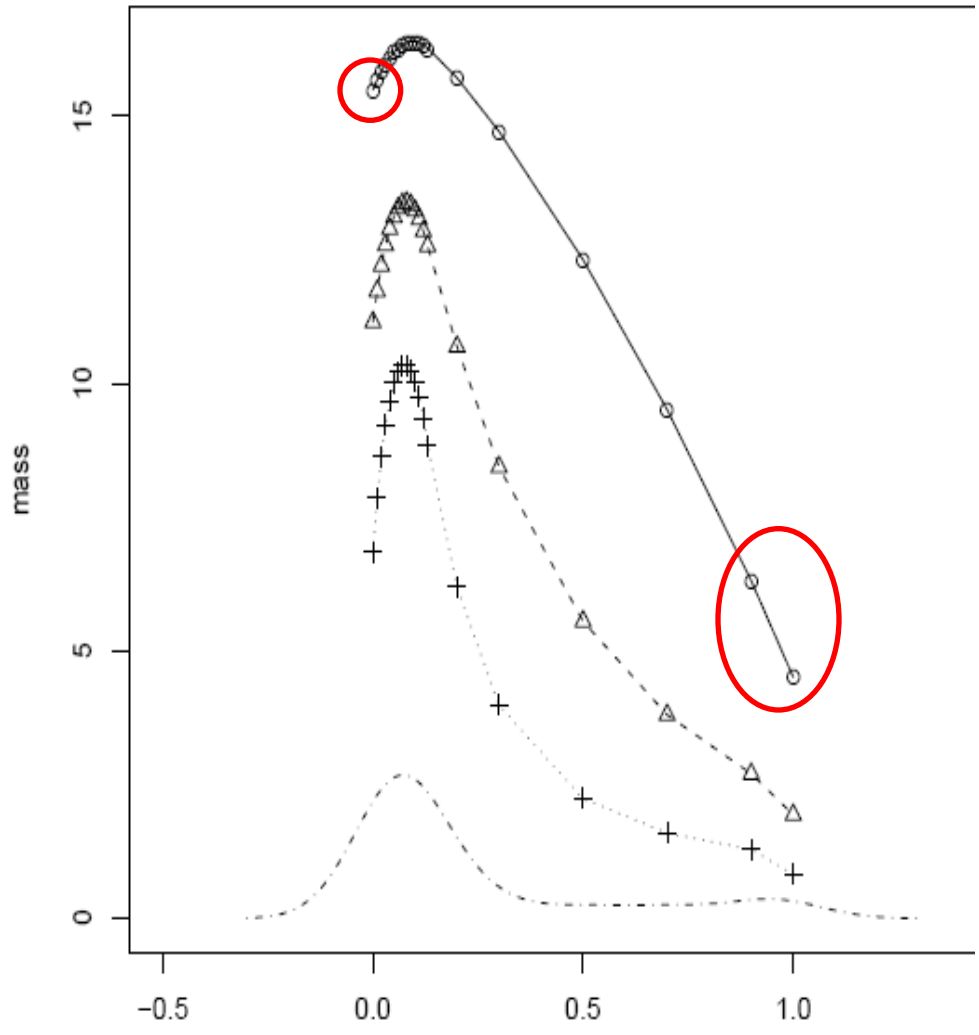


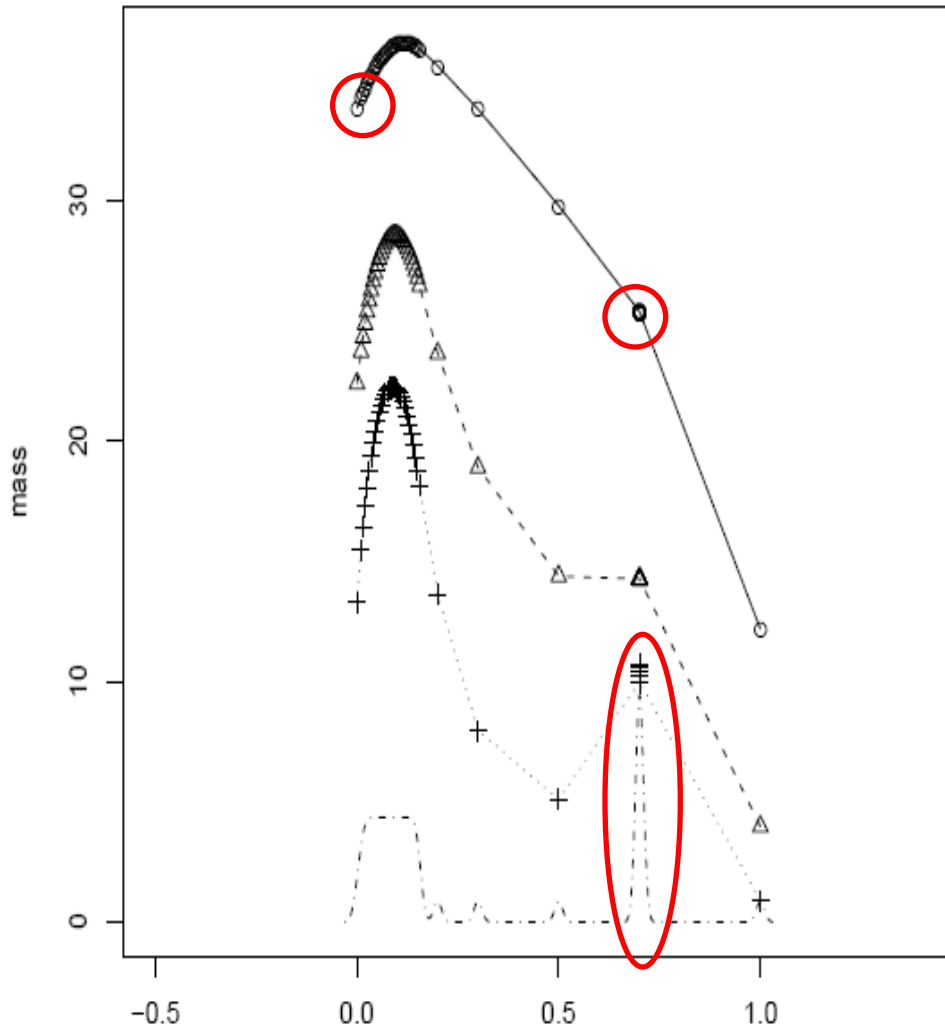
Mass estimation vs density estimation (1)





Mass estimation vs density estimation (2)





Applications of Mass

TASK	
Anomaly Detection	High mass signifies normal points Low mass signifies anomalies
Information Retrieval	High (low) mass signifies that a database object is highly (less) relevant to the query. The same framework can be applied for relevance feedback.
Regression	SVR is able to exploit the 'stretched' mass space to improve its predictive performance.

How Mass can be applied

TASK	How mass is applied
Regression	Indirect Approach
Information Retrieval	Indirect and Direct Approaches (Multimedia Data Mining Workshop)
Anomaly Detection	Direct Approach

Mass Estimation

Algorithm 1 : Mass_Estimation(D, ψ, h, t)

Inputs: D - input data; ψ - data size for $\mathcal{D}_k$; h - level of mass distribution; t - number of mass distributions.

Output: $\widetilde{\mathbf{mass}}(\mathbf{x}) \rightarrow \mathcal{R}^t$ - a function consists of t mass distributions, $mass(x^d, h|\mathcal{D}_k)$.

1: **for** $k = 1$ to t **do**
2: $\quad \mathcal{D}_k \leftarrow$ a random subset of size ψ from D;
3: $\quad d \leftarrow$ a randomly selected dimension from { 1,. . . ,u };
4: $\quad$ Build $mass(x^d, h|\mathcal{D}_k)$;
5: **end for**

where $\mathbf{x}_i = [x_i^1, \ldots, x_i^u]$

Indirect Approach

Algorithm 3 : Perform task in `MassSpace`(D, ψ, h, t)

Inputs: D - input data; ψ - data size for $\mathcal{D}$; h - level of mass distribution; t - number of mass distributions.

Output: Task-specific model.

1: $\widetilde{\mathbf{mass}}(.) \leftarrow$ Mass_Estimation(D, ψ, h, t);

2: $D' \leftarrow$ Mass_Mapping$(D, \widetilde{\mathbf{mass}})$;

3: Perform task (information retrieval or regression) in the mapped mass space using D';

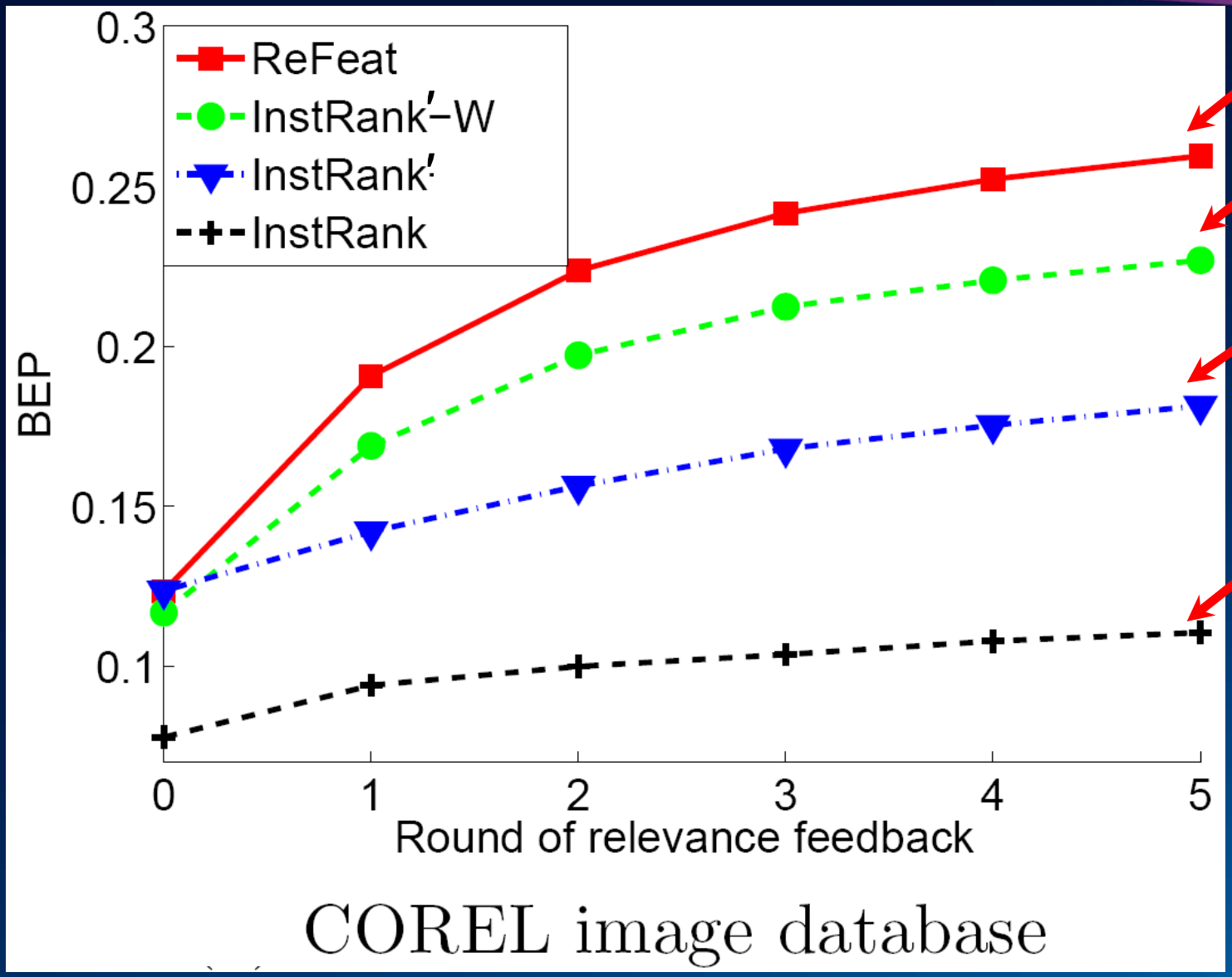
Result of Indirect Approach : Content- based Image Retrievel

BEP ($\times 10^{-2}$)

Mass Space is better for information retrieval than the original space

- MRBIR – Manifold-based method; Qsim and InstR – two recent techniques to improve similarity calculation.
- COREL database of 10000 images, with 100 categories.
- 67 features: 11 shape, 24 texture, 32 colour.
- 5 images from each category as initial queries (5 x 100 runs)
- 5 rounds of relevant feedback; each 2 +ve & 2 –ve feedbacks
- Break-even-point (BEP) of the precision-recall curve.

Result of Direct Approach : Content- based Image Retrievel



Mass Estimation has Constant Time and Space Complexities

Table 9: Runtime (second) for sampling, $mass(x, 1|\mathcal{D})$ and $mass(x, 3|\mathcal{D})$, where $t = 1000$ and $\psi = 8$.

	data size	sampling	$mass(x, 1\|\mathcal{D})$	$mass(x, 3\|\mathcal{D})$
Http	567497	185.21	0.57	17.15
Shuttle	49097	12.47	0.59	17.37
COREL	10000	2.34	0.53	17.28
tic	9822	2.28	0.56	17.23
concrete	1030	0.36	0.48	17.28

Characteristics of Mass Estimation

1) **Require a much smaller sample size**
2) **Utilise no distance or density measures**
   - Eliminates major computational cost in distance and density calculation
3) **Scale up to handle extremely large data size**

Contributions

1. **Introduce a base measure, mass, and delineate its three properties:**
   (i) Mass distribution stipulates an ordering
   (ii) This ordering accentuates fringe points with a concave function.
   (iii) It is a constant-time-and-space-complexities estimation method.
2. **A formalism to apply mass for different tasks.**
   Mass estimation has the potentials in applications as diverse as density estimation has applied now.

Final Remark and Future work

- Express the problem in terms of mass rather than density
- Apply mass directly

1. Multi- dimensional mass estimation
2. Improve/extend the formalism
3. Purposes of mass estimation and density estimation are different—it is thus important identify areas for which each is best suited. This will ascertain areas in which density has been a mismatch, unbeknown thus far

Acknowledgements

This is a joint work with

- **Guang- Tong Zhou**
  (visiting from Shandong University)
- **Tony Liu**
- **James Tan Swee Chuan**
  (now at SIM University)

Funding supports:

- **US Air Force Office of Scientific Research and Asian Office of Aerospace Research & Development.**
- **Shandong University: Guang- Tong Zhou**

Thank you for your attention