Clustering methods: Part 7

Outlier removal

Pasi Fränti

16.5.2017

Machine Learning

University of Eastern Finland

UNIVERSITY OF EASTERN FINLAND

Outlier detection methods

Distance-based methods

- Knorr & Ng

Density-based methods

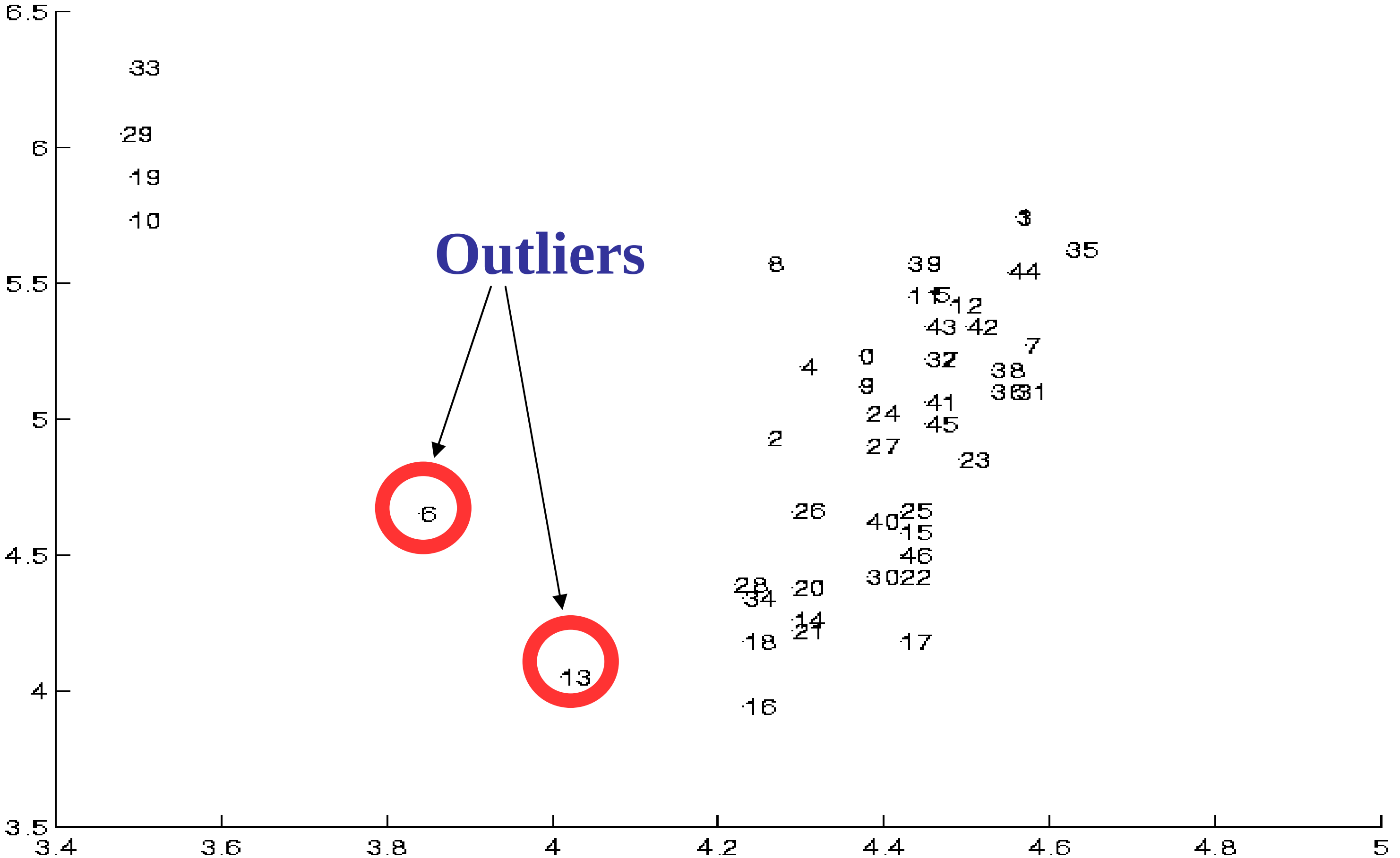
- KDIST: K^{th} nearest distance
- MeanDIST: Mean distance

Graph-based methods

- MkNN: Mutual K-nearest neighbor
- ODIN: Indegree of nodes in k-NN graph

What is outlier?

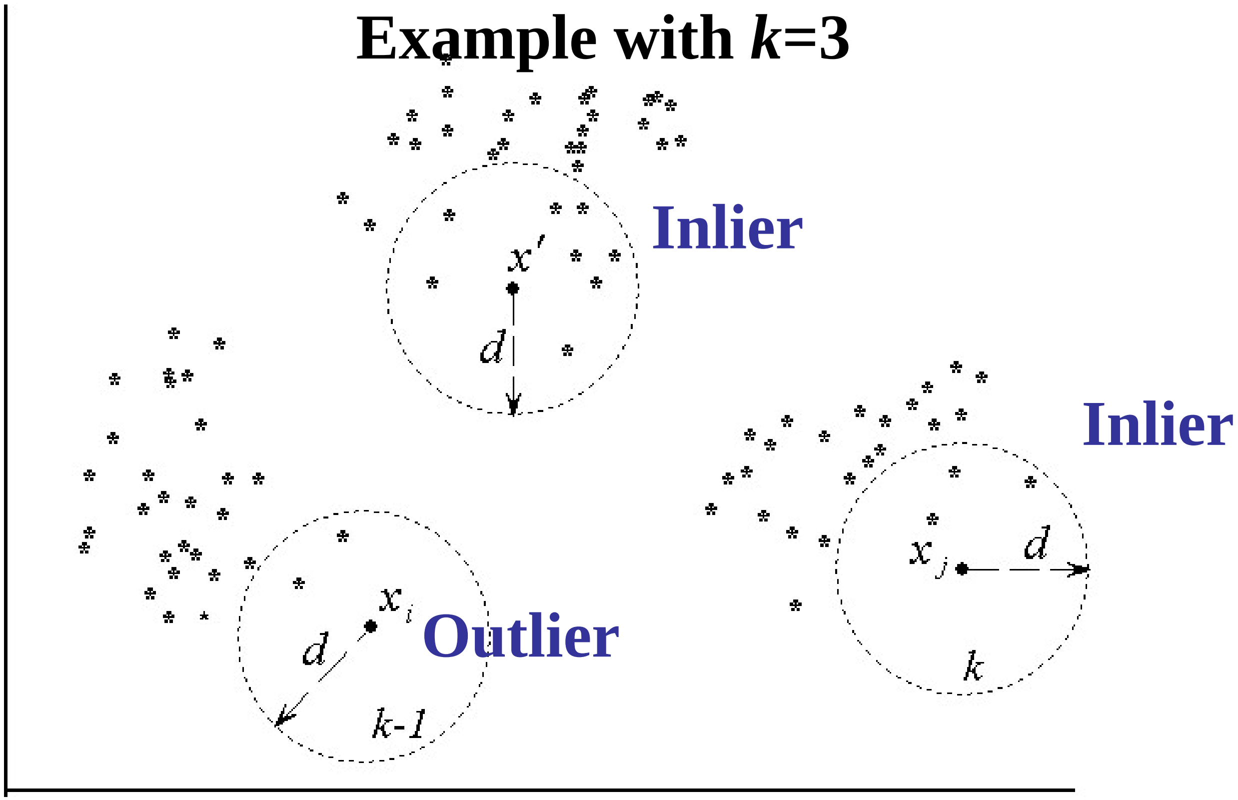
One definition: Outlier is an observation that deviates from other observations so much that it is expected to be generated by a different mechanism.



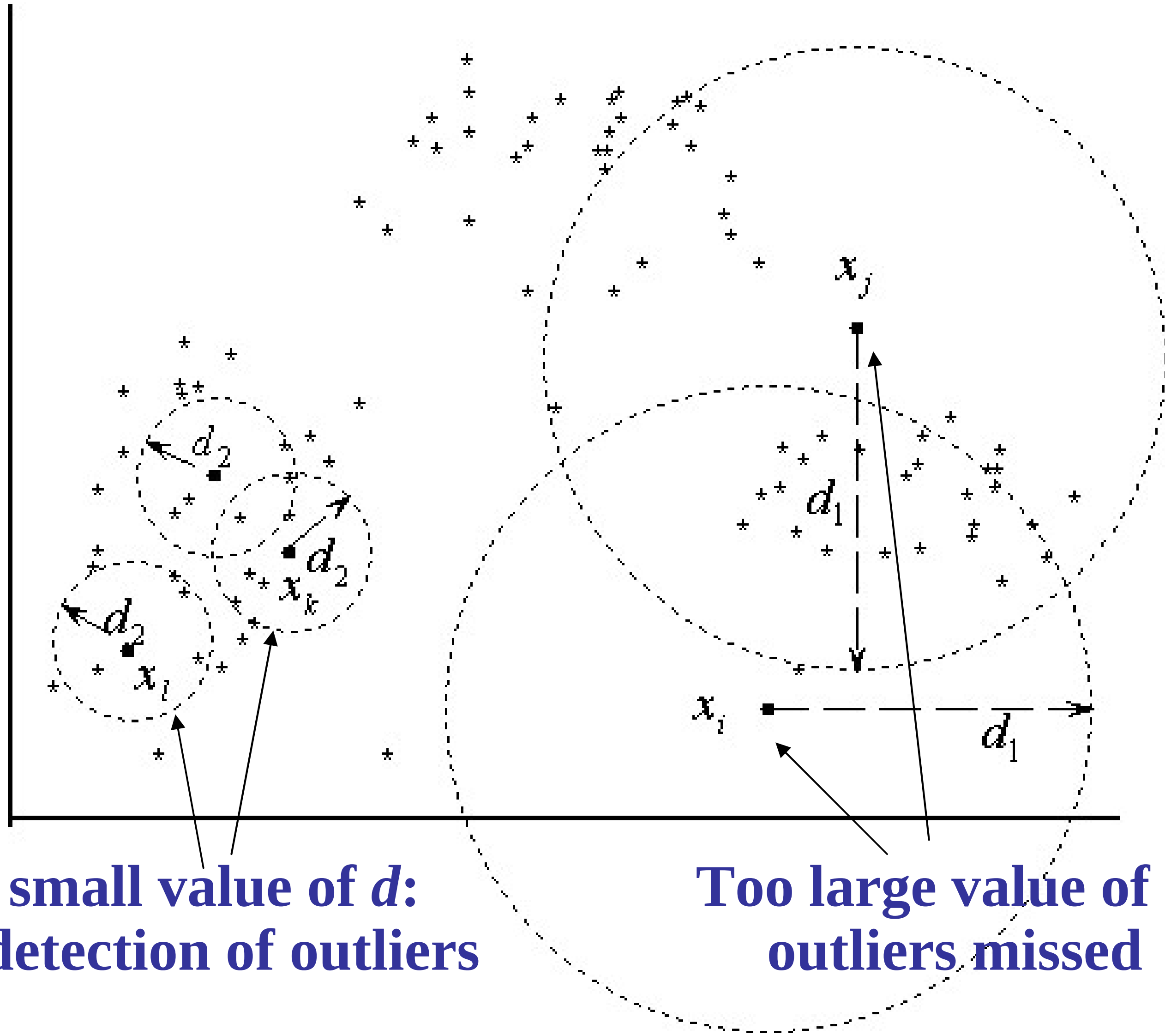
Distance-based method

[Knorr and Ng , CASCR 1997]

Definition: Data point x is an outlier if at most k points are within the distance d from x.



Selection of distance threshold



Too small value of *d*: false detection of outliers

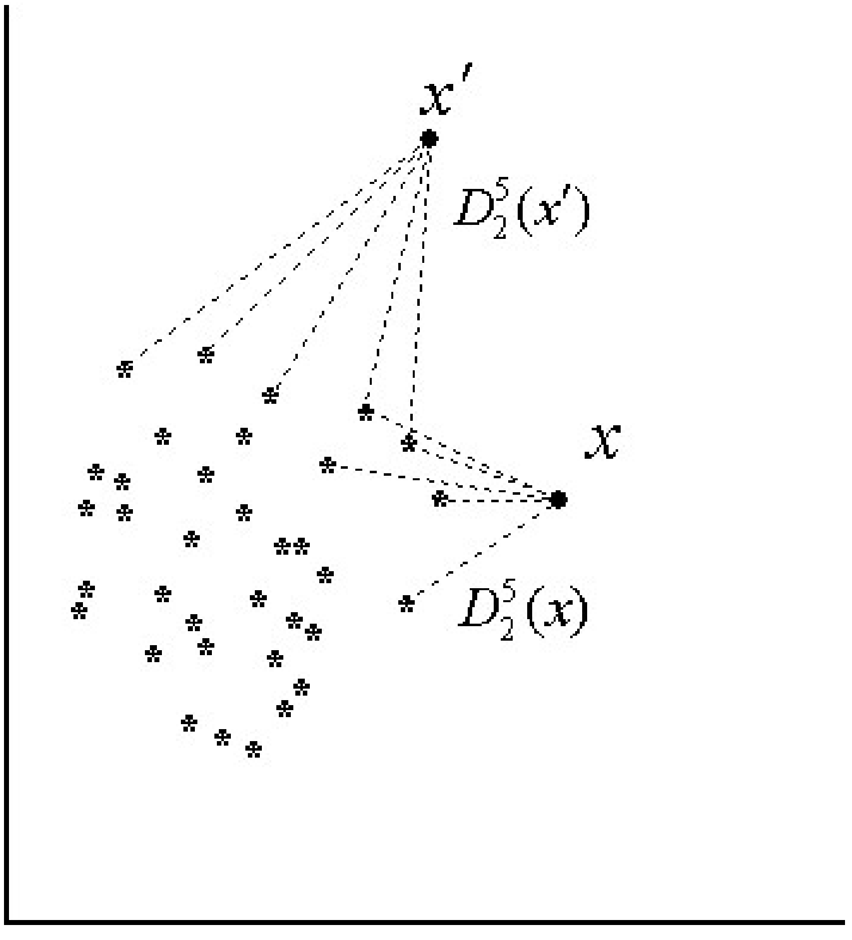
Too large value of *d*: outliers missed

Density-based method: KDIST

[Ramaswamy *et al.*, ACIM SIGMOD 2000]

• Define *KDIST* as distance to the k^{th} nearest point.

• Points are sorted by their *KDIST* distance. The last *n* points in the list are classified as outliers.



Density-based: MeanDist

[Hautamäki *et al.*, ICPR 2004]

MeanDIST = the mean of *k* nearest distances.

User parameters: Cutting point *k*, and local threshold *t:*

$$T = \max\left(L_i - L_{i-1}\right)\cdot t$$

Algorithm 2 MeanDIST

Compute T using Eq. 1 with t

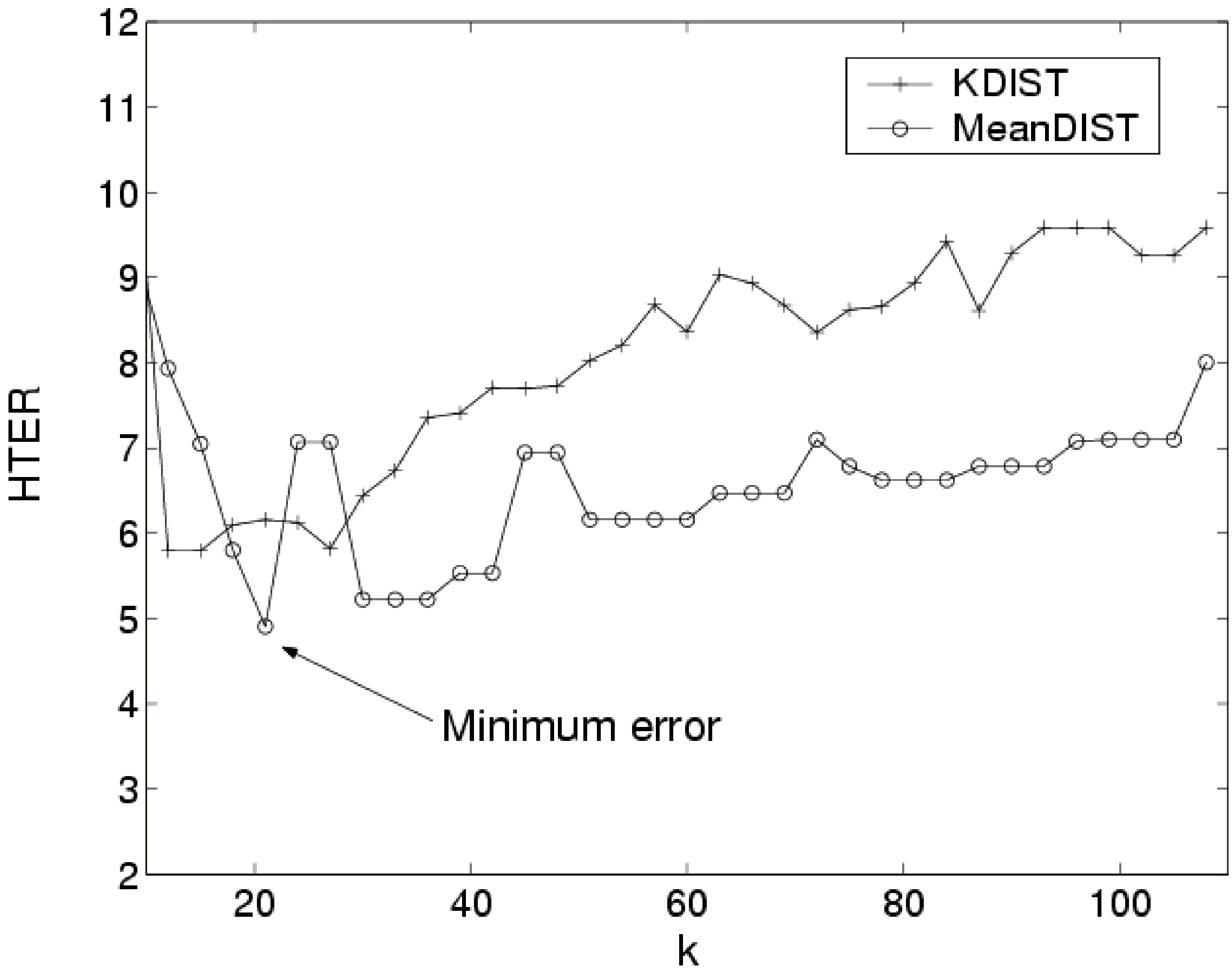
Calculate kNN graph of S

$L \leftarrow$ Sort vectors in ascending order by kNN density

Find smallest i for which $L_i - L_{i-1} \geq T$

Mark $L_i, \ldots, L_{|S|}$ as outliers

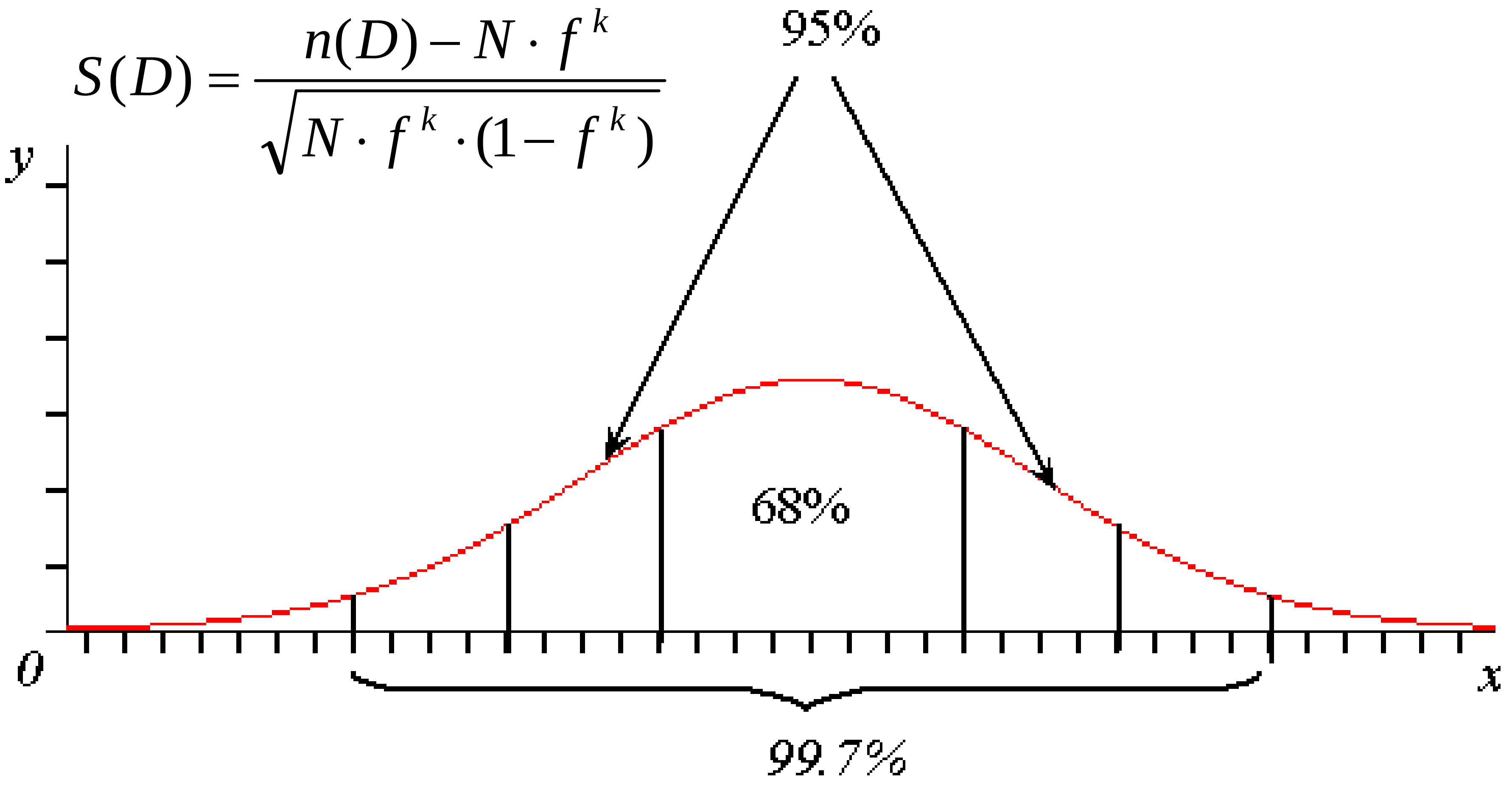
Comparison of KDIST and MeanDIST



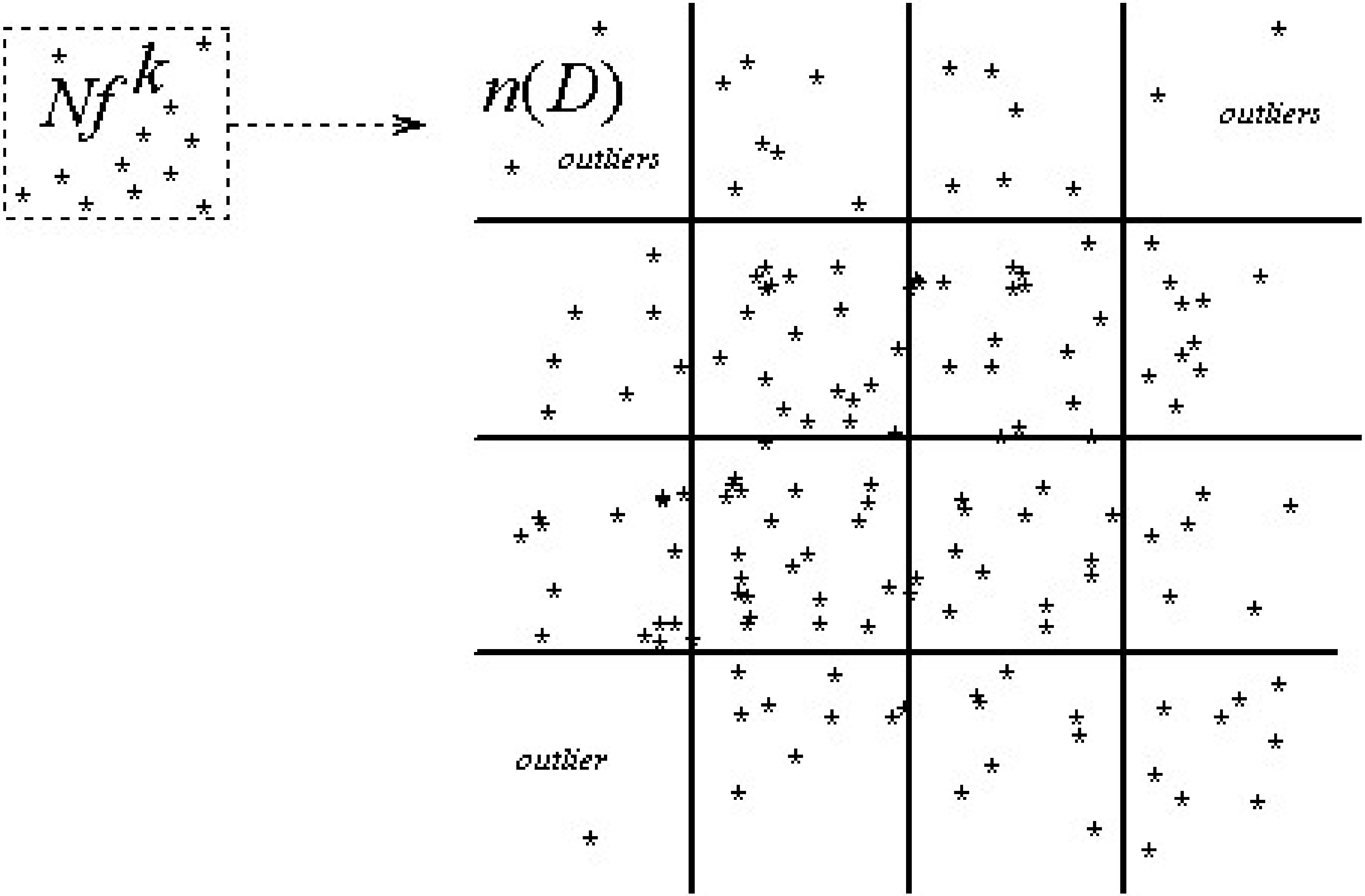
Distribution-based method

[Aggarwal and Yu, ACM SIGMOD, 2001]

$$S(D) = \frac{n(D) - N \cdot f^k}{\sqrt{N \cdot f^k \cdot (1 - f^k)}}$$



Detection of sparse cells



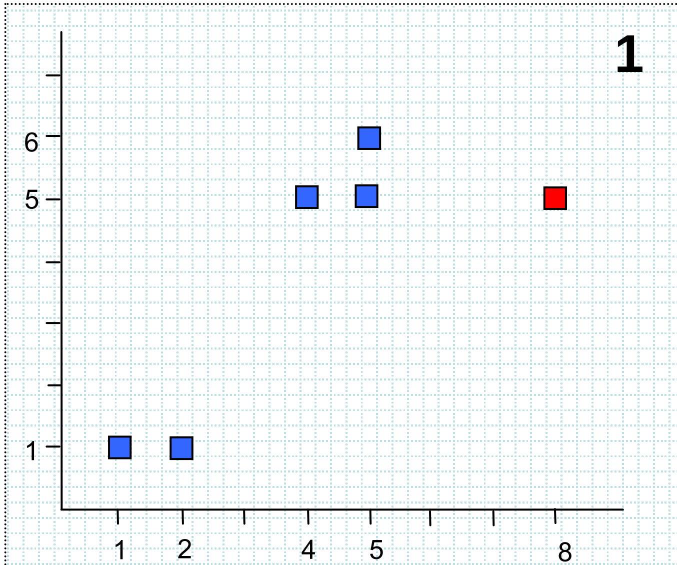
Mutual *k*-nearest neighbor

[Brito et al., *Statistics & Probability Letters*, 1997]

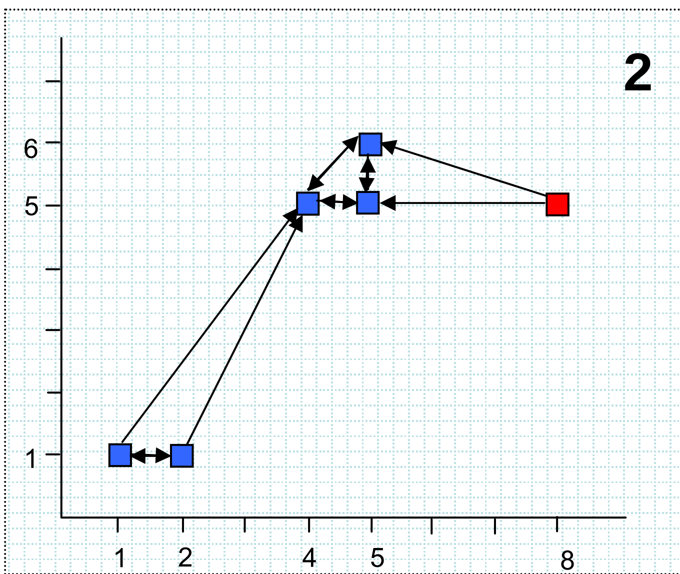
- Generate directed k-NN graph.
- Create undirected graph:
  1. Points *a* and *b* are *mutual neighbors* if both links $a \rightarrow b$ and $b \rightarrow a$ exist.
  2. Change all mutual links $a \leftrightarrow b$ to undirected link a—b.
  3. Remove the rest.
- Connected components are clusters.
- Isolated points as outliers.

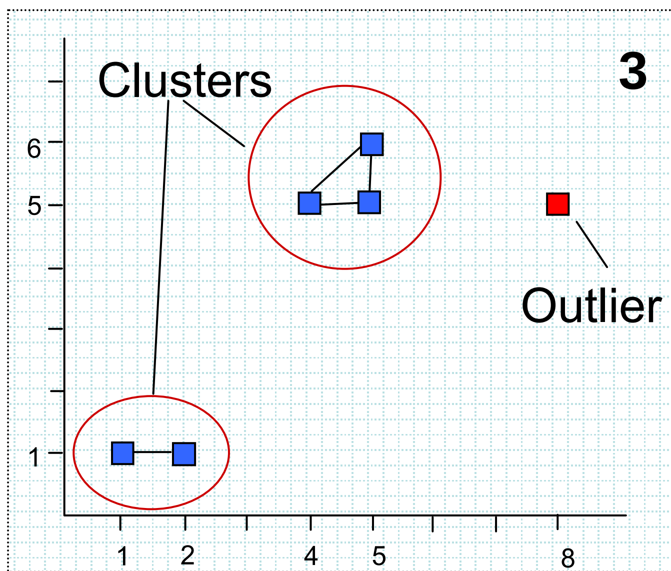
Mutual *k*-NN example

$k = 2$



1. Given a data with one outlier.
2. For each point find two nearest neighbours and create <u>*directed 2-NN*</u> graph.
3. For each pair of points, create link if both a→b and b→a exist.





ODIN: Outlier detection using indegree

[Hautamäki et al., ICPR 2004]

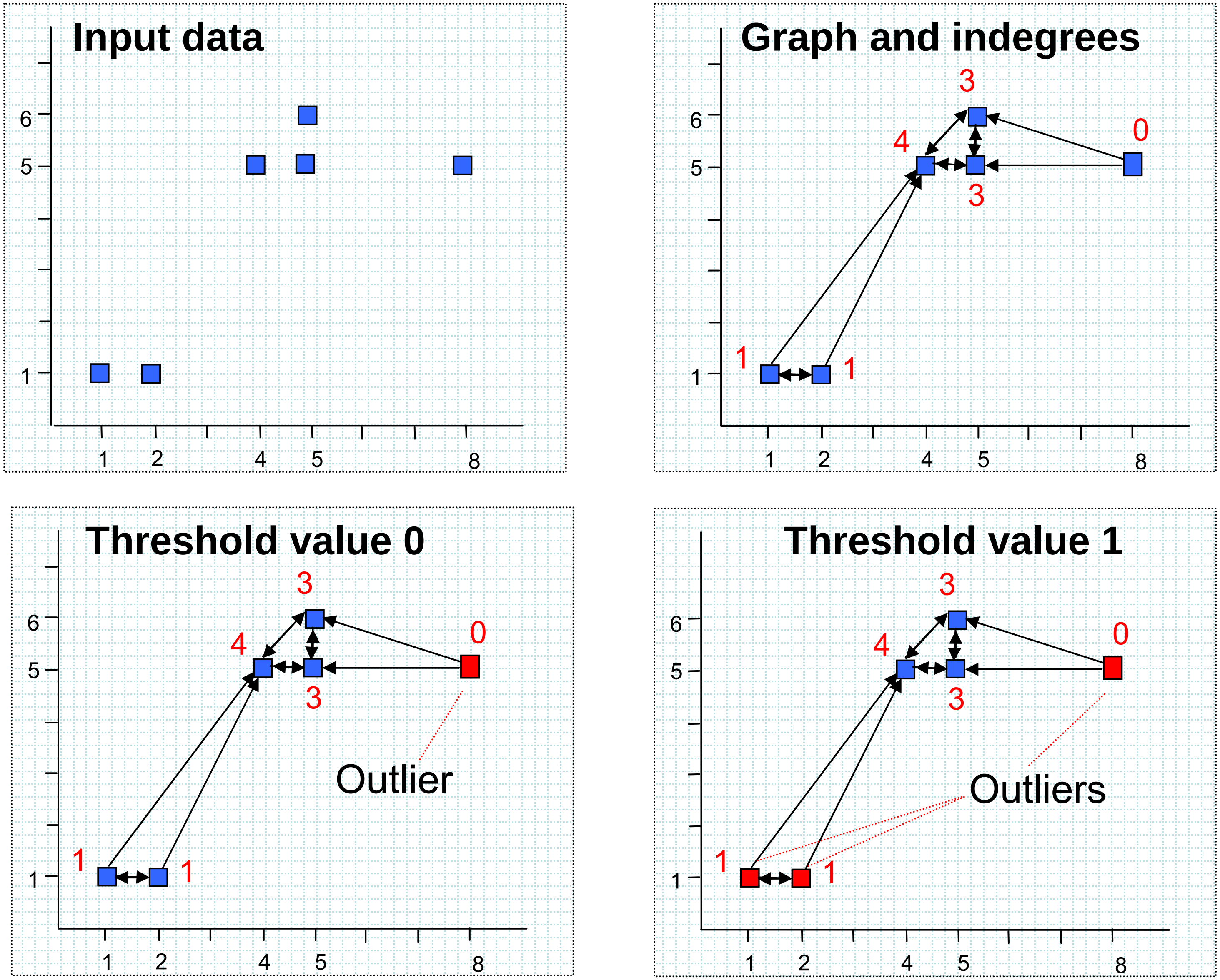
Definition: Given kNN graph, classify data point *x* as an outlier its indegree ≤ *T*.

Algorithm 1 ODIN

T is indegree threshold
Calculate kNN graph of S
for $i = 1$ to $|S|$ **do**
  if indegree of $v_i \leq T$ **then**
    Mark v_i as outlier
  end if
end for

Example of ODIN

$k = 2$

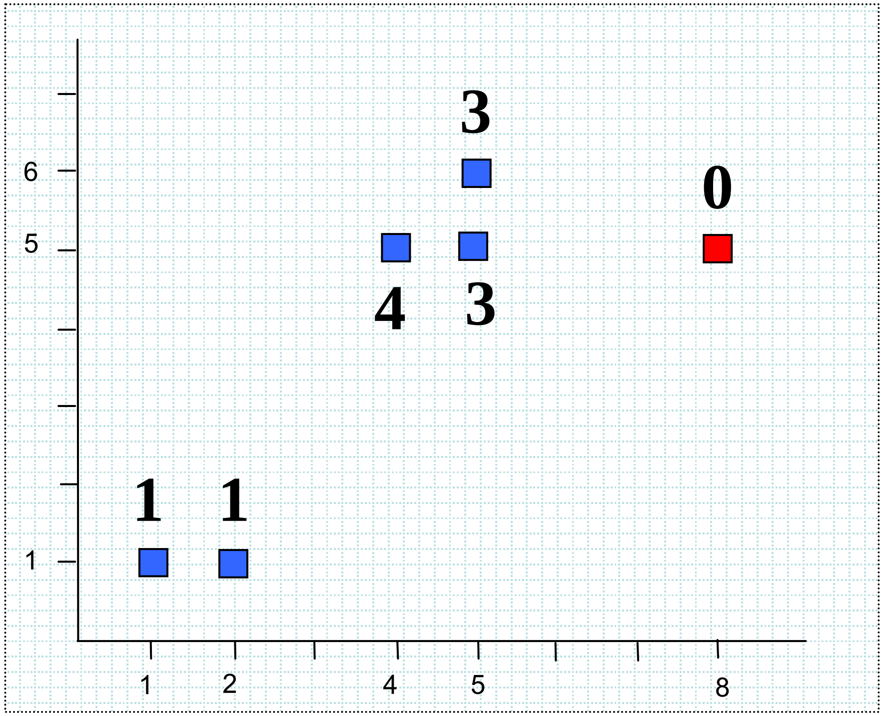


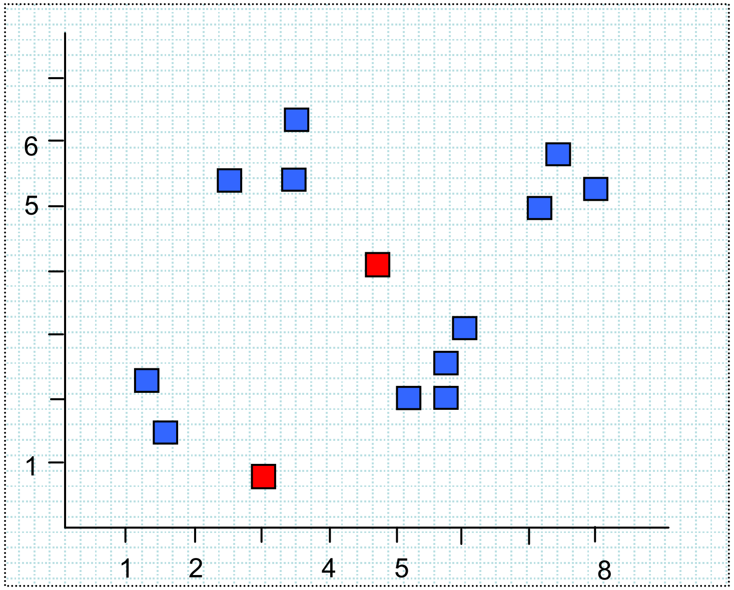
Example of FA and FR

$k = 2$

T	False Acceptance	False Rejection
0	0/1	0/5
1	0/1	2/5
2	0/1	2/5
3	0/1	4/5
4	0/1	5/5
5	0/1	5/5
6	0/1	5/5

Detected as outlier with different threshold values (*T*)





6
5
1
1
2
4
5
8

Experiments

Measures

- False acceptance (FA):
  - Number of outliers that are not detected.
- False rejection (FR):
  - Number of good points wrongly classified as outlier.
- Half total error rate:
  - HTER = (FR+FA) / 2

Comparison of graph-based methods

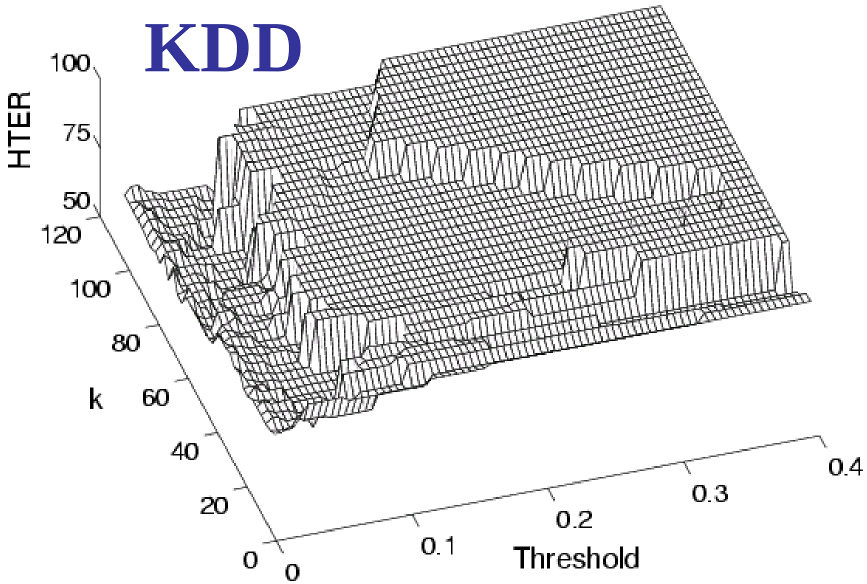
Name	N	d	Outliers
HR [12]	47	2	2
KDD [9]	60318	3	486
NHL1 [8]	681	3	2
NHL2 [8]	731	3	1
synthetic	5165	2	165

Method	synthetic	KDD	HR	NHL1	NHL2
MkNN [1]	50.0 (13)	77.0 (1)	25.0 (5)	25.0 (29)	44.4 (28)
ODIN	9.0 (190,26)	49.6 (1,2)	**0.0** (7, 1)	**0.0** (87, 9)	**0.0** (36, 2)
MeanDIST	**4.9** (21, 0.05)	49.6 (232, 0.19)	30.0 (1, 0.15)	16.7 (20, 0.05)	43.8 (1, 0.57)
KDIST [11]	5.7 (12, 0.06)	**48.6** (72, 0.40)	30.0 (1, 0.15)	30.0 (1, 0.02)	41.7 (7, 0.75)

Difficulty of parameter setup

MeanDIST:

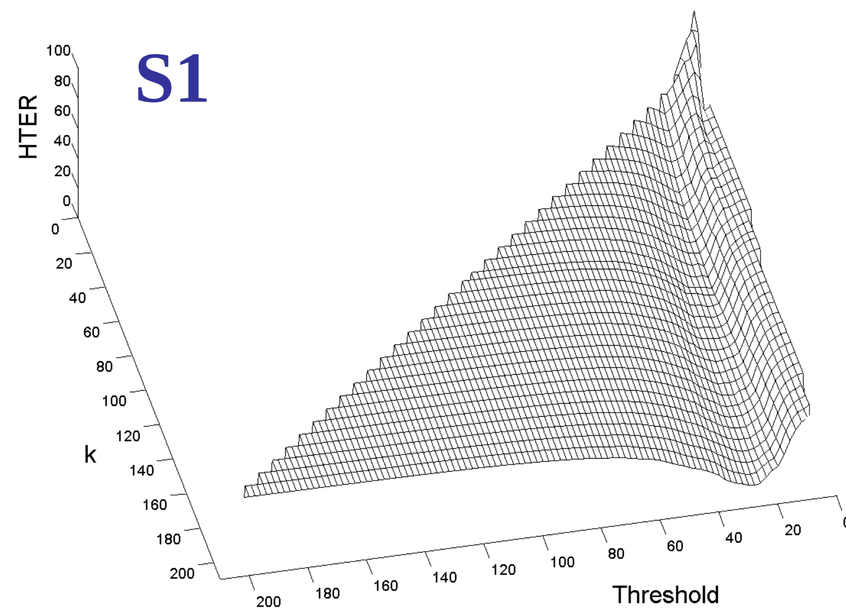
KDD



Value of *k* is not important as long as threshold below 0.1.

ODIN:

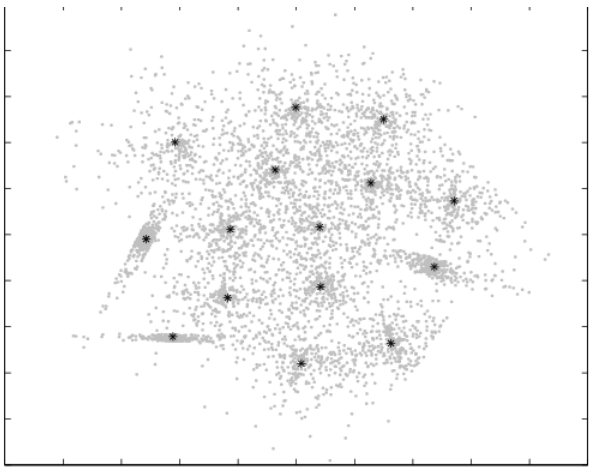
S1



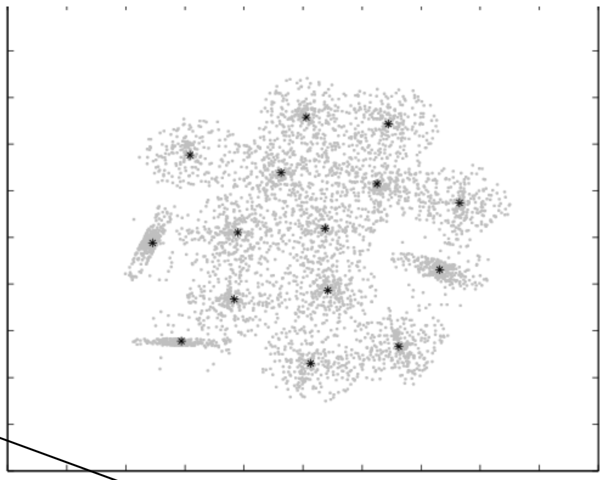
A clear valley in error surface between 20-50.

Improved k-means using outlier removal

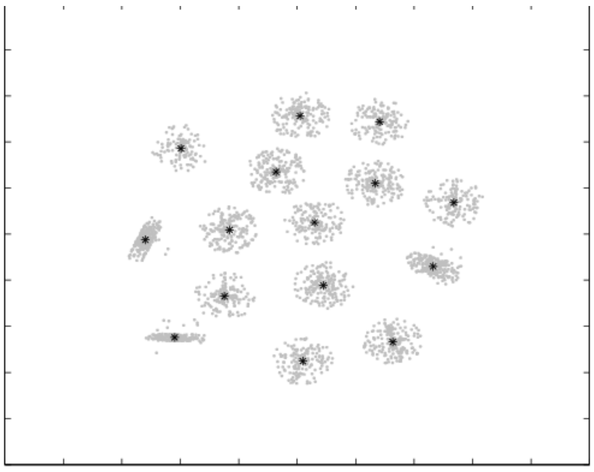
Original



After 40 iterations



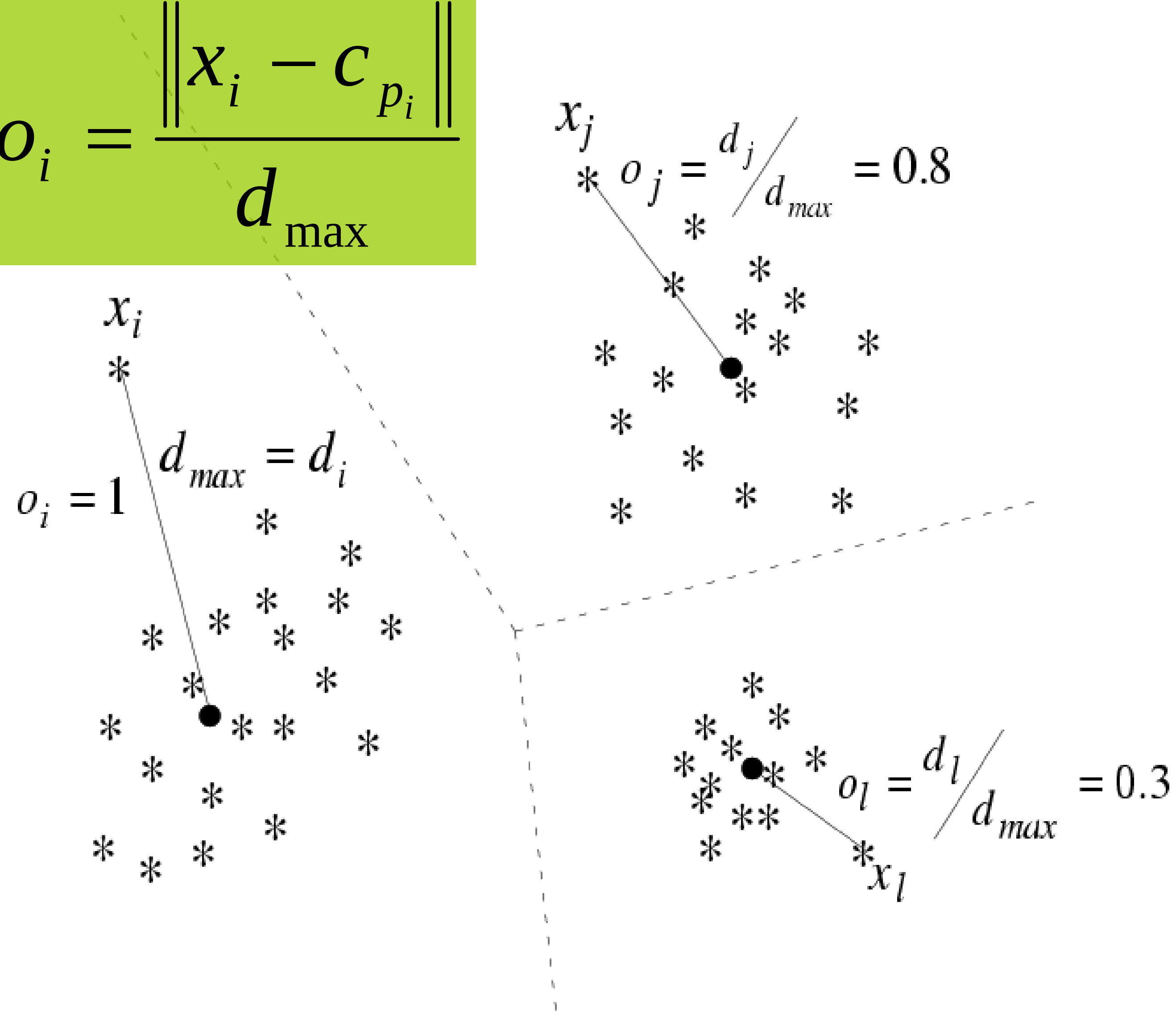
After 70 iterations



At each step, remove most diverging data objects and construct new clustering.

Example of removal factor

$$o_i = \frac{\left\| x_i - c_{p_i} \right\|}{d_{\max}}$$



CERES algorithm

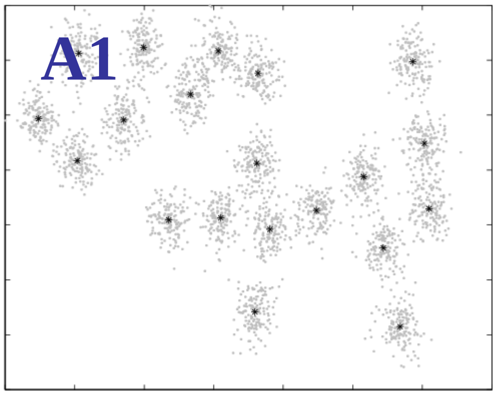
[Hautamäki et al., SCIA 2005]

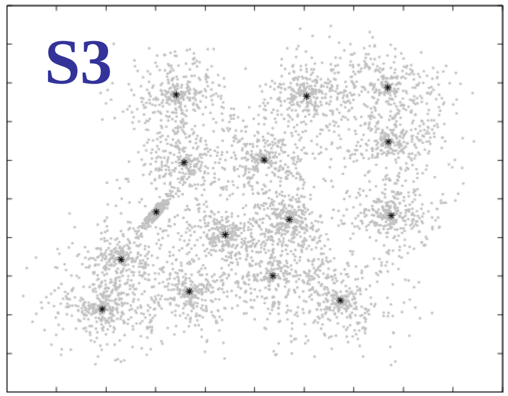
Algorithm 1 CERES(I, T)

$C \leftarrow$ Initialize codebook
for $j \leftarrow 1, \ldots, I$ **do**
  $d_{\max} \leftarrow \max_i\{\|\vec{x}_i - \vec{c}_{p_i}\|\}$
  for $i \leftarrow 1, \ldots, N$ **do**
    $o_i = \|\vec{x}_i - \vec{c}_{p_i}\|/d_{\max}$
    if $o_i > T$ **then**
      $X \leftarrow X \setminus \{\vec{x}_i\}$
    end if
  end for
  $(C, P) \leftarrow$ K-means(X, C)
end for

Experiments

Artificial data sets





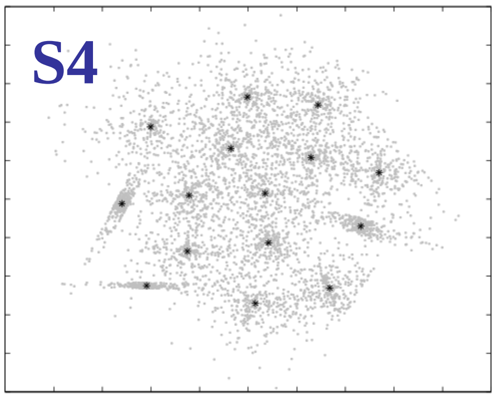
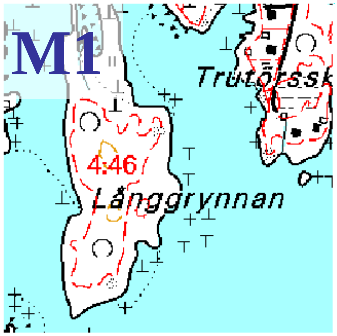
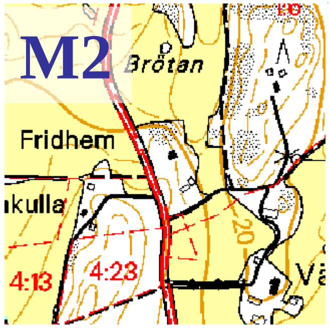
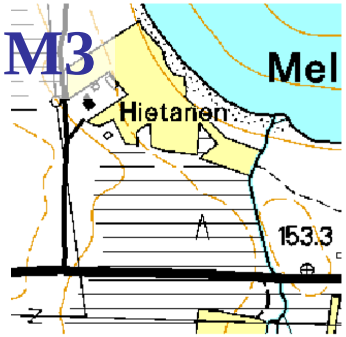


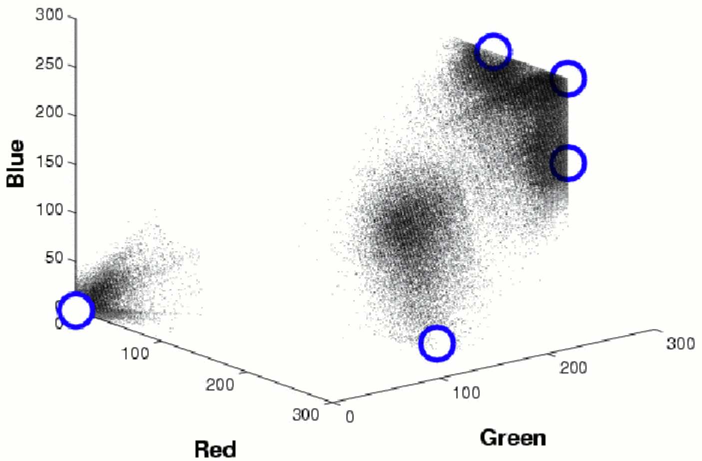
Image data sets







Plot of M2



Comparison

Algorithm	A1	S3	S4	M1	M2	M3
K-means	60	5719	7100	47	32	26
EM	525	3586	3507	46	49	35
CERES	56	3329	2813	45	13	23

Literature

1. D.M. Hawkins, Identification of Outliers, Chapman and Hall, London, 1980.
2. W. Jin, A.K.H. Tung, J. Han, "Finding top-n local outliers in large database", In *Proc. 7th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pp. 293-298, 2001.
3. E.M. Knorr, R.T. Ng, "Algorithms for mining distance-based outliers in large datasets", In *Proc. 24th Int. Conf. Very Large Data Bases*, pp. 392-403, New York, USA, 1998.
4. M.R. Brito, E.L. Chavez, A.J. Quiroz, J.E. Yukich, "Connectivity of the mutual k-nearest-neighbor graph in clustering and outlier detection", *Statistics & Probability Letters*, 35 (1), 33-42, 1997.

Literature

5. C.C. Aggarwal and P.S. Yu, "Outlier detection for high dimensional data", *Proc. Int. Conf. on Management of data ACM SIGMOD*, pp. 37-46, Santa Barbara, California, United States, 2001.
6. V. Hautamäki, S. Cherednichenko, I. Kärkkäinen, T. Kinnunen and P. Fränti, Improving K-Means by Outlier Removal, In *Proc. 14th Scand. Conf. on Image Analysis* (SCIA'2005), 978-987, Joensuu, Finland, June, 2005.
7. V. Hautamäki, I. Kärkkäinen and P. Fränti, "Outlier Detection Using k-Nearest Neighbour Graph", In *Proc. 17th Int. Conf. on Pattern Recognition* (ICPR'2004), 430-433, Cambridge, UK, August, 2004.