Anomaly Detection

Lecture Notes for Chapter 9

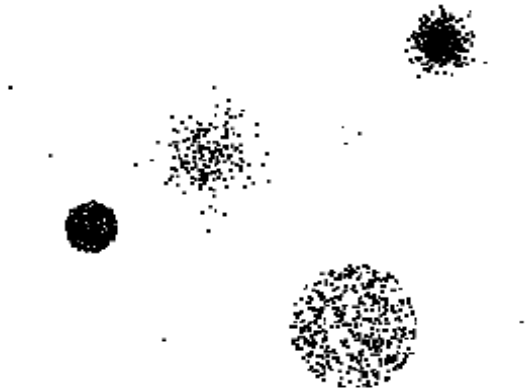
Introduction to Data Mining, 2nd Edition

by

Tan, Steinbach, Karpatne, Kumar

Anomaly/Outlier Detection

- What are anomalies/outliers?
  - The set of data points that are considerably different than the remainder of the data

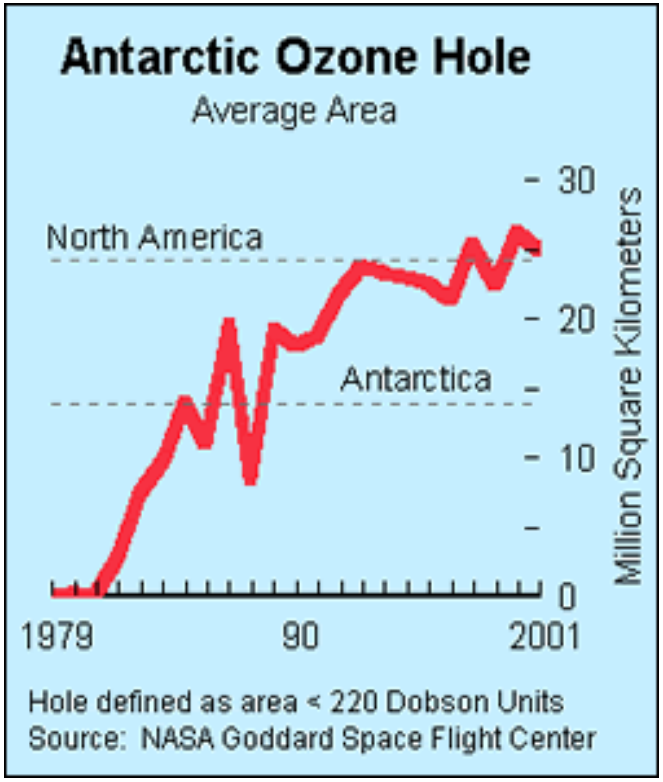


- Natural implication is that anomalies are relatively rare
  - One in a thousand occurs often if you have lots of data
  - Context is important, e.g., freezing temps in July
- Can be important or a nuisance
  - 10 foot tall 2 year old
  - Unusually high blood pressure

Importance of Anomaly Detection

Ozone Depletion History

- In 1985 three researchers (Farman, Gardinar and Shanklin) were puzzled by data gathered by the British Antarctic Survey showing that ozone levels for Antarctica had dropped 10% below normal levels
- Why did the Nimbus 7 satellite, which had instruments aboard for recording ozone levels, not record similarly low ozone concentrations?
- The ozone concentrations recorded by the satellite were so low they were being treated as outliers by a computer program and discarded!



Sources:
http://exploringdata.cqu.edu.au/ozone.html
http://www.epa.gov/ozone/science/hole/size.html

Causes of Anomalies

- Data from different classes
  - Measuring the weights of oranges, but a few grapefruit are mixed in
- Natural variation
  - Unusually tall people
- Data errors
  - 200 pound 2 year old

Distinction Between Noise and Anomalies

- Noise is erroneous, perhaps random, values or contaminating objects
  - Weight recorded incorrectly
  - Grapefruit mixed in with the oranges

- Noise doesn't necessarily produce unusual values or objects
- Noise is not interesting
- Anomalies may be interesting if they are not a result of noise
- Noise and anomalies are related but distinct concepts

General Issues: Number of Attributes

- Many anomalies are defined in terms of a single attribute
  - Height
  - Shape
  - Color
- Can be hard to find an anomaly using all attributes
  - Noisy or irrelevant attributes
  - Object is only anomalous with respect to some attributes
- However, an object may not be anomalous in any one attribute

General Issues: Anomaly Scoring

- Many anomaly detection techniques provide only a binary categorization
  - An object is an anomaly or it isn't
  - This is especially true of classification-based approaches
- Other approaches assign a score to all points
  - This score measures the degree to which an object is an anomaly
  - This allows objects to be ranked
- In the end, you often need a binary decision
  - Should this credit card transaction be flagged?
  - Still useful to have a score
- How many anomalies are there?

Other Issues for Anomaly Detection

- Find all anomalies at once or one at a time
  - Swamping
  - Masking
- Evaluation
  - How do you measure performance?
  - Supervised vs. unsupervised situations
- Efficiency
- Context
  - Professional basketball team

Variants of Anomaly Detection Problems

- Given a data set D, find all data points $\mathbf{x} \in D$ with anomaly scores greater than some threshold t
- Given a data set D, find all data points $\mathbf{x} \in D$ having the top-n largest anomaly scores
- Given a data set D, containing mostly normal (but unlabeled) data points, and a test point **x**, compute the anomaly score of **x** with respect to D

Model-Based Anomaly Detection

- Build a model for the data and see
  - Unsupervised
    - Anomalies are those points that don't fit well
    - Anomalies are those points that distort the model
    - Examples:
      - Statistical distribution
      - Clusters
      - Regression
      - Geometric
      - Graph
  - Supervised
    - Anomalies are regarded as a rare class
    - Need to have training data

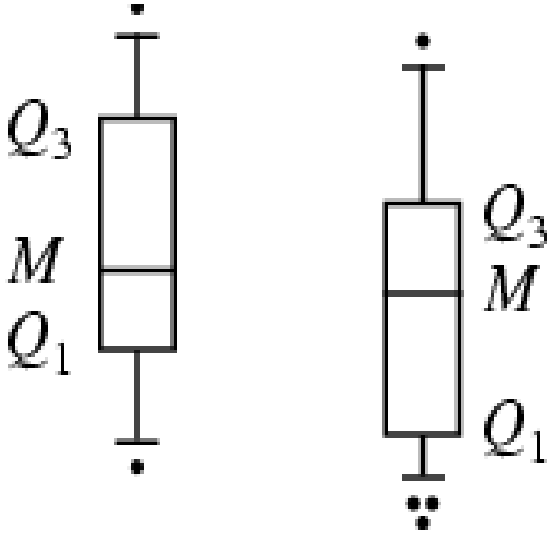
Additional Anomaly Detection Techniques

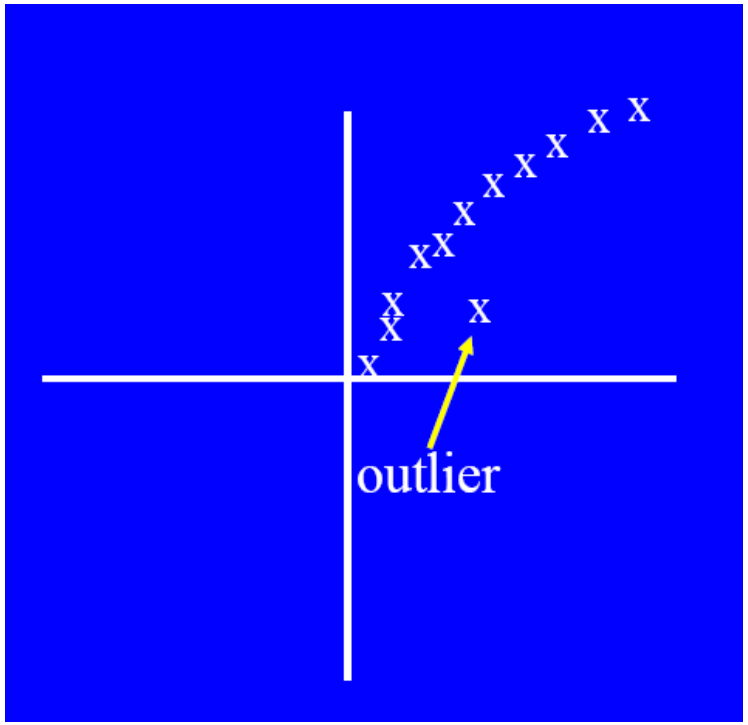
- Proximity-based
  - Anomalies are points far away from other points
  - Can detect this graphically in some cases
- Density-based
  - Low density points are outliers
- Pattern matching
  - Create profiles or templates of atypical but important events or objects
  - Algorithms to detect these patterns are usually simple and efficient

Visual Approaches

- Boxplots or scatter plots

- Limitations
  - Not automatic
  - Subjective



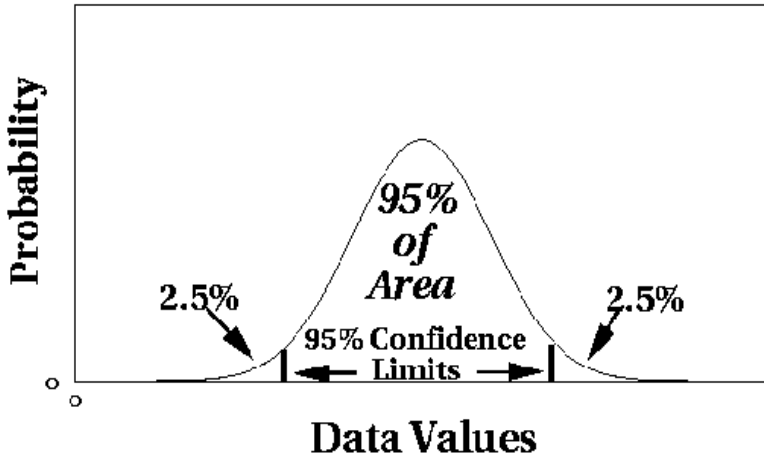


Statistical Approaches

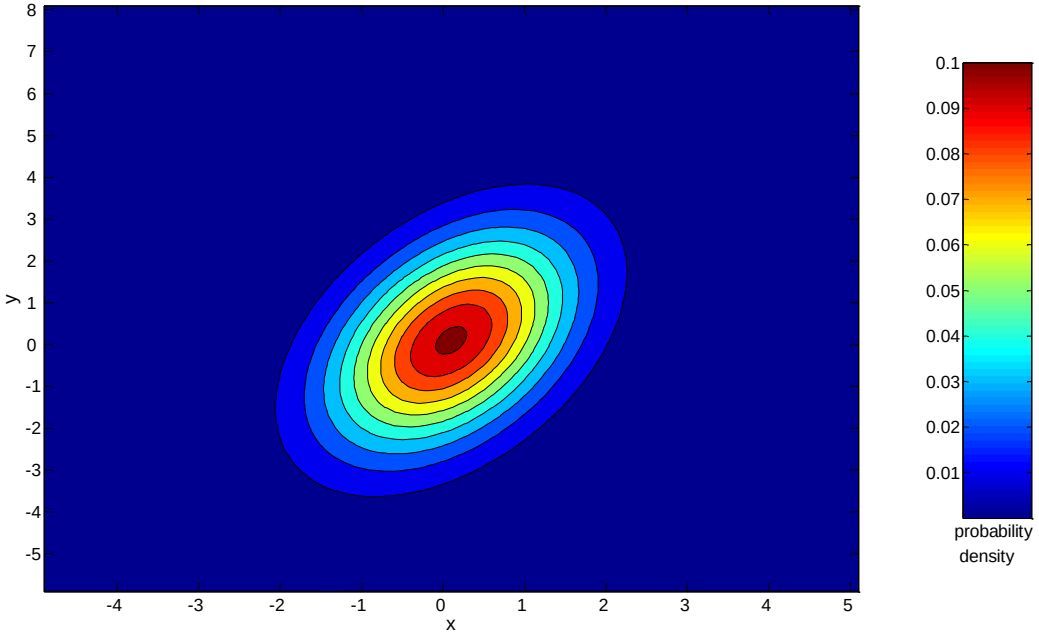
Probabilistic definition of an outlier: An outlier is an object that has a low probability with respect to a probability distribution model of the data.

- Usually assume a parametric model describing the distribution of the data (e.g., normal distribution)
- Apply a statistical test that depends on
  - Data distribution
  - Parameters of distribution (e.g., mean, variance)
  - Number of expected outliers (confidence limit)
- Issues
  - Identifying the distribution of a data set
    - Heavy tailed distribution
  - Number of attributes
  - Is the data a mixture of distributions?

Normal Distributions



One-dimensional Gaussian



Two-dimensional Gaussian

Grubbs' Test

- Detect outliers in univariate data
- Assume data comes from normal distribution
- Detects one outlier at a time, remove the outlier, and repeat
  - H_0: There is no outlier in data
  - H_A: There is at least one outlier
- Grubbs' test statistic:

$$G = \frac{\max\left|X - \overline{X}\right|}{s}$$

- Reject H_0 if:

$$G > \frac{(N-1)}{\sqrt{N}} \sqrt{\frac{t^2_{(\alpha/N,N-2)}}{N-2+t^2_{(\alpha/N,N-2)}}}$$

Find the G Critical Value

n	g_{crit} α=0.05	g_{crit} α=0.01	n	g_{crit} α=0.05	g_{crit} α=0.01	n	g_{crit} α=0.05	g_{crit} α=0.01
3	1.1531	1.1546	15	2.4090	2.7049	80	3.1319	3.5208
4	1.4625	1.4925	16	2.4433	2.7470	90	3.1733	3.5632
5	1.6714	1.7489	17	2.4748	2.7854	100	3.2095	3.6002
6	1.8221	1.9442	18	2.5040	2.8208	120	3.2706	3.6619
7	1.9381	2.0973	19	2.5312	2.8535	140	3.3208	3.7121
8	2.0317	2.2208	20	2.5566	2.8838	160	3.3633	3.7542
9	2.1096	2.3231	25	2.6629	3.0086	180	3.4001	3.7904
10	2.1761	2.4097	30	2.7451	3.1029	200	3.4324	3.8220
11	2.2339	2.4843	40	2.8675	3.2395	300	3.5525	3.9385
12	2.2850	2.5494	50	2.9570	3.3366	400	3.6339	4.0166
13	2.3305	2.6070	60	3.0269	3.4111	500	3.6952	4.0749
14	2.3717	2.6585	70	3.0839	3.4710	600	3.7442	4.1214

ID	Age	Wage		
1	24	7522	0.632865	0.321847
2	22	9294	0.720157	0.261909
3	40	100860	0.065469	2.83533
4	21	11188	0.763803	0.197844
5	90	5586	2.247763	0.387332
6	38	9134	0.021823	0.267321
7	50	2000	0.501928	0.50863
8	19	8128	0.851095	0.301349
9	60	9816	0.938387	0.244252
10	21	6842	0.763803	0.344848
AVERAGE	38.5	17037		
STD	22.91167	29563.76		

ID	Age	Wage		
1	24	7522	0.588868	0.321847
2	22	9294	0.723041	0.261909
3	40	100860	0.484512	2.83533
4	21	11188	0.790127	0.197844
5		5586		0.387332
6	38	9134	0.350339	0.267321
7	50	2000	1.155374	0.50863
8	19	8128	0.924299	0.301349
9	60	9816	1.826236	0.244252
10	21	6842	0.790127	0.344848
AVERAGE	32.77778	17037		
STD	14.90619	29563.76		

ID	Age	Wage		
1	24	7522	0.588868	0.074045
2	22	9294	0.723041	0.577652
3	40		0.484512	
4	21	11188	0.790127	1.274219
5		5586		0.786058
6	38	9134	0.350339	0.518808
7	50	2000	1.155374	2.104901
8	19	8128	0.924299	0.148826
9	60	9816	1.826236	0.769631
10	21	6842	0.790127	0.324133
AVERAGE	32.77778	7723.333		
STD	14.90619	2719.051		

Statistical-based – Likelihood Approach

- Assume the data set D contains samples from a mixture of two probability distributions:
  - M (majority distribution)
  - A (anomalous distribution)
- General Approach:
  - Initially, assume all the data points belong to M
  - Let $L_t(D)$ be the log likelihood of D at time t
  - For each point x_t that belongs to M, move it to A
    - Let $L_{t+1}(D)$ be the new log likelihood.
    - Compute the difference, $\Delta = L_t(D) - L_{t+1}(D)$
    - If $\Delta > c$ (some threshold), then x_t is declared as an anomaly and moved permanently from M to A

Statistical-based – Likelihood Approach

- Data distribution, D = (1 – λ) M + λ A
- M is a probability distribution estimated from data
  - Can be based on any modeling method (naïve Bayes, maximum entropy, etc)
- A is initially assumed to be uniform distribution
- Likelihood at time t:

$$L_t(D) = \prod_{i=1}^{N} P_D(x_i) = \left((1-\lambda)^{|M_t|} \prod_{x_i \in M_t} P_{M_t}(x_i) \right) \left(\lambda^{|A_t|} \prod_{x_i \in A_t} P_{A_t}(x_i) \right)$$

$$LL_t(D) = |M_t| \log(1-\lambda) + \sum_{x_i \in M_t} \log P_{M_t}(x_i) + |A_t| \log \lambda + \sum_{x_i \in A_t} \log P_{A_t}(x_i)$$

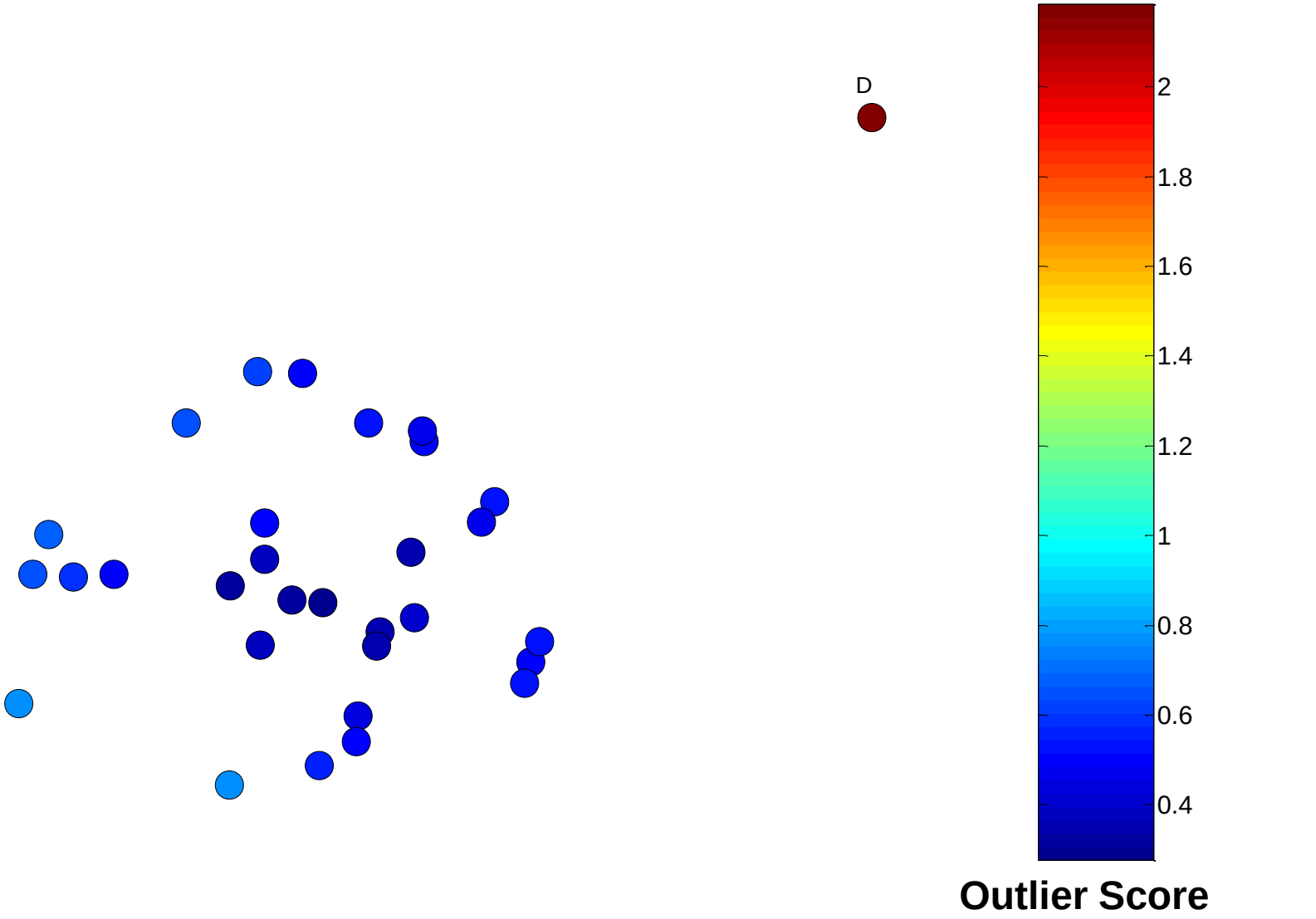
Strengths/Weaknesses of Statistical Approaches

- Firm mathematical foundation
- Can be very efficient
- Good results if distribution is known
- In many cases, data distribution may not be known
- For high dimensional data, it may be difficult to estimate the true distribution
- Anomalies can distort the parameters of the distribution

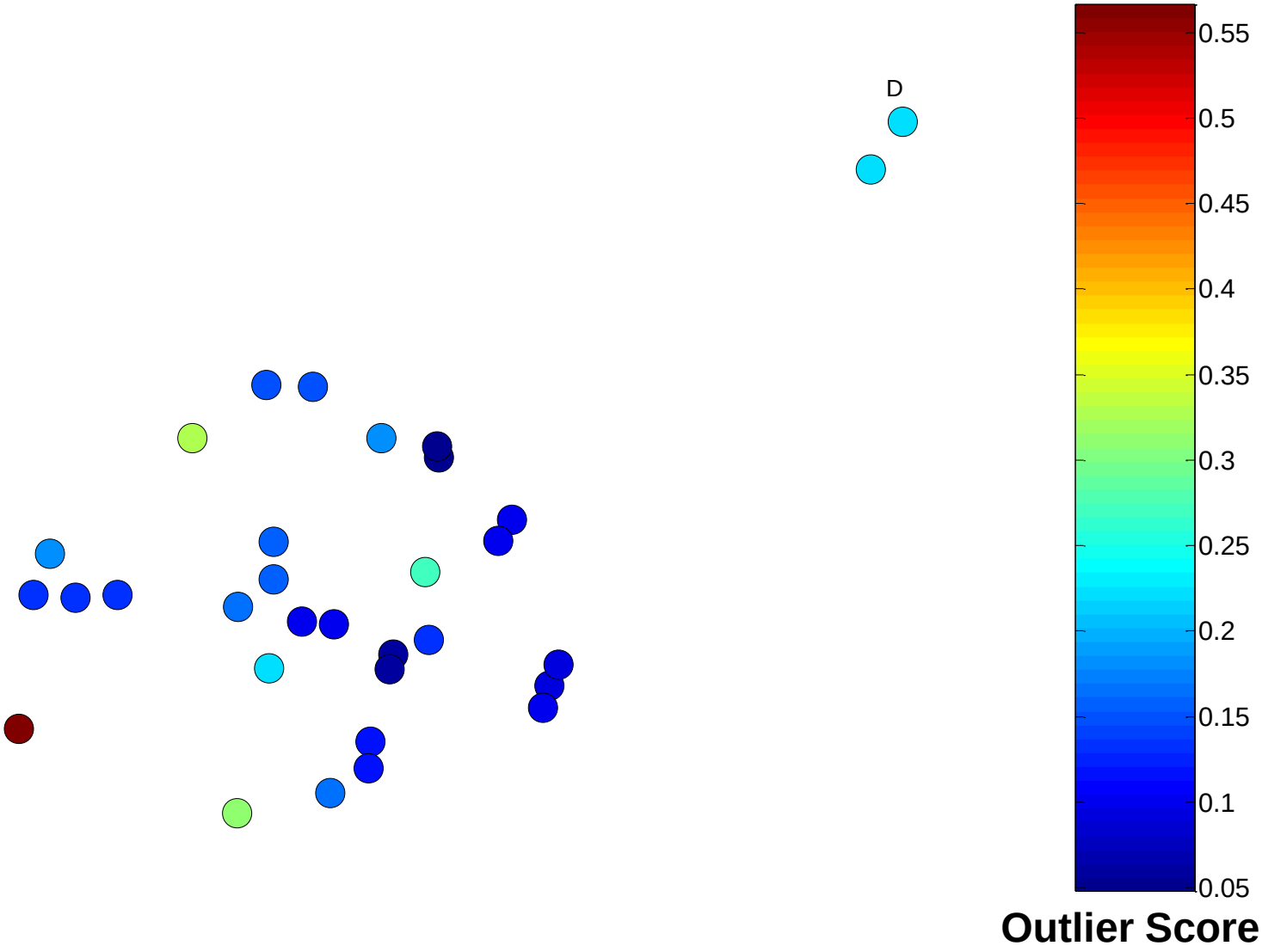
Distance-Based Approaches

- Several different techniques
- An object is an outlier if a specified fraction of the objects is more than a specified distance away (Knorr, Ng 1998)
  - Some statistical definitions are special cases of this
- The outlier score of an object is the distance to its kth nearest neighbor

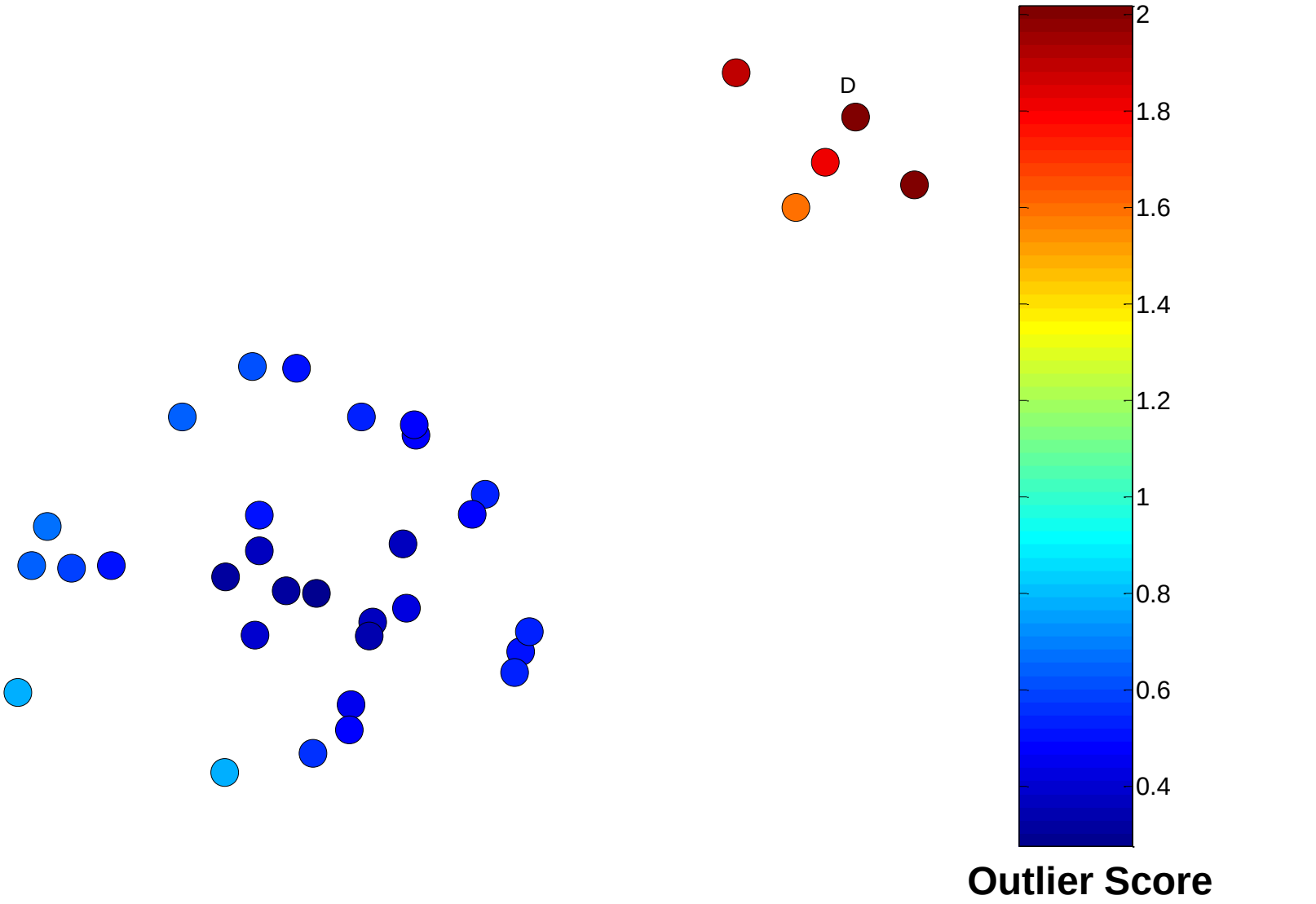
One Nearest Neighbor - One Outlier



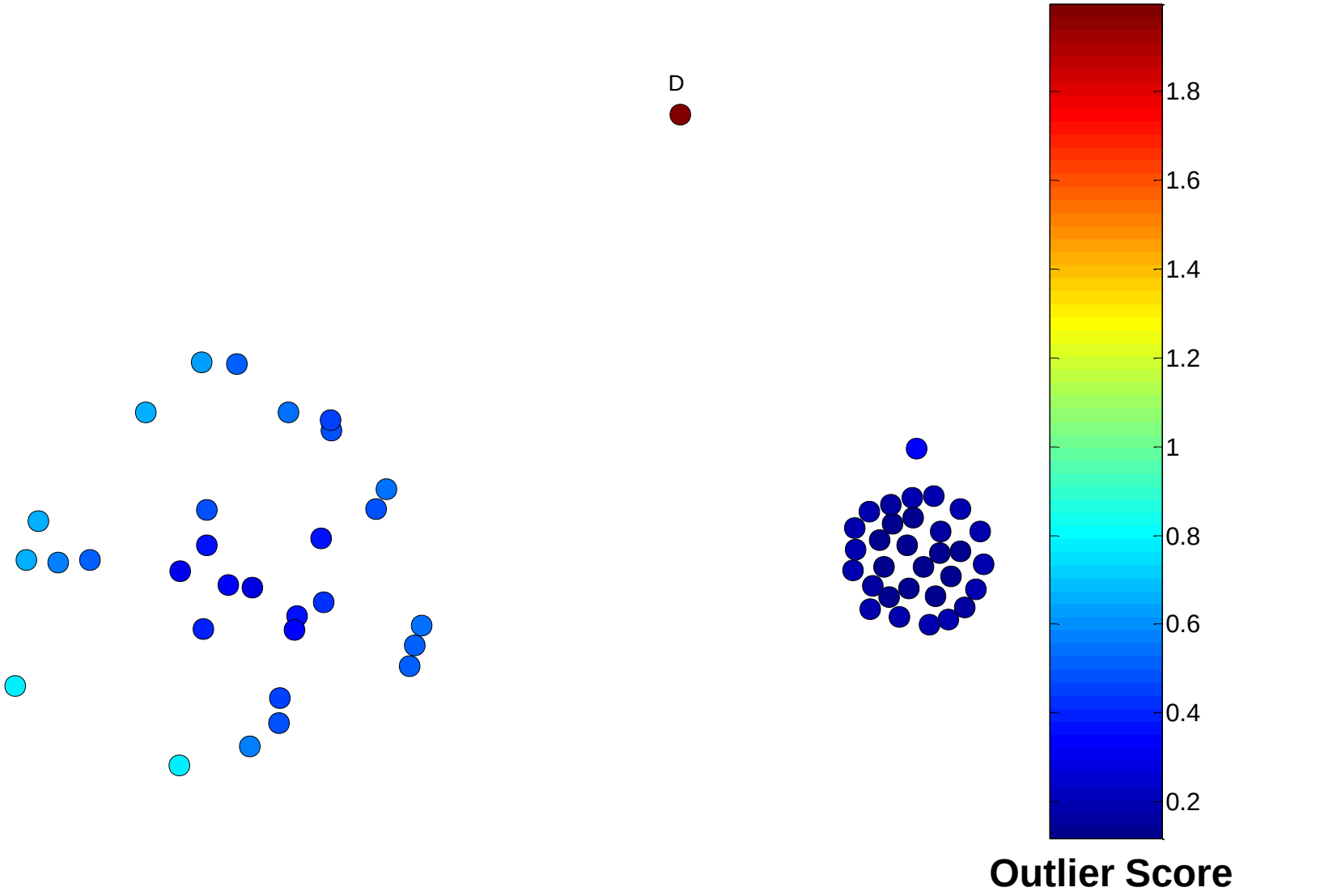
One Nearest Neighbor - Two Outliers



Five Nearest Neighbors - Small Cluster



Five Nearest Neighbors - Differing Density



Strengths/Weaknesses of Distance-Based Approaches

- Simple
- Expensive – $O(n^2)$
- Sensitive to parameters
- Sensitive to variations in density
- Distance becomes less meaningful in high-dimensional space

Density-Based Approaches

- **Density-based Outlier:** The outlier score of an object is the inverse of the density around the object.
  - Can be defined in terms of the k nearest neighbors
  - One definition: Inverse of distance to kth neighbor
  - Another definition: Inverse of the average distance to k neighbors
  - DBSCAN definition

- If there are regions of different density, this approach can have problems

Relative Density

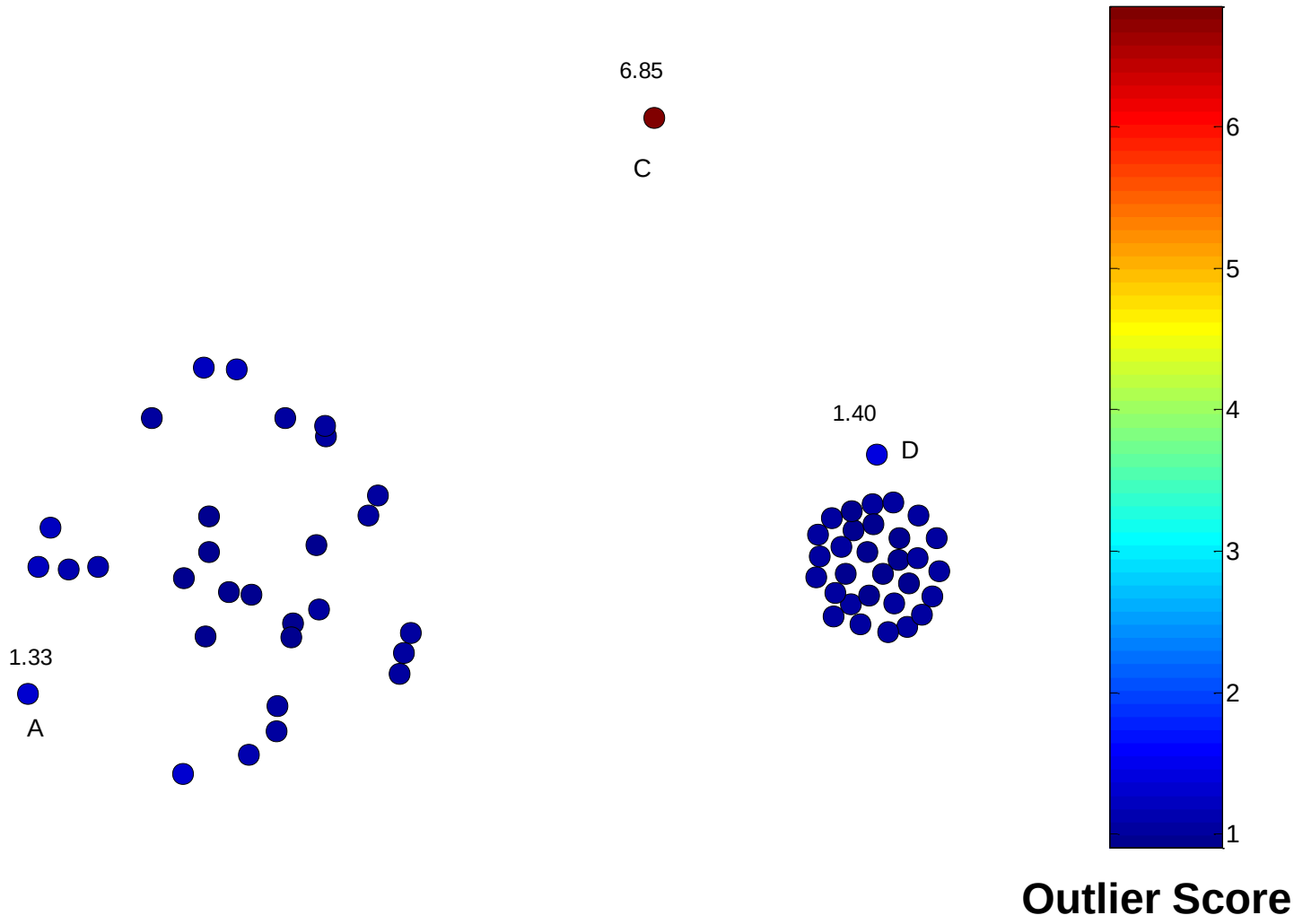
- Consider the density of a point relative to that of its k nearest neighbors

$$average\ relative\ density(\mathbf{x}, k) = \frac{density(\mathbf{x}, k)}{\sum_{\mathbf{y} \in N(\mathbf{x},k)} density(\mathbf{y}, k)/|N(\mathbf{x}, k)|}. \quad (10.7)$$

Algorithm 10.2 Relative density outlier score algorithm.

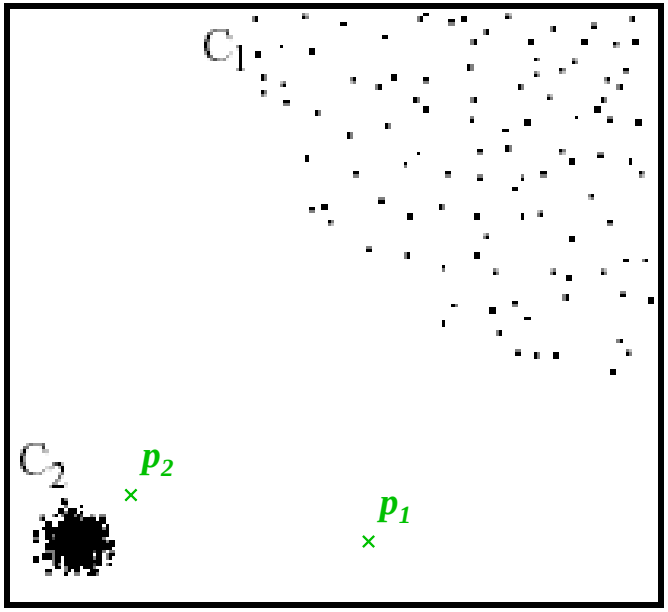
1: $\{k$ is the number of nearest neighbors$\}$
2: **for all** objects $\mathbf{x}$ **do**
3: Determine $N(\mathbf{x}, k)$, the k-nearest neighbors of $\mathbf{x}$.
4: Determine $density(\mathbf{x}, k)$, the density of $\mathbf{x}$, using its nearest neighbors, i.e., the objects in $N(\mathbf{x}, k)$.
5: **end for**
6: **for all** objects $\mathbf{x}$ **do**
7: Set the $outlier\ score(\mathbf{x}, k) = average\ relative\ density(\mathbf{x}, k)$ from Equation 10.7.
8: **end for**

Relative Density Outlier Scores



Density-based: LOF approach

- For each point, compute the density of its local neighborhood
- Compute local outlier factor (LOF) of a sample *p* as the average of the ratios of the density of sample *p* and the density of its nearest neighbors
- Outliers are points with largest LOF value



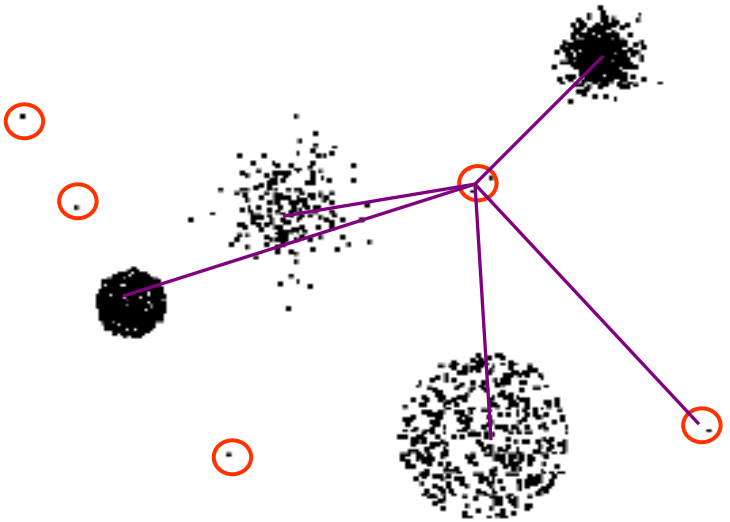
In the NN approach, p_2 is not considered as outlier, while LOF approach find both p_1 and p_2 as outliers

Strengths/Weaknesses of Density-Based Approaches

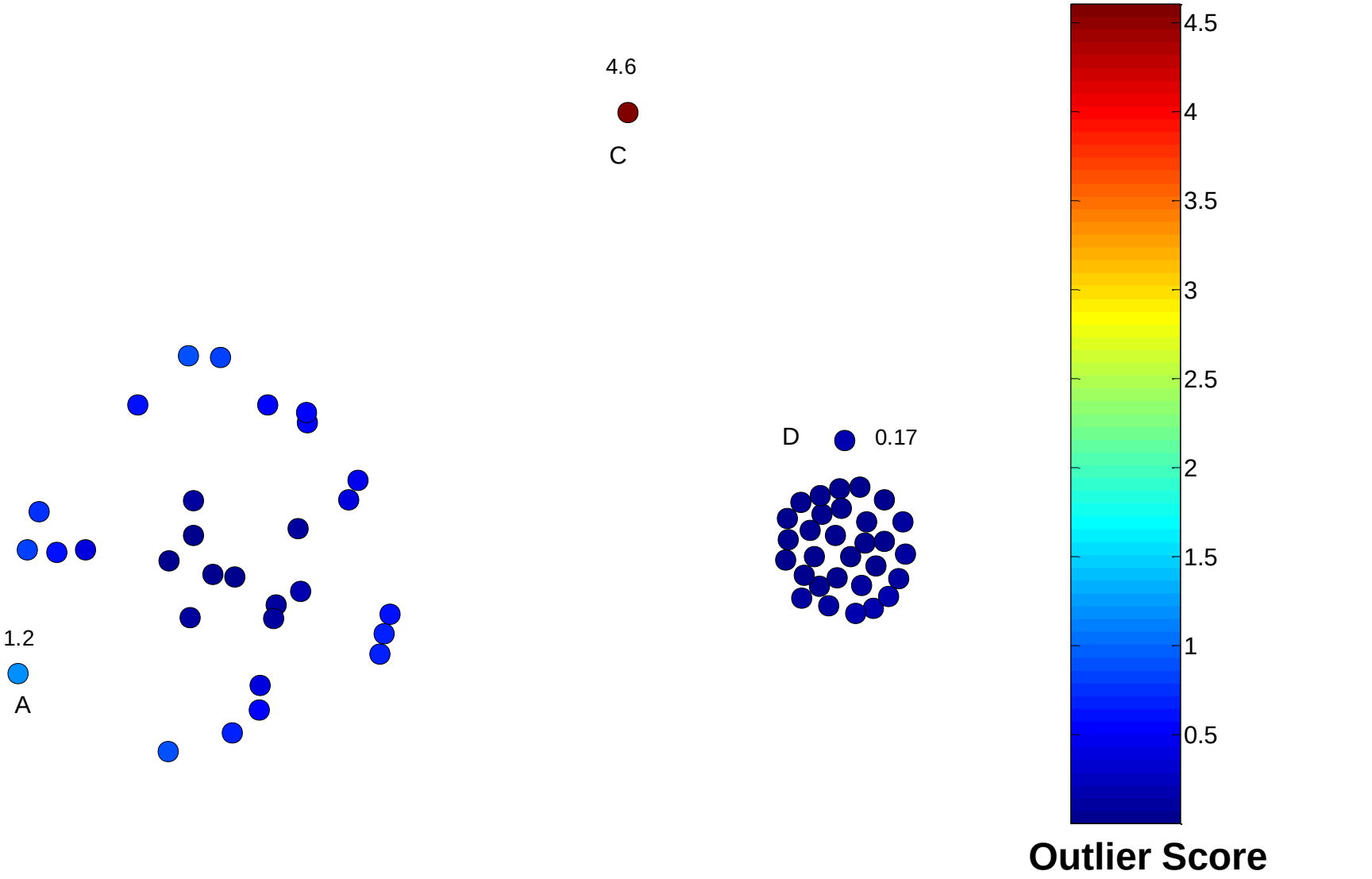
- Simple
- Expensive – $O(n^2)$
- Sensitive to parameters
- Density becomes less meaningful in high-dimensional space

Clustering-Based Approaches

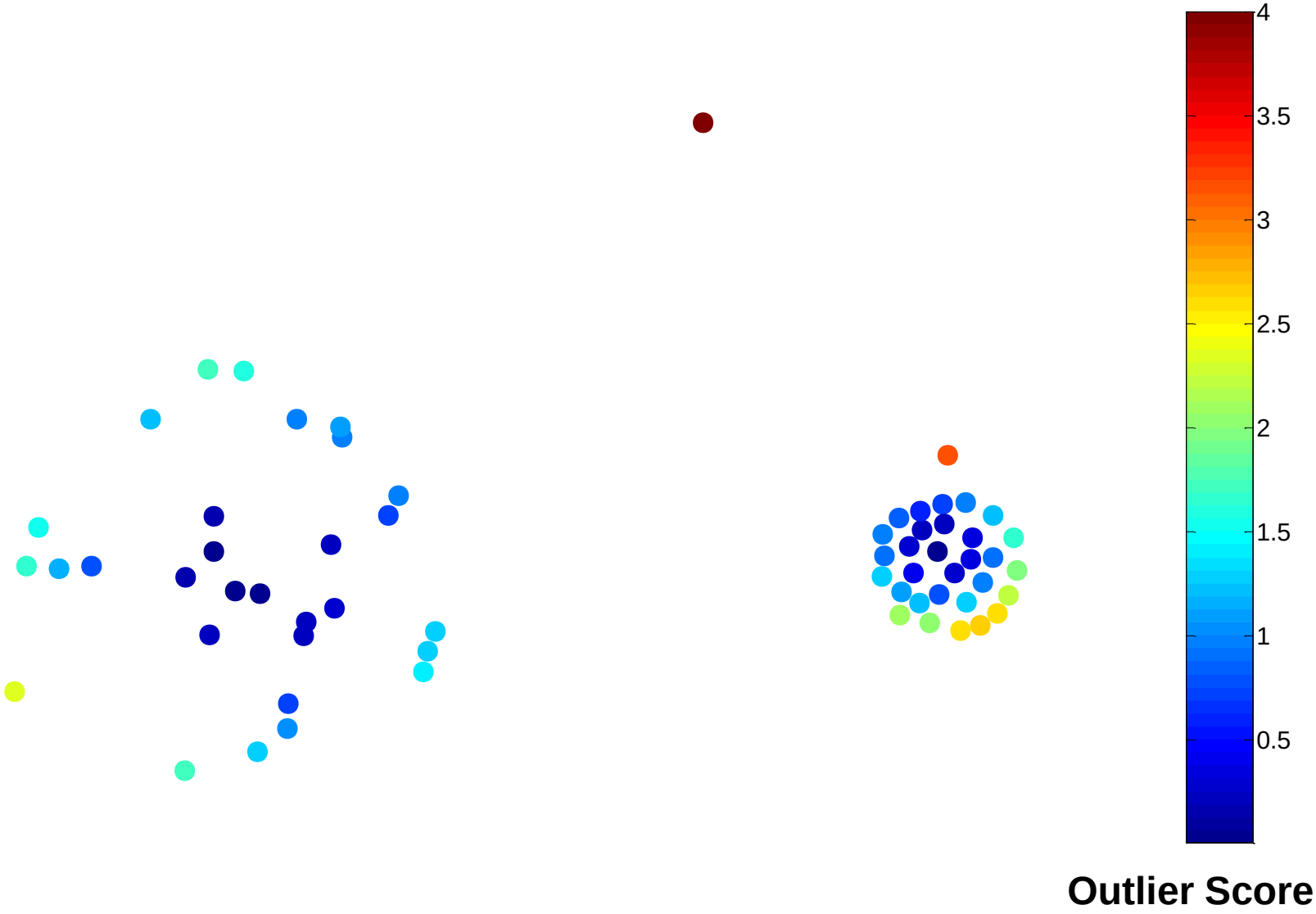
- **Clustering-based Outlier:** An object is a cluster-based outlier if it does not strongly belong to any cluster
  - For prototype-based clusters, an object is an outlier if it is not close enough to a cluster center
  - For density-based clusters, an object is an outlier if its density is too low
  - For graph-based clusters, an object is an outlier if it is not well connected
- Other issues include the impact of outliers on the clusters and the number of clusters



Distance of Points from Closest Centroids



Relative Distance of Points from Closest Centroid



Strengths/Weaknesses of Distance-Based Approaches

- Simple
- Many clustering techniques can be used
- Can be difficult to decide on a clustering technique
- Can be difficult to decide on number of clusters
- Outliers can distort the clusters