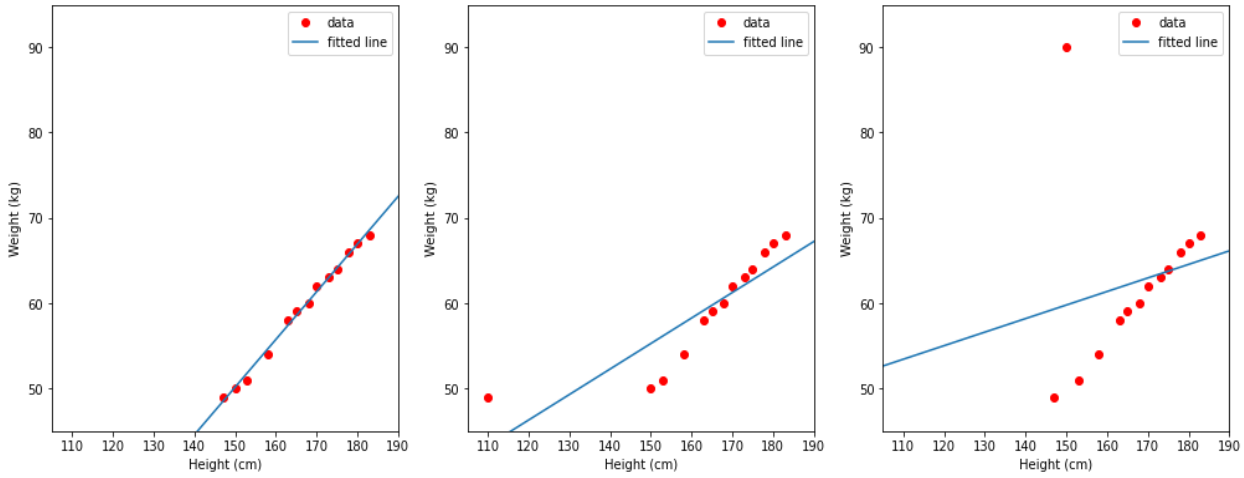
REPORT OUTLIER DETECTION

Outlier

Outlier (điểm dị biệt, điểm ngoại lai) là những mẫu dữ liệu đặc biệt, cách xa so với phần còn lại của dữ liệu. Ví dụ: Với dạng số, dữ liệu ngoại lệ có thể là một giá trị phi thực tế như số tuổi âm, hoặc một giá trị khác xa với phần còn lại của các giá trị trong trường đó. Với dạng hạng mục, dữ liệu ngoại lệ có thể là một giá trị phi thực tế như một hạng mục nằm ngoài những khả năng có thể xảy ra như một địa danh không có trên bản đồ.

Outlier khác với noise (nhiễu). Noise là một sai số ngẫu nhiên hoặc phương sai trong một biến đo được, như là một class bị phân lớp sai hay một lỗi trong một thuộc tính. Đơn giản noise chỉ là những dữ liệu bị lỗi, trong khi đó outlier không chỉ bao gồm các lỗi mà còn bao gồm các dữ liệu không thống nhất (cách xa khỏi các dữ liệu còn lại), các dữ liệu đó có thể vẫn là một dữ liệu hợp lệ. Ví dụ như việc thầy cùng làm bài kiểm tra đột xuất (đề của chính thầy ra) với học sinh. Điểm của thầy sẽ được xem như là một outlier, nhưng nếu điểm thầy bằng 0 thì đó sẽ là một noise. “Nhiễu có thể là điểm ngoại lai nhưng điểm ngoại lai chưa chắc là nhiễu”



Nguyên nhân xuất hiện outlier có thể do:

- Lỗi con người: sai sót trong quá trình nhập liệu.
- Lỗi dụng cụ: sai sót khi đo lường.

- Lỗi thực nghiệm: sai sót trong quá trình thu thập, xử lý, phân tích hay lấy mẫu dữ liệu
- Tự nhiên: có thể không phải lỗi, chỉ là một ngẫu nhiên trong phân phối hay là một thuộc tính mới của dữ liệu.

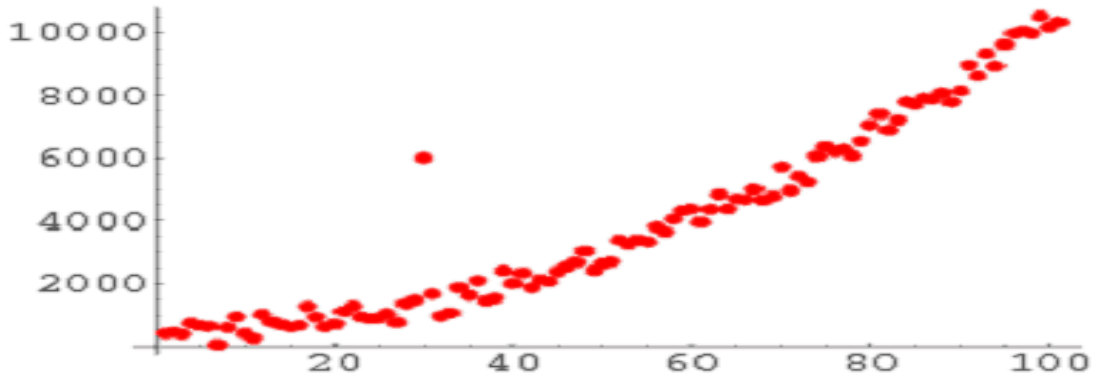
Các loại outlier

Ta có thể chia outlier theo số lượng biến:

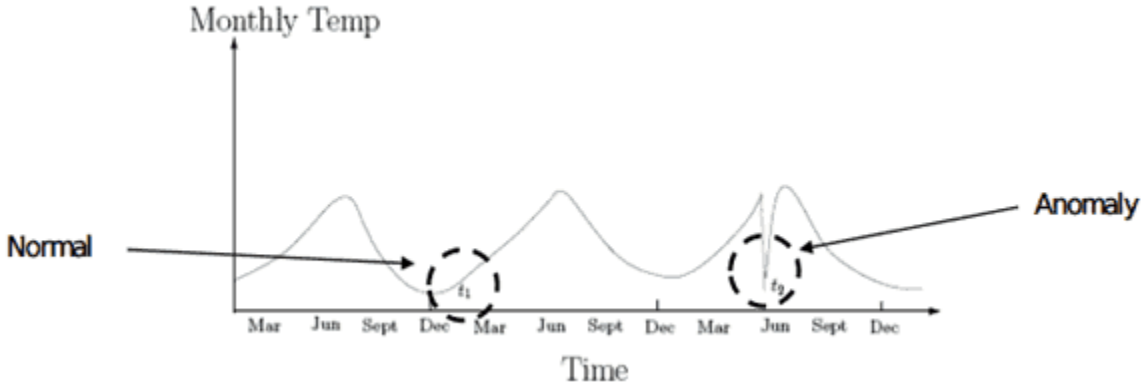
- **Univariate outlier**: điểm dị biệt khi xét trong không gian đặc trưng 1 biến riêng lẻ (1-D). Ví dụ, chúng ta có ba câu hỏi hỏi về giới tính, thâm niên làm việc, sự hài lòng trong công việc, thì ba câu này sẽ phân tích điểm dị biệt tách riêng nhau, không có sự liên quan nào giữa ba câu hỏi. Do đó, kết quả chúng ta sẽ có điểm dị biệt của biến giới tính, điểm dị biệt của biến thâm niên, điểm dị biệt của biến sự hài lòng.
- **Multivariate outlier**: điểm dị biệt khi xét trong không gian đặc trưng kết hợp nhiều biến (n-D). Ví dụ, khi chúng ta xem xét mối quan hệ giữa thâm niên làm việc và sự hài lòng, sẽ có những điểm dị biệt xuất phát từ sự kết hợp giữa hai biến này với nhau. Điểm dị biệt này có thể trùng với điểm dị biệt đơn lẻ của mỗi biến.

Ngoài ra ta có thể phân loại điểm ngoại lai như sau:

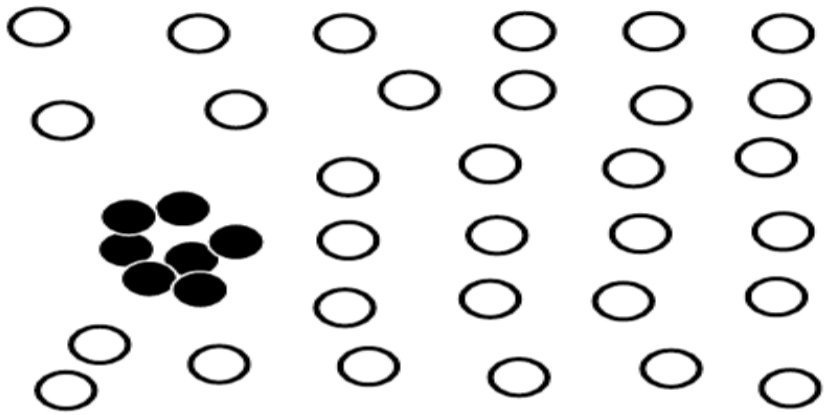
- **Point outlier (Global outlier)**: đây là loại bất thường so với toàn bộ tập dữ liệu. Có thể hiểu đây như là một điểm dữ liệu ở rất xa trong một phân phối, ví dụ như giá trị đơn hàng cao hẳn hay thấp hẳn so với bình thường.



- **Contextual outlier (conditional outlier)**: Trường hợp giá trị của field đang xét tới vẫn là bình thường nếu theo tiêu chí Global outlier nhưng lại bất thường nếu kèm theo ngữ cảnh hay điều kiện cụ thể nào đó. Ví dụ vào tháng 6, số user hoạt động đột ngột giảm mạnh là bất thường (so với tại các thời điểm khác số user đó lại là bình thường).



- **Collective outlier**: Loại outlier này xét trên một tập hợp các điểm dữ liệu mà có sự khác biệt so với toàn dữ liệu thay vì từng điểm riêng biệt. Ví dụ: doanh thu 1 ngày 2000-2500$ thì bình thường, nhưng doanh thu 5 ngày liên tiếp là 2000-2500$ là bất thường.



Tại sao cần phát hiện Outlier

Các điểm dị biệt xuất hiện trong bộ dữ liệu sẽ làm ảnh hưởng các ước lượng thống kê (trung bình, độ lệch chuẩn) hay độ chính xác trong kết luận của mô hình dựa trên thống kê. Ta sẽ cần phải loại bỏ các điểm dữ liệu như vậy ra khỏi

bộ dữ liệu để đảm bảo độ ổn định của kết luận từ nhóm quan sát mà chúng ta quan tâm. Tuy nhiên không phải outlier nào cũng là một lỗi (mang đến tác dụng xấu), outlier cũng có thể chứa các tiềm năng về dữ liệu để chúng ta nghiên cứu sâu hơn. Vì thế nên việc phát hiện outlier để đưa ra phương án xử lý thích hợp là rất cần thiết trong quá trình khai thác dữ liệu.

Phát hiện outlier có rất nhiều ý nghĩa và ứng dụng trong rất nhiều lĩnh vực:

- Fraud detection: phát hiện sai phạm trong việc sử dụng thẻ tín dụng hay thuê bao điện thoại.
- Loan processing: phát hiện hồ sơ khách hàng có vấn đề.
- Intrusion detection: phát hiện xâm nhập trái phép vào mạng máy tính.
- Network performance monitoring: giám sát hoạt động của mạng máy tính, chẳng hạn để phát hiện tắc nghẽn mạng.
- Medical: ứng dụng trong phân tích y tế, như giám sát sự sống, phát hiện tác dụng phụ của thuốc trên người.
- Marketing: xác định bản chất chi tiêu của khách hàng, xác định bất thường trong hành vi của khách hàng.
- Phát hiện những dòng dữ liệu không mong muốn trong cơ sở dữ liệu

Khó khăn

Mặc dù có rất ứng dụng hữu ích, nhưng việc phát hiện outlier không phải là một việc đơn giản và phải đối diện với những thách thức cần phải giải quyết:

- **Xác định được các điểm bình thường và điểm ngoại lai một cách hiệu quả:** Việc phân biệt rạch ròi dữ liệu bình thường với outlier không phải là một việc đơn giản, việc xây dựng một mô hình toàn diện về tính chuẩn của dữ liệu là rất khó khăn, nếu không muốn nói là không thể.Biên giới giữa tính bình thường của dữ liệu và sự bất thường (ngoại lệ) thường không rõ ràng, chúng chỉ là những vùng xám nhập nhằng, khó phân định. Do đó, trong khi có những phương pháp phát hiện outlier tập trung vào việc gán nhãn “bình thường” hay “ngoại lai” cho dữ liệu, thì lại có những phương pháp tiến hành đo lường “sự ngoại lai” của đối tượng.

- **Xác định outlier phụ thuộc vào đặc thù của bài toán:** Việc lựa chọn cách đo độ tường đồng/khoảng cách giữa các đối tượng, mô hình để biểu diễn các đối tượng dữ liệu, xác định số lượng thuộc tính và các thông tin nền (ngữ cảnh, yêu cầu bài toán) để xác định outlier đóng vai trò rất quan trọng trong việc phát hiện ngoại lệ. Thật không may, những lựa chọn này lại phụ thuộc vào bài toán đang phải giải quyết, những bài toán khác nhau sẽ có những yêu cầu riêng trong việc phát hiện outlier. Ví dụ, trong phân tích dữ liệu y tế, một sai lệch nhỏ cũng đủ quan trọng để ta coi đó là một outlier. Ngược lại, trong phân tích thị trường, các đối tượng thường chịu sự biến động lớn hơn, và do đó cần có một độ lệch lớn hơn đáng kể để biện minh cho một giá trị ngoại lệ.

- **Xử lý nhiễu**: Dữ liệu ngoại lai khác với dữ liệu nhiễu. Trong thức tế, chất lượng của tập dữ liệu thường thấp vì quá trình thu thập thường lẫn dữ liệu nhiễu dưới dạng là một giá trị bị sai lệch hay thiếu. Chất lượng dữ liệu thấp và sự hiện diện của nhiễu mang lại thách thức lớn cho việc phát hiện ngoại lệ. Chúng làm sai lệch dữ liệu, làm mờ sự phân biệt giữa các đối tượng bình thường và các đối tượng ngoại lai. Không những thế, nhiễu và dữ liệu bị thiếu có thể “che giấu” các điểm ngoại lệ và làm giảm hiệu quả của việc phát hiện ngoại lệ - một ngoại lệ có thể “ngụy trang” thành một điểm nhiễu và việc phát hiện ngoại lệ có thể xác định nhầm một điểm nhiễu là một điểm ngoại lệ.

- **Hiểu được phương pháp xác định outlier**: ta không chỉ muốn phát hiện ra các ngoại lệ, mà còn cần hiểu tại sao các đối tượng được phát hiện lại là ngoại lệ. Để đáp ứng yêu cầu về tính dễ hiểu, một phương pháp phát hiện ngoại lệ phải cung cấp một số lý do cho việc phát hiện. Ví dụ, một phương pháp phát hiện dựa trên thống kê có thể được sử dụng để xác định mức độ mà một đối tượng có thể là một ngoại lệ dựa trên khả năng đối tượng được tạo ra bởi cùng một cơ chế đã tạo ra phần lớn dữ liệu. Khả năng xảy ra càng nhỏ, đối tượng càng có nhiều là ngoại lệ.

Phương pháp xác định outlier

Trước khi bắt tay vào xác định outlier, ta cần phải xác định được ngữ cảnh của bài toán và cố gắng trả lời câu hỏi sau “Tại sao ta cần phải xác định outlier ?”, “Đây sẽ là univariate hay multivariate outlier ?”, ”Thuộc tính nào cần quan tâm

trong bài toán nay ?". Những câu hỏi trên rất quan trọng để ta xác định được phương pháp thích hợp trong những phương pháp sẽ được giới thiệu ở dưới.

Hiện tại, có rất nhiều phương pháp xác định outlier, ta có thể phân loại chúng theo cách mà các phương pháp đưa ra giả định về outlier so với phần còn lại của dữ liệu.

- **Nhóm phương pháp dựa trên thống kê (Statistical Methods):** Nhóm phương pháp thống kê (nhóm phương pháp dựa trên model - model-based) xác định outlier dựa trên các giả định về tính chuẩn (normality) của dữ liệu. Các dữ liệu bình thường sẽ tuân theo một mô hình và các outlier sẽ là những điểm dữ liệu không tuân theo mô hình này. Tính hiệu quả của các phương pháp thống kê phụ thuộc nhiều vào việc mô hình thống kê có đúng với tập dữ liệu đã cho hay không.
  Ngoài ra, việc xác định các đối tượng bất thường và bình thường có thể được xem như là xác định hai lớp riêng biệt, ta có thể áp dụng các **kỹ thuật phân loại (Classification-based Methods)** để xây dựng mô hình của hai lớp này. Tất nhiên, các kỹ thuật phân loại chỉ có thể được sử dụng nếu các nhãn lớp có sẵn một tập huấn luyện để xây dựng mô hình phân loại dữ liệu. Ngoài ra, các dị thường tương đối hiếm và điều này cần được tính đến khi lựa chọn cả kỹ thuật phân loại và các biện pháp đánh giá mô hình.

- **Nhóm phương pháp dựa trên vùng lân cận (Proximity-Based Methods):** nhóm này xác định một đối tượng là một ngoại lệ nếu các lân cận gần nhất của đối tượng ở xa trong không gian đặc trưng - khoảng cách của đối tượng với các đối tượng lân cận của nó lệch đáng kể so với khoảng cách của hầu hết các đối tượng khác với các đối tượng lân cận của chúng trong cùng một bộ dữ liệu. Hiệu quả của các phương pháp này phụ thuộc rất nhiều vào cách xác định khoảng cách (không dễ dàng để có được cách xác định phù hợp với bài toán). Hơn nữa, các phương pháp dựa trên vùng lân cận thường gặp khó khăn trong việc phát hiện một nhóm các ngoại lệ gần nhau.

- **Nhóm phương pháp dựa vào phân cụm (Clustering-Based Methods):** xác định outlier bằng cách tìm các điểm dữ liệu thuộc những cụm (cluster) nhỏ, thưa thớt hoặc không thuộc bất cứ cụm nào. Phân cụm là một hoạt động khai thác dữ liệu tốn kém. Việc điều chỉnh phương pháp phân cụm

để phù hợp với phát hiện ngoại lệ có thể rất tốn kém, dẫn đến khó khăn trong quá trình mở rộng quy mô cho các tập dữ liệu lớn.

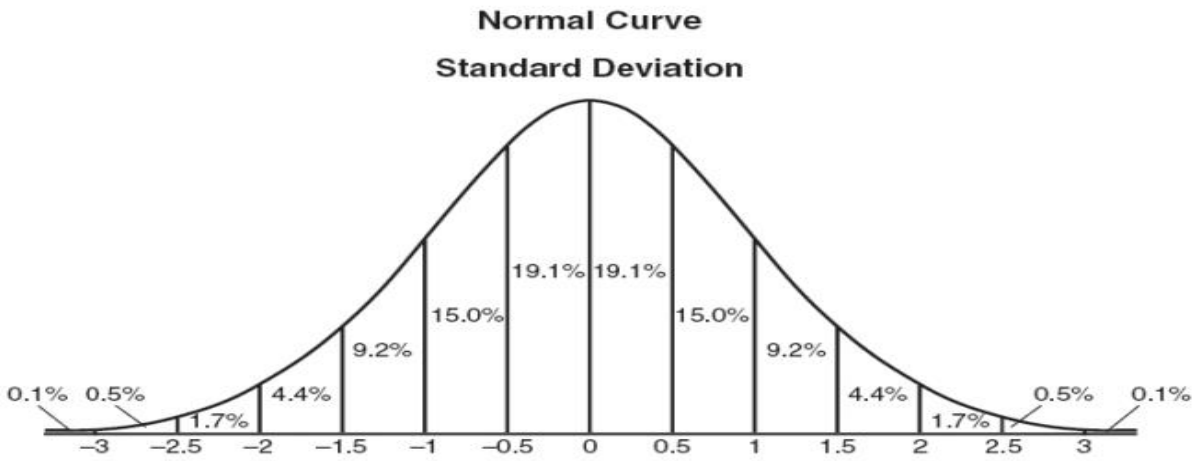
Phương pháp dựa trên thống kê

Như đã nói ở trên, nhóm phương pháp này sẽ xác định outlier dựa trên tính chuẩn của dữ liệu. Ý tưởng đằng sau các phương pháp thống kê để phát hiện outlier là tìm một mô hình tổng quát phù hợp với tập dữ liệu đã cho, và sau đó xác định các đối tượng đó trong các vùng có xác suất thấp của mô hình là outlier.

Đối với những phương pháp thuộc nhóm này, ta cần phải xác định được dạng phân phối thích hợp với tập dữ liệu của bài toán (Gaussian, Poisson, nhị thức, ...). Ngoài ra, ta cần phải quan tâm đến số lượng thuộc tính để xác định outlier (một số phương pháp chỉ dành riêng cho đơn biến - univariate, trong khi số còn lại dành cho đa biến - multivariate), ta cũng có thể phối hợp nhiều dạng phân phối với nhau để thích hợp với bài toán.

Phân phối chuẩn

Phân phối chuẩn (Gauss distribution, Normal distribution), phụ thuộc vào hai tham số μ (giá trị trung bình) và σ (độ lệch chuẩn), ký hiệu là N(μ, σ). Trong đó, μ thể hiện kỳ vọng hay giá trị mong muốn của phân phối, σ cho ta thông tin về mức độ phân tán xác suất.



Trong phân phối xác suất, người ta còn sử dụng khái niệm điểm chuẩn (z-score) thể hiện cho khoảng cách từ một điểm tới điểm trung bình của theo đơn vị là độ lệch chuẩn, có công thức như sau *z-score* = $\frac{x-\mu}{\sigma}$.

Những điểm là outlier trong phân phối chuẩn, sẽ là những điểm dữ liệu mà có xác suất xảy ra thấp.

Maximum Likelihood Estimation (MLE)

Ta có thể sử dụng MLE để ước lượng các hệ số μ và σ theo công thức sau:

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Trong đó n là số mẫu trong bộ dữ liệu. Ta sẽ coi những giá trị cách xa μ một đoạn lớn hơn 3σ sẽ là những giá trị outlier.

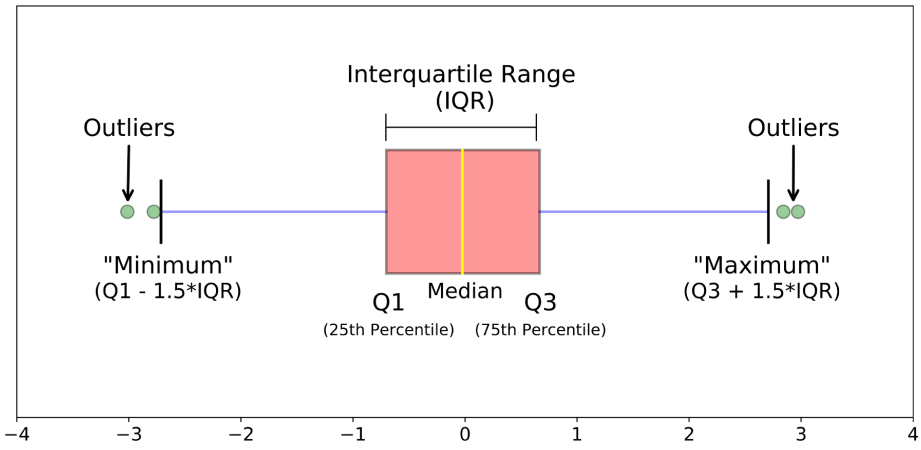
Ví dụ: nhiệt độ của một thành phố tại tháng 7, trong vòng 10 năm, được sắp xếp tăng dần như sau.

Nhiệt độ (◦C)	24	28.9	28.9	29	29.1	29.1	29.2	29.2	29.3	29.4

Ta có thể dễ dàng tính được μ = 28.61 và σ = 1.51 theo công thức. Xét giá trị phân cực nhất là 24, cách giá trị trung bình 4.61, 4.61 / σ > 3. Theo phân phối chuẩn trong đoạn μ ± 3σ chứa 99.7% mẫu dữ liệu, xác suất để 24 xuất hiện trong phân phối chuẩn sẽ thấp hơn 0.15%, nên ta có thể coi 24 là một outlier.

Interquartile Range (IQR)

Boxplot là một phương pháp để biểu diễn dữ liệu dựa vào 5 giá trị trong tập dữ liệu, bao gồm: min, Q1 (điểm phần tư dưới), Q2 (trung vị), Q3 (điểm phần tư trên), max.



IQR được định nghĩa là nằm trong đoạn [Q1,Q3], |IQR| = |Q3 - Q1|, 99.3% dữ liệu nằm trong đoạn [Q1 - 1,5 x IQR, Q3 + 1,5 x IQR] (giống với MLE trong khoảng μ ± 3σ), những dữ liệu nằm ngoài khoảng trên sẽ được coi như là outlier.

Grubb's test (Maximum normed residual test)

Trong thực tế, ta không thể xác định μ và σ cho toàn bộ quần thể dữ liệu. Thay vào đó ta sẽ tính z-score, như sau:

$$z = \frac{|x - \bar{x}|}{s},$$

Trong đó $\bar{x}$ là giá trị trung bình, s là độ lệch chuẩn của tập dữ liệu đang xét. Lúc này giá trị x của dữ liệu, sẽ là outlier nếu:

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

Với $t_{\alpha/(2N),N-2}$ là giá trị của t-distribution với bậc α/(2N) và N là số lượng mẫu trong tập dữ liệu.

Dữ liệu liên quan đến hai hoặc nhiều thuộc tính hoặc biến là dữ liệu đa biến. Nhiều phương pháp phát hiện ngoại lệ đơn biến có thể được mở rộng để xử lý dữ liệu đa biến. Ý tưởng để xử lý với dữ liệu đa biến là chuyển đổi nhiệm vụ phát hiện ngoại lệ đa biến thành một vấn đề phát hiện ngoại lệ đơn biến.

Khoảng cách Mahalanobis

Khoảng cách Mahalanobis là một độ đo đa biến có thể biểu diễn hiệu quả khoảng cách của một điểm dữ liệu tới một phân phối.

Cho một tập dữ liệu đa biến, có $\bar{x}$ là vector trung bình, với từng điểm dữ liệu x, khoảng cách Mahalanobis từ x đến $\bar{x}$ được định nghĩa bằng công thức sau:

$$MDist(x, \bar{x}) = (x - \bar{x})^T S^{-1} (x - \bar{x}), với\ S\ là\ ma\ trận\ hiệp\ phương\ sai$$

Sau khi tính khoảng cách Mahalanobis, ta có thể áp dụng các phương pháp xác outlier cho dữ liệu đơn biến, chuyển vấn đề từ đa biến sang đơn biến. Nếu $MDist(x, \bar{x})$ được xác định là 1 outlier thì, x sẽ là một outlier.

Chi-Square Statistic ($\chi 2$ -statistic)

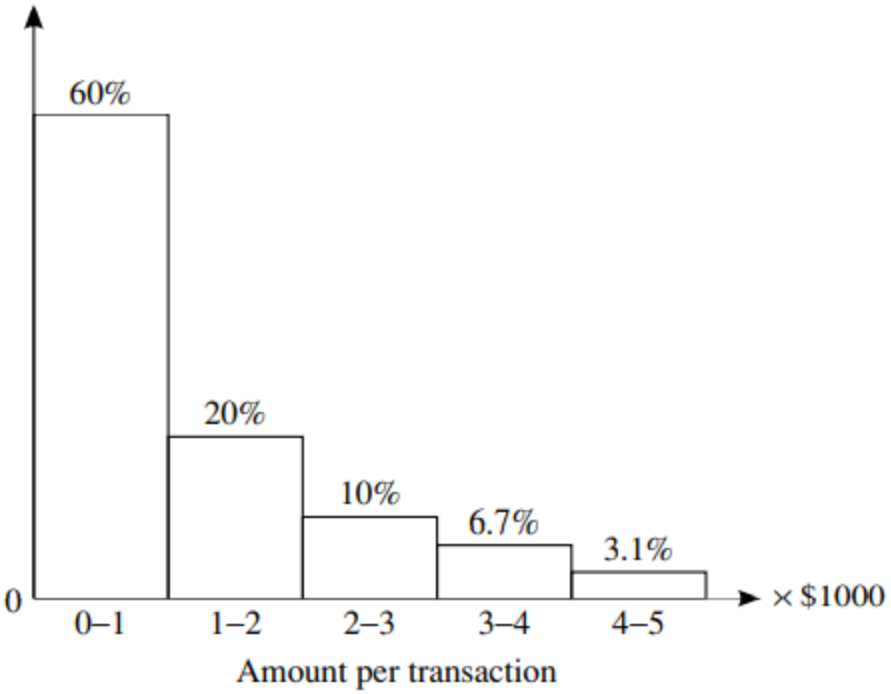
Chi-square được dùng để đo lường sự khác biệt giữa tần số quan sát và tần số mong đợi của các kết quả của một tập hợp các sự kiện hoặc biến số. Chi-square được tính như sau:

$$\chi^2 = \sum_{i=1}^{n} \frac{(o_i - E_i)^2}{E_i},$$

Với o_j là giá trị của dữ liệu o ở không gian j, E_j là giá trị trung bình của không gian j, n là số chiều dữ liệu. Nếu dữ liệu có chi-square lớn, thì có thể xem như là một outlier.

Histogram

Histogram là một mô hình thống kê phi tham số được sử dụng thường xuyên có thể được sử dụng để phát hiện các ngoại lệ. Lưu ý rằng mặc dù các phương pháp phi tham số không cần bất kỳ mô hình thống kê tiên nghiệm nào, nhưng chúng thường yêu cầu các tham số do người dùng chỉ định để học các mô hình từ dữ liệu, đối với histogram nói riêng thì người dùng cần phải xác định loại histogram, kích thước bin và số lượng bin.



Biểu đồ thể hiện giá tiền trên từng giao dịch của khách hàng.

Quá trình xác định outlier từ histogram bao gồm 2 bước chính:

1. Tạo histogram từ dữ liệu đầu vào. Ta cần quan tâm đến việc chọn kích thước trùng bin phù hợp với tập dữ liệu (nếu quá nhỏ có thể nhầm dữ liệu thường với outlier, nếu quá lớn có thể không chọn được outlier).
2. Xác định outlier. Để có thể xác định outlier, ta có thể tính khả năng trở thành outlier dựa trên dữ liệu. Ví dụ: Theo histogram ở trên một giao dịch lớn hơn 5000$ chỉ chiếm khoảng 0.2% số lượng giao dịch, vậy một giao dịch có giá trị cỡ 7500$ sẽ có khả năng là một outlier nhiều hơn một giao dịch có giá trị 385$ khoảng $\frac{1}{0.2\%} \div \frac{1}{60\%} = 300$ lần, ta có thể xem đó như là một outlier.

Phương pháp dựa trên vùng lân cận

Cho một tập các đối tượng trong không gian đặc trưng, một thước đo khoảng cách có thể được sử dụng để định lượng mức độ giống nhau giữa các đối tượng. Mặc dù có một số biến thể về ý tưởng phát hiện điểm bất thường dựa trên vùng lân cận nhưng về cơ bản đều dựa trên nhận xét “một đối tượng là dị thường nếu nó ở xa hầu hết các điểm”.

Cách tiếp cận này tổng quát hơn và dễ áp dụng hơn các cách tiếp cận thống kê, vì nó dễ dàng xác định một thước đo khoảng có ý nghĩa cho một tập dữ liệu hơn là xác định phân phối thống kê của nó.

Distance-Based

Cho tập dữ liệu D, người dùng đặt threshold cho khoảng cách r, đối với mỗi đối tượng dữ liệu o, ta sẽ đếm số đối tượng dữ liệu nằm trong vùng lân cận r của o. Nếu hầu hết các đối tượng khác xa o hơn một khoảng r, nghĩa là o chính là một outlier. Ta có thể định nghĩa cách làm này bằng công thức như sau:

$$\frac{số\ đối\ tượng\ nằm\ trong\ khoảng\ r\ của\ o}{số\ đối\ tượng\ trong\ D} = n,$$

với n là threshold $0 \leq n \leq 1$, r là khoảng cách vùng lân cận $r \geq 0$

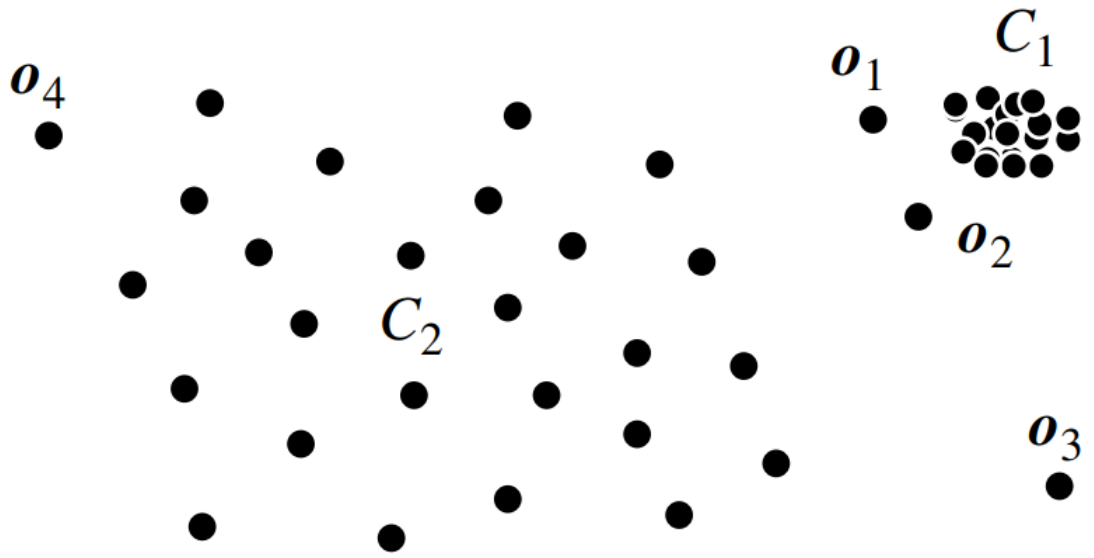
Một cách tương đương, ta có thể xác định một điểm o, có là outlier hay không bằng cách kiểm tra xem khoảng cách từ o đến k hàng xóm gần nhất (k-nearest neighbor) o_k, với k = n ||D||. o sẽ là một outlier nếu có một hàng xóm thỏa mãn dist(o,o_k) > r.

Để cài đặt cho thuật toán này, ta sẽ cần một vòng lặp duyệt qua từng phần tử i trong tập D, một vòng lặp con duyệt qua các phần tử j còn lại để xác định xem i có phải là một outlier hay không, bằng cách tính khoảng cách d từ i đến j, nếu số phần tử có d < r nhiều hơn k, thì i sẽ là một phần tử bình thường, nếu duyệt hết j mà vẫn không đủ k phần tử thỏa mãn, thì i sẽ là một outlier.

Một vài vấn đề phải cân nhắc là độ phức tạp thuật toán là O(n^2) và số lượng phần tử giữ trong bộ nhớ, có thể gặp phải những vấn đề về tốc độ khi gặp phải dữ liệu lớn. Một vài cải tiến có thể áp dụng là gom nhóm dữ liệu theo dựa theo độ lân cận và duyệt theo từng nhóm phần tử (CELL - Grid-based).

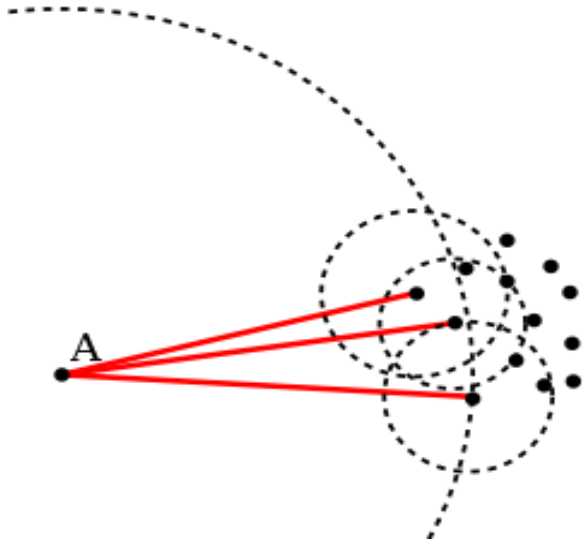
Density-Based

Đối với Distance-Based, các outlier mang tính toàn cục so với tập dữ liệu. Vì một điểm là một outlier khi và chỉ khi nó đạt đủ số lượng điểm dữ liệu cách xa, số lượng cần thiết được xác định bởi 2 biến toàn cục trên cả tập dữ liệu là r và n. Trong thực tế, việc phân bố dữ liệu thường rất phức tạp, nơi các đối tượng có thể được coi là ngoại lai đối với các vùng lân cận địa phương của chúng, thay vì đối với phân phối dữ liệu toàn cục.



Cho một phân bố dữ liệu như trên, bao gồm 2 nhóm dữ liệu C1 (đặc) và C2 (thưa) trong đó ta có o3 là một outlier toàn cục, o1 và o2 là outlier cục bộ của C1 (o1 và o2 gần C1 hơn C2) và o4 là một phần tử bình thường của C1. Nếu xét theo Distance-Based, ta chỉ có thể xác định được o3 là một global outlier. Ta cần phải tìm ra một cách để xác định o1 và o2 là một outlier so với C1 mà vẫn đảm bảo các o4 và dữ liệu trong C2 là bình thường.

Một cách để giải quyết vấn đề này là dựa vào mật độ (density) xung quanh điểm dữ liệu - mật độ xung quanh một đối tượng ngoại lai khác đáng kể so với mật độ xung quanh các đối tượng lân cận của nó. Ta sẽ sử dụng mật độ tương đối của một đối tượng so với các đối tượng lân cận của nó để chỉ ra mức độ mà đối tượng đang xét là một ngoại lệ hay không.



Ý tưởng cơ bản của LOF: so sánh mật độ cục bộ của một điểm với các mật độ của các điểm lân cận (hàng xóm) của nó. A có mật độ thấp hơn nhiều so với mật độ của các điểm lân cận của nó. Lúc này A có khả năng cao là một outlier.

Gọi khoảng từ o đến k lân cận gần nhất (k-NN) là dist_k(o), với k-NN được định nghĩa như sau $N_k(o) = \{o' \mid o' \in D, dist(o, o') \leq dist_k(o)\}$, N_k(o) có thể lớn hơn k. Để có thể xác định mật độ cục bộ quanh o, ta sử dụng khái niệm "khoảng cách có thể tiếp cận được" (Reachability distance), được tính như sau,

$$reachdist_k(o \leftarrow o') = max\{dist_k(o), dist(o, o')\},$$

Với k do người dùng chỉ định.

Sau đó ta sẽ xác định mật độ khả năng tiếp cận cục bộ (local reachability density), đây là một thước đo mật độ k điểm gần nhất xung quanh một điểm được tính bằng cách lấy nghịch đảo của tổng tất cả $reachdist$ của tất cả k điểm lân cận gần nhất. Các điểm càng gần nhau, khoảng cách càng nhỏ và mật độ càng cao.

$$lrd_k(o) = \frac{||N_k(o)||}{Sum(reachdist_k(o \leftarrow o')), \forall o' \in N_k(o)}$$

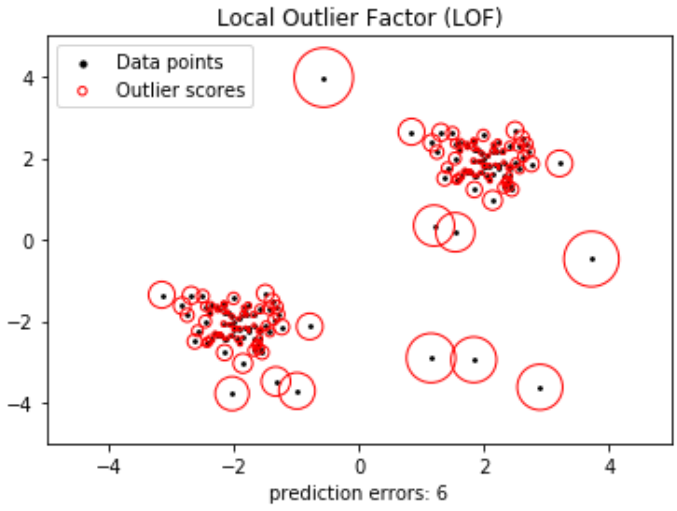
Cuối cùng ta có thể tính được hệ số tách biệt địa phương (local outlier factor) là giá trị trung bình của tỷ số giữa mật độ khả năng tiếp cận cục bộ của o và của

các nước láng giềng gần nhất của o.

$$LOF_k(o) = \frac{\sum_{o' \in N_k(o)} \frac{lrd_k(o')}{lrd_k(o)}}{\|N_k(o)\|} = \sum_{o' \in N_k(o)} lrd_k(o') \cdot \sum_{o' \in N_k(o)} reachdist_k(o' \leftarrow o).$$

Khi mật độ xung quanh o càng thấp và mật độ các điểm k-NN của o càng cao thì LOF càng lớn. Nhờ vậy ta có thể xác định chính xác một ngoại lai cục bộ có mật độ cục bộ thấp so với mật độ cục bộ của k-láng giềng gần nhất của nó. Ngoài ra LOF còn đảm bảo các đối tượng ở trong một cluster (đặc hoặc thưa) sẽ không bị coi như là một outlier. Từ kết quả của LOF ta có thể xác định, mối quan hệ giữa các điểm dữ liệu như sau:

- LOF ~ 1 => Các điểm đang xét tương đồng (cùng nằm trong một cluster phân bố ổn định)
- LOF < 1 => Các điểm đang xét là local inlier (cùng nằm trong một cluster có mật độ đặc)
- LOF > 1 => Điểm đang xét là một local outlier

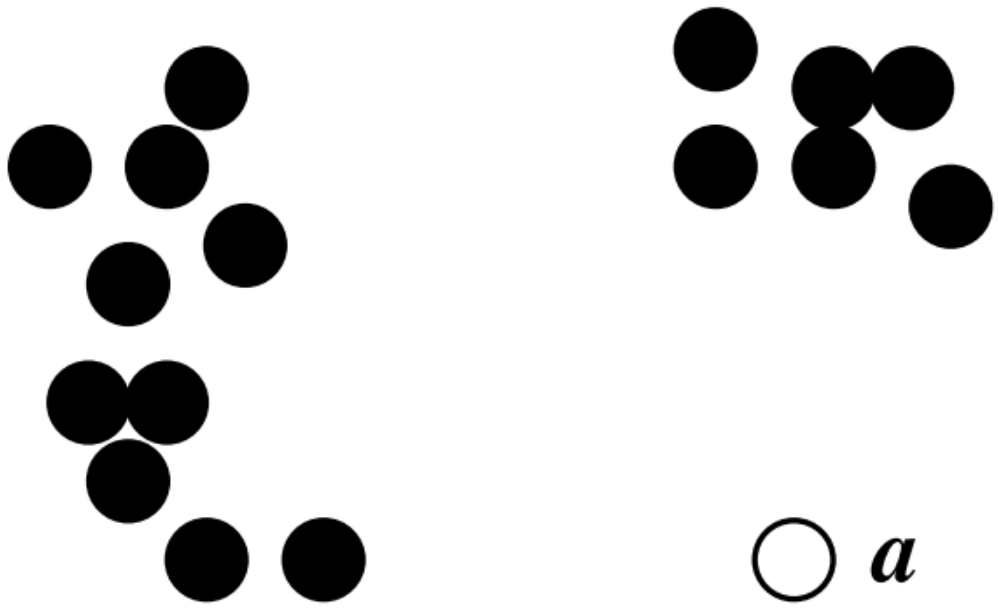


Phương pháp dựa trên phân cụm

Các phương pháp tiếp cận dựa trên phân cụm phát hiện các ngoại lệ bằng cách kiểm tra mối quan hệ giữa các đối tượng và các cụm. Dựa vào đó ta có 3 cách để xác định điểm ngoại lai.

Không thuộc bất kỳ cluster nào

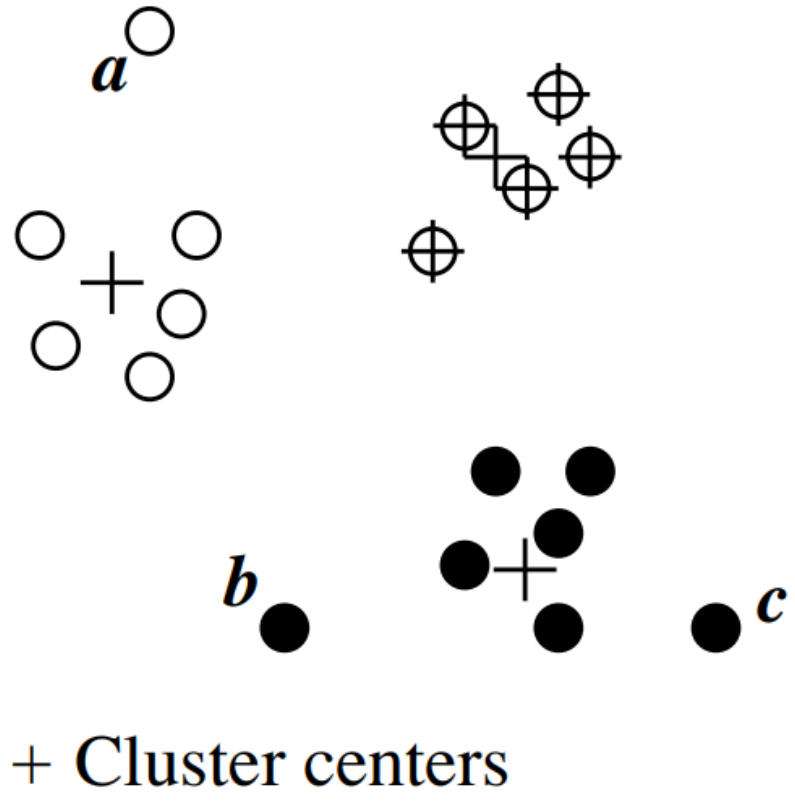
Điểm không nằm trong cluster nào có khả năng là một outlier.
Ta sẽ dựa vào các thuật toán phân cụm dựa vào mật độ như DBSCAN. Điểm dữ liệu nào không thuộc bất kỳ cluster nào sẽ được xem như là một outlier.



Khoảng cách đến cluster gần nhất

Điểm nằm ở xa so với cluster gần nhất với nó có thể là một outlier.
- Sử dụng k-mean để phân cụm dữ liệu, đồng thời xác định trung tâm của từng cụm.
- Với từng điểm dữ liệu o sử dụng tỉ số giữa khoảng cách từ o đến trung tâm cụm gần nhất c, dist(o,c) và trung bình khoảng cách các điểm dữ liệu đến trung tâm cụm là, l. Tỉ lệ dist(o,c) / l càng lớn, o càng ở xa so với cụm dữ liệu gần nhất,

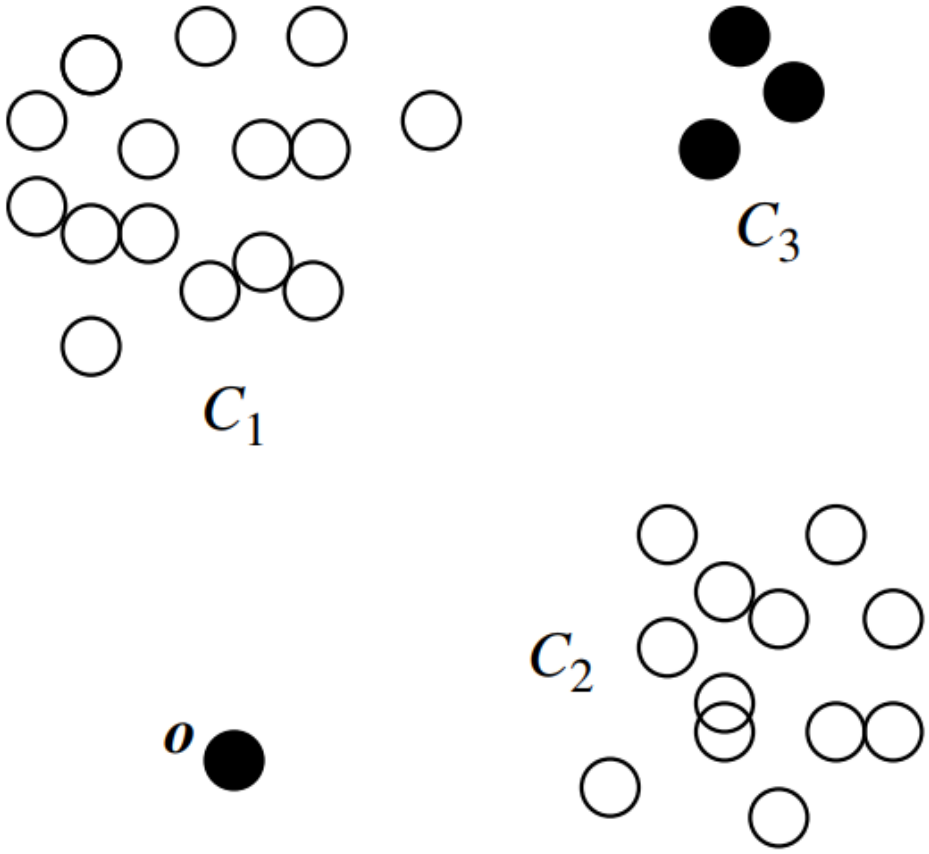
o càng có khả năng là một outlier.



Dựa vào xác định các cluster nhỏ

Nếu điểm dữ liệu thuộc về một cụm nhỏ hoặc thưa, rất có thể cả cụm đó là một cụm outlier.
Ta có thể kiểm tra xem điểm dữ liệu hay một cluster có phải là outlier hay không bằng cách sử dụng CBLOF (cluster based local reachability density). CBLOF có thể cho ta biết sự giống nhau giữa một điểm và một cụm theo cách thống kê thể hiện xác suất điểm đó thuộc về cụm. Quá trình xác định như sau:
- Xác định các cluster ở trong bộ dữ liệu, sau đó sắp xếp chúng theo thứ tự tăng dần.
- Sử dụng hệ số α ($0 \leq \alpha \leq 1$) để phân loại các các cụm dữ liệu lớn (số phần tử trong cụm > α * số phần tử trong dataset) và cụm dữ liệu nhỏ (các cụm còn lại).
- Với từng điểm dữ liệu o, tính CBLOF tương ứng như sau:
    Nếu o thuộc cụm lớn, CBLOF = cluster_size X similarity_o_cluster
    Nếu o thuộc cụm nhỏ, CBLOF = cluster_size X similarity_o_nearestLargeCluster

- Những điểm có CBLOF thấp sẽ là outlier.



Phương pháp xác định dựa trên phần cụm có thể xác định outlier không giám sát (không cần label trên dữ liệu). Nếu đã có sẵn thông tin về các cluster, việc xác định outlier sẽ khá nhanh và không tốn quá nhiều tài nguyên. Tuy nhiên, một vài hạn chế của phương pháp này là phải dựa vào hiệu quả của việc lựa chọn phương pháp gom cụm, dễ bị nghẽn cổ chai khi gặp các tập dữ liệu lớn.

Tổng kết

Ta có thể xem một tập dữ liệu được tạo ra bởi một quy trình thống kê bất kỳ, trong đó outlier (ngoại lai) là những điểm dữ liệu đặc biệt, có sai lệch đáng kể so với phần còn lại của các đối tượng, như thể nó được tạo ra bởi một cơ chế khác. Có nhiều loại outlier, một đối tượng dữ liệu có thể thuộc một hoặc nhiều loại outlier. Có rất nhiều cách để có thể xác định các điểm dữ liệu ngoại lai và ta cần phải xác định rõ các yếu tố liên quan đến bài toán (phân bố dữ liệu, thuộc tính ảnh hưởng đến outlier, mục đích xác định outlier, ...) để có thể chọn ra phương pháp phù hợp nhất. Việc xác định outlier vẫn đang phải đối diện với

những thách thức cần phải giải quyết, nhưng nó cũng đã và đang mang lại những ứng dụng thực tế cho con người.