Introduction of Isolation Forest

Anomaly Detection Series

報告者：吳俊輝

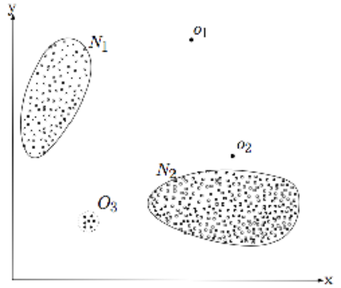
專案背景介紹

- 與洗防合作，希望可以透過關注名單在ATM上的行爲來觀察出新的Pattern (加速調查過程)
- 例如同帳號大量跨縣市提領的行為是否代表著可疑？
- 清晨6點鐘的大量提領是否代表可疑？
- 問題：洗錢是稀少事件、且全行客戶相較於關注名單數目是極度不平衡的數據
- 因此採用 Anomaly Detection方式，嘗試找到 Anomaly 與 名單的交集

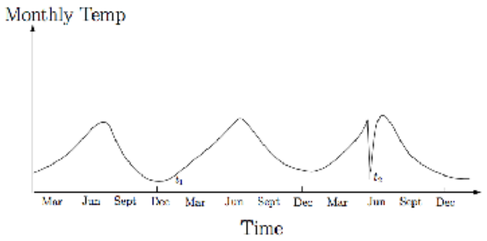
異常偵測

Anomaly Detection 分成三種大類，依需求不同
可以使用不同的方式處理

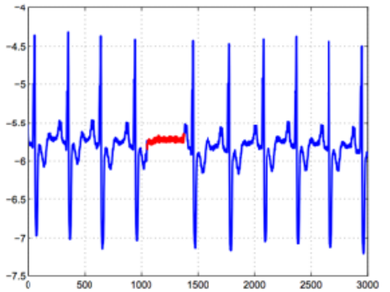
- Point Anomaly Detection
  - 著重在Data點的分佈偵測



- Contextual Anomaly Detection
  - 根據 前後文 決定該點在整串序列中是否可疑



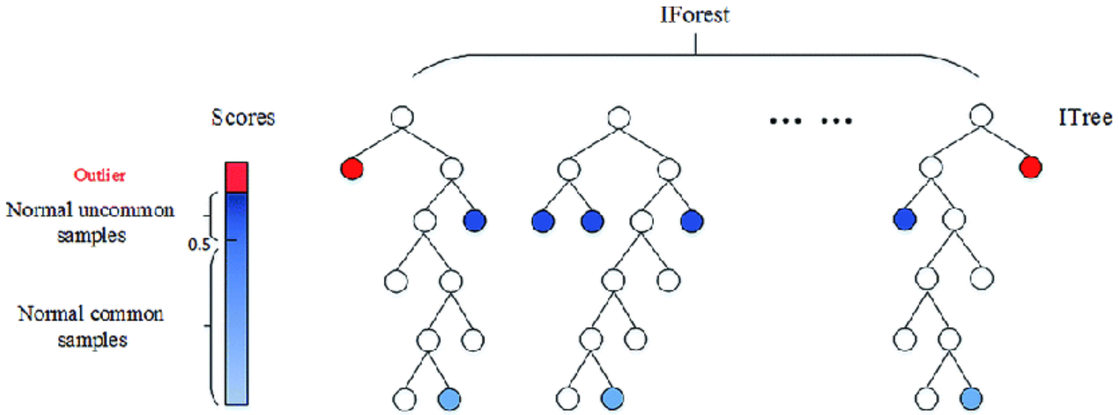
- Collective Anomaly Detection
  - 一串序列中哪些子序列(sub-series)相較於整體序列是可疑的



孤立森林

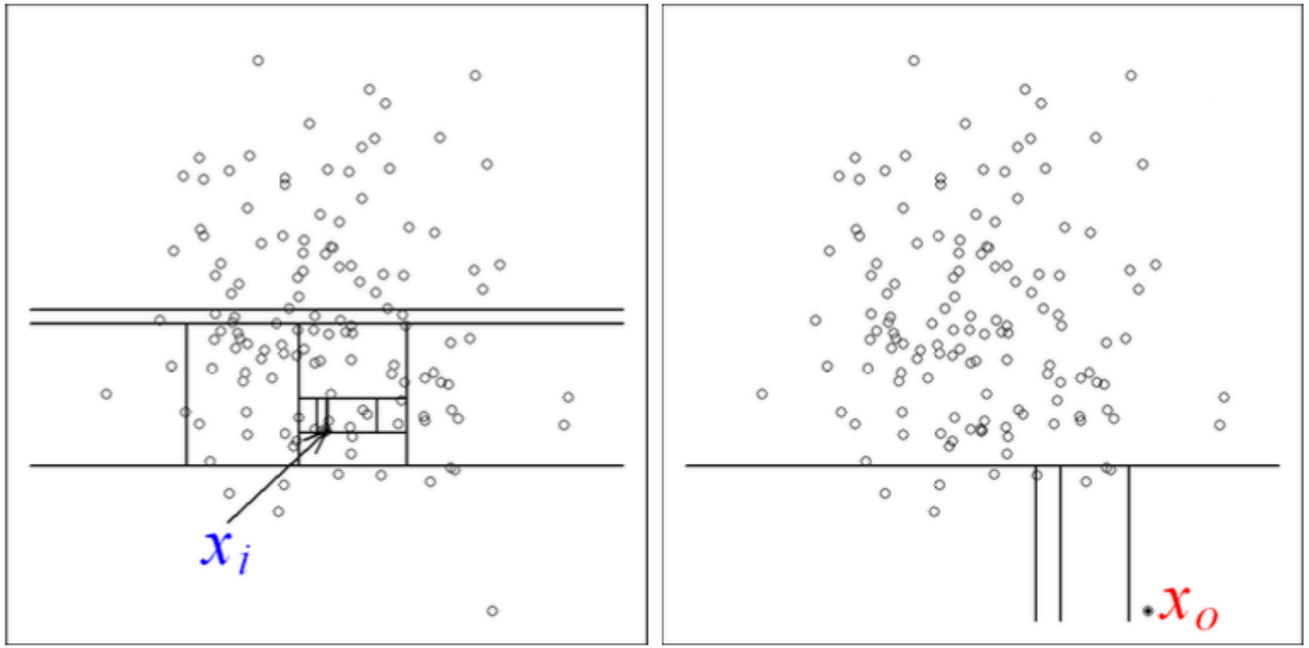
Point Anomaly Detection Method

- 模型假設： 異常值稀少且在feature上與一般的資料不一樣，因此使得異常值容易透過切分的方式被分別出來
- 使用Subsampling，Bagging方式建立模型，因此可以避免Masking，Swamping問題
- Linear Time
- Masking：異常值太多、密度太高導致難以切分
- Swamping：異常值跟正常值太接近難以切分



理論介紹

直覺：異常點在超平面切分後很容易就被區隔出來



(a) Isolating x_i (b) Isolating x_o

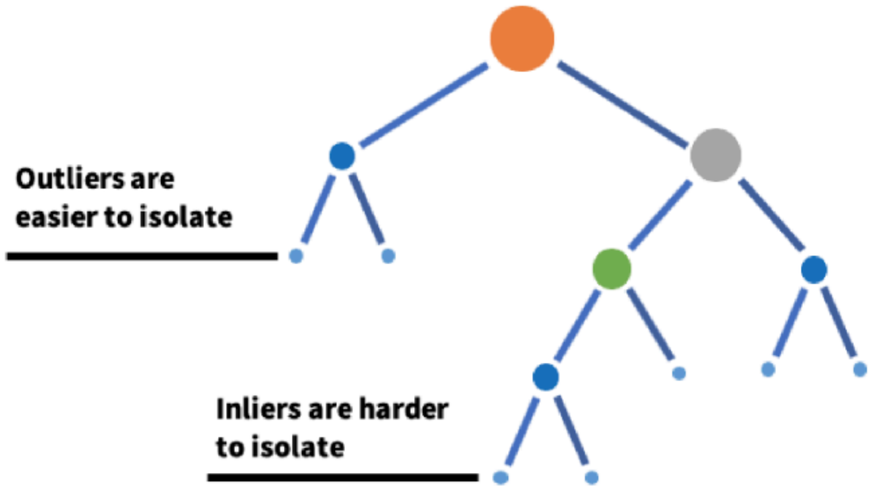
理論介紹

Isolation Tree 的建構方法

1. 假設資料集 $X=\{x_1,x_2 \dots x_n\}$ 特徵維度為d

2. 選取特徵q 並選取切分點 p
   (隨機在特徵q的最大最小範圍內取值) ，
   將X根據該切分點切分為左右節點

3. 若以下三個條件其中之一到達則停止：
   (i) 樹達到限定高度
   (ii) 節點僅剩一筆資料集
   (iii) 節點中得所有數據都有一樣的值



理論介紹

異常分數的計算

- c(n) 代表 n 個 data進入二元搜尋樹的平均搜尋路徑
- 用以作為 Base Line 與期望層數作比較
- 數值越接近1代表越可疑
- 數值越接近 0 代表越不可疑

unsuccessful search in BST. We borrow the analysis from BST to estimate the average path length of iTree. Given a data set of n instances, Section 10.3.3 of [9] gives the average path length of unsuccessful search in BST as:

$$c(n) = 2H(n-1) - (2(n-1)/n), \quad (1)$$

where $H(i)$ is the harmonic number and it can be estimated by $\ln(i) + 0.5772156649$ (Euler's constant). As $c(n)$ is the average of $h(x)$ given n, we use it to normalise $h(x)$. The anomaly score s of an instance x is defined as:

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}}, \quad (2)$$

where $E(h(x))$ is the average of $h(x)$ from a collection of isolation trees. In Equation (2):

- when $E(h(x)) \to c(n)$, $s \to 0.5$;
- when $E(h(x)) \to 0$, $s \to 1$;
- and when $E(h(x)) \to n-1$, $s \to 0$.

D. M. Rocke and D. L. Woodruff. Identification of outliers in multivariate data. Journal of the American Statistical As- sociation, 91(435):1047–1061, 1996.

理論介紹

Isolation Tree 的建構方法

{姓名:小明 , 身高:120 , 體重: 70}

{姓名:小美 , 身高:160 , 體重: 40}

{姓名:小華 , 身高:175 , 體重: 60}

{姓名:小泰 , 身高:173 , 體重: 70}

{姓名:小世 , 身高:160 , 體重: 66}

{姓名:小國 , 身高:180 , 體重: 50}

理論介紹

Isolation Tree 的建構方法

{姓名:小世 , 身高:160 , 體重: 66}

{姓名:小明 , 身高:120 , 體重: 70}

{姓名:小美 , 身高:160 , 體重: 40}

身高> 170

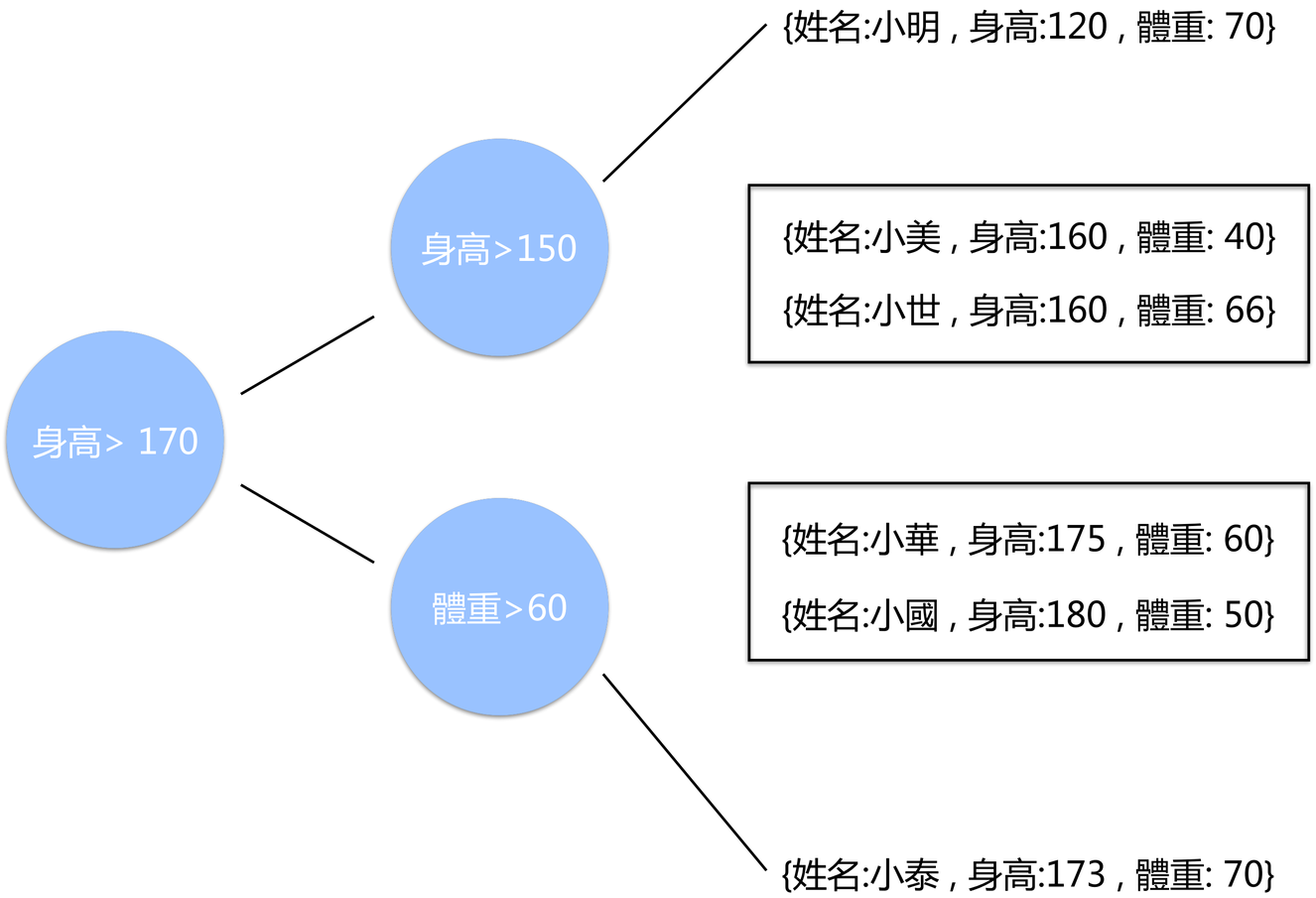
{姓名:小泰 , 身高:173 , 體重: 70}

{姓名:小華 , 身高:175 , 體重: 60}

{姓名:小國 , 身高:180 , 體重: 50}

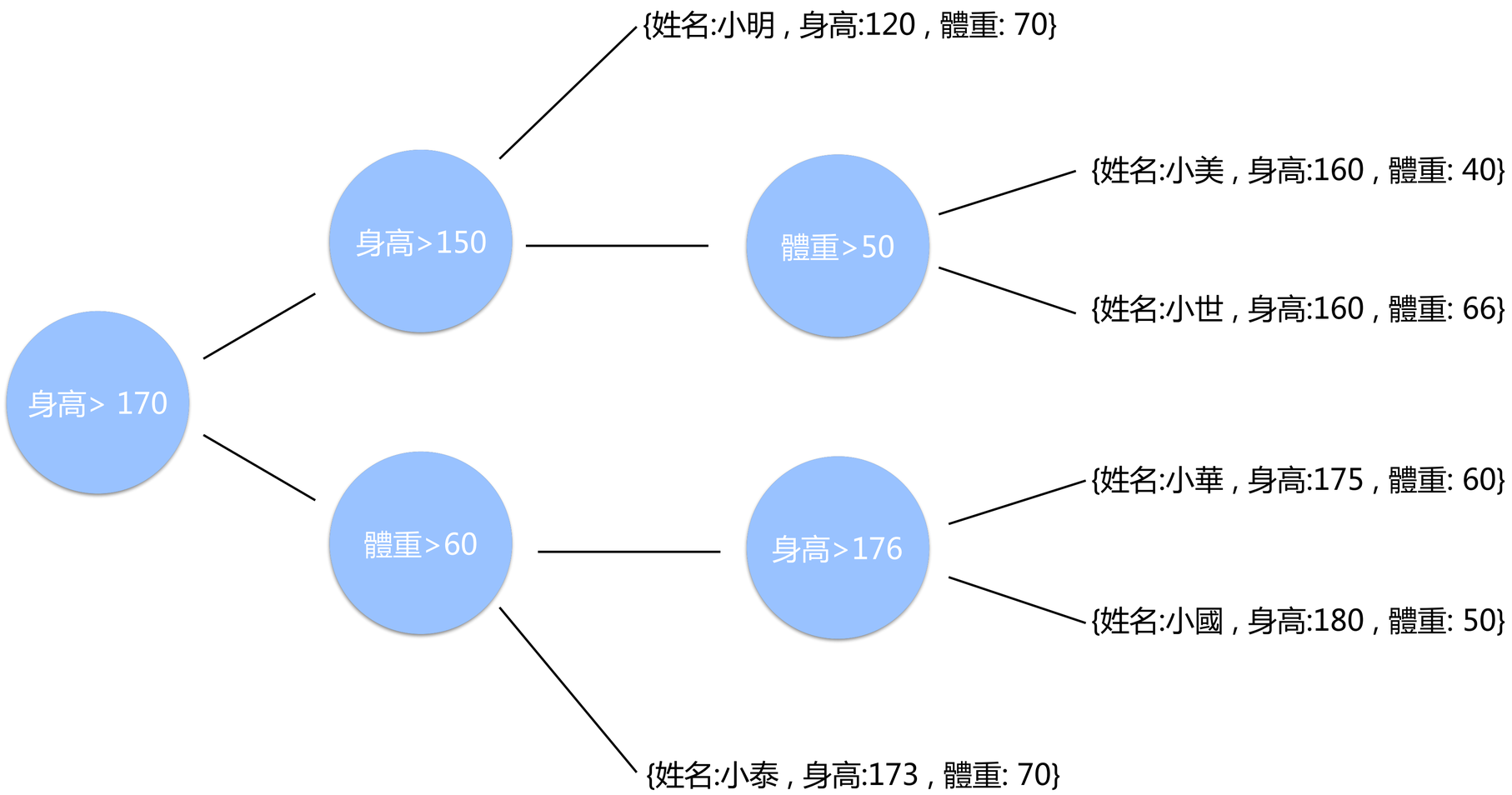
理論介紹

每次分割都抽取其中一個Feature, 並挑選候選節點中最大最小之間的值
直到每個節點都達到終止條件。



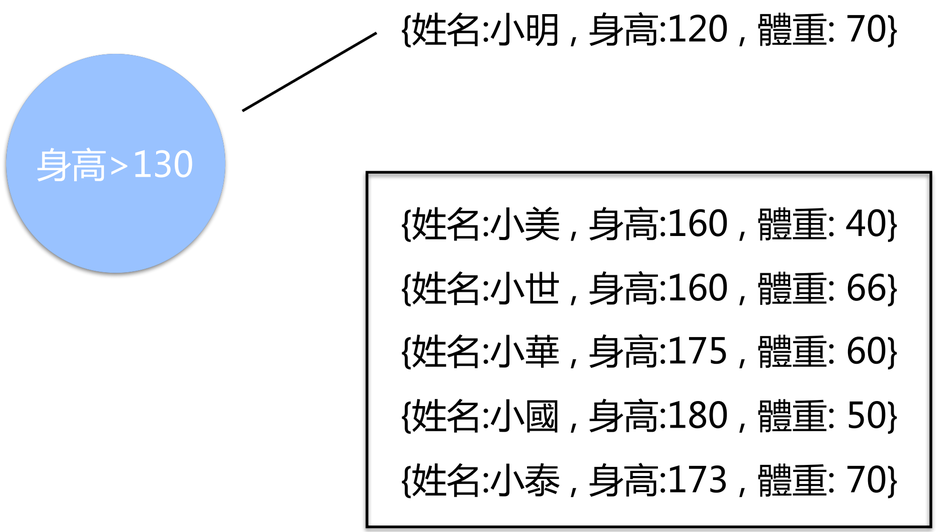
理論介紹

每次分割都抽取其中一個Feature, 並挑選候選節點中最大最小之間的值
直到每個節點都達到終止條件。



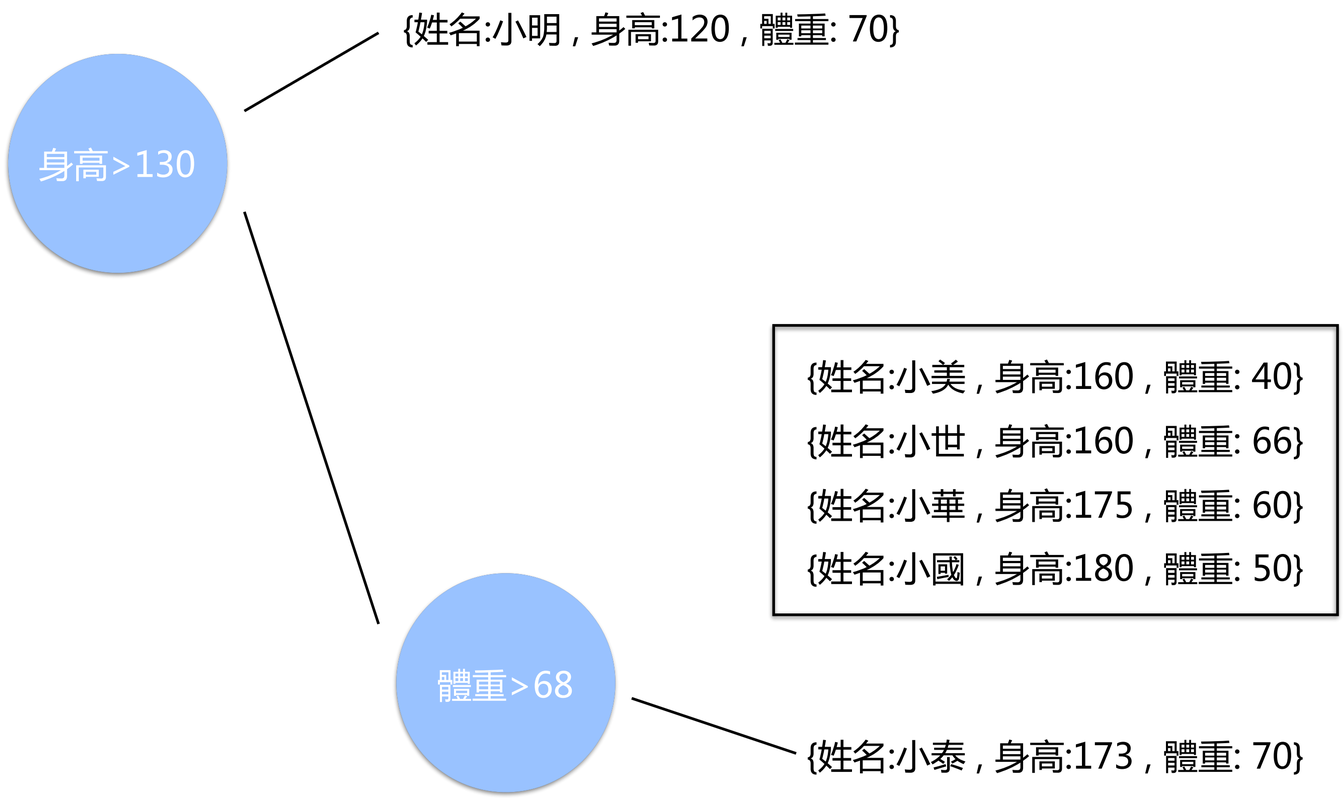
理論介紹

建立多個iTree ，組成iForest



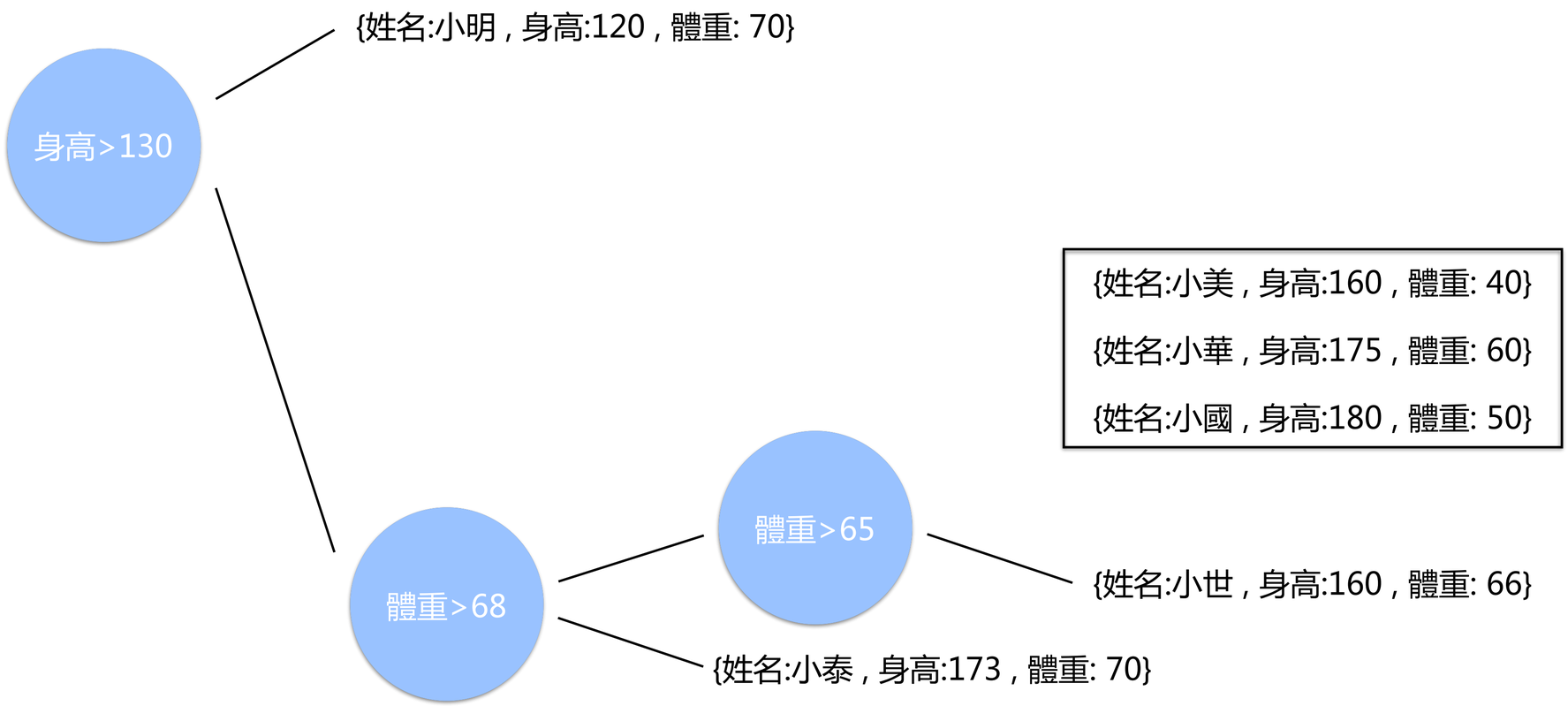
理論介紹

最高層數4層的例子



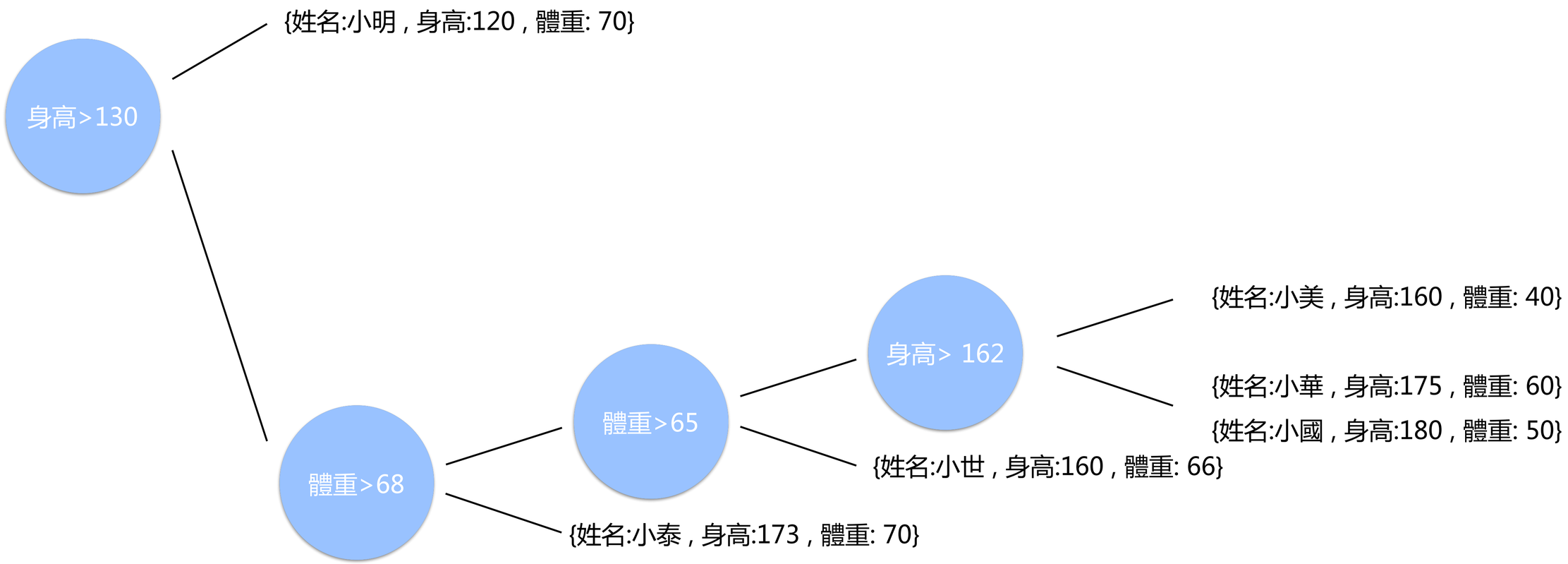
理論介紹

建立多個iTree ，組成iForest



理論介紹

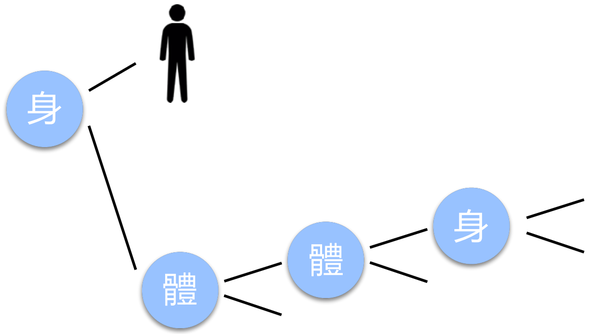
建立多個iTree ，組成iForest

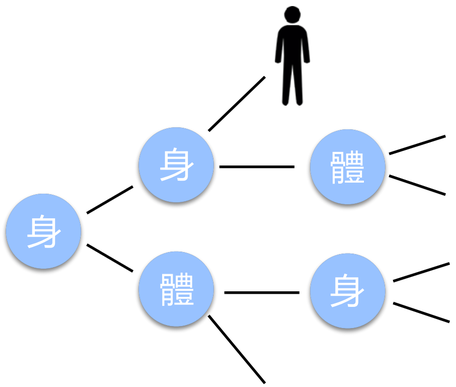


理論介紹

異常值的計算方式

{
姓名:小明 ,
身高:120 ,
體重: 70
}





- 小明的數據在第一棵樹被分到第二層
- 小明的數據在第二棵樹被分到第一層
- 平均層數為 1.5
- c(6) = 2H(6 – 1) – (2(6 – 1)/6) =2.72
- Anomaly Score = 2^-(1.5/2.72) = 0.682

Pros and Cons

Pros

- 線性時間、不需計算距離
- Scalable 可以平行運算
- 避免Swamping ,Masking

Cons

- 數值變數才能用
- Curse of Dimension
- 僅對Global Outlier敏感
  [Improving iForest with Relative Mass]
- 數據量過大反而會效果不好

Application

- Anomaly 代表的是某些值跟現有的資料集不一樣
  (大量的提領到底是潛在的商機還是潛在的犯罪？)
- Anomaly Score 要怎樣跟業務單位的目的找到交集？
  (經過業務單位的經驗訪談轉換爲具備解釋性及業務意義的變數)