資料前處理技術

國立宜蘭大學資訊工程系
吳政瑋 助理教授
wucw@niu.edu.tw

資料前置處理簡介

- **資料前置處理**是指在進行資料探勘之前，為了讓資料更適合進行探勘的工作，對於資料所做的預先處理動作。
- 在整個資料探勘的過程當中，**資料前置處理所需要花費的功夫通常是最多的**，同時資料前置處理也是對**探勘品質**影響最大的一個關鍵步驟。
- 資料前置處理的主要目的就是解決**資料品質不良的問題**，使得探勘結果的品質得以提升。

導致資料品質不佳的一些情況

- 未經處理的資料(Raw Data)可能存在許多品質不佳的情況，例如:
  - 資料不完整(Data Incomplete)
  - 資料有雜訊(Noise Data)
  - 資料不一致(Data Inconsistency)
  - 資料重複(Data Redundancy)

資料不完整(Data Incomplete)

- 資料不完整的情況最常見的便是資料中有某些屬性值有遺缺。
  - 例如：某顧客填寫會員資料表時，可能遺漏了填寫年齡這一欄。
- 在一般的資料庫系統中，除非管理者將資料庫中的每一個欄位均設定為不可接受空值(Null)，否則即有可能在某些欄位出現資料遺缺的情況。

資料有雜訊(Dara Noise)

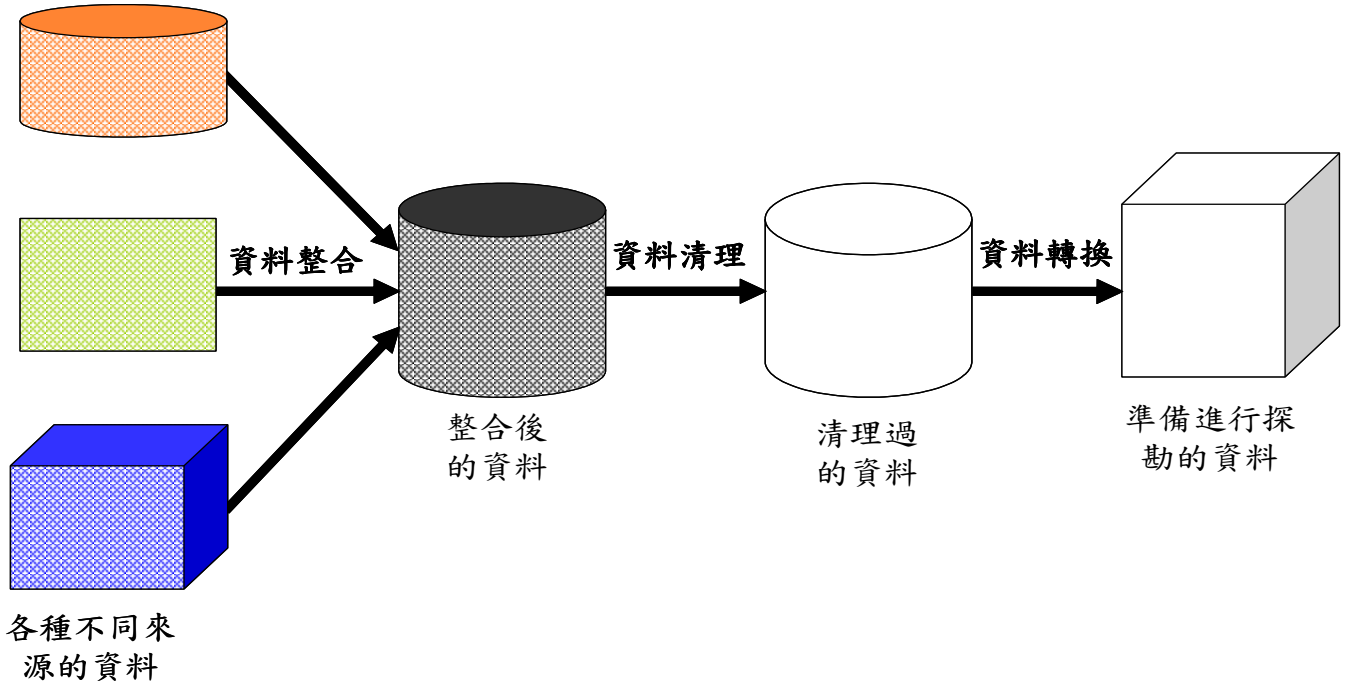
- 此問題多半是因資料有**錯誤**或是**特例(Outlier)**所造成的。
  - 例如：顧客填寫會員資料表時，有可能因為要保護自己的隱私而故意填寫錯誤的資料。
- 雜訊不一定全是故意填錯造成的，也有可能是因為填寫資料時不小心填錯或是資料原本就包含特例而產生的。
  - 例如：一般男生的身高大多介於165~185公分之間，然而有一位顧客的身高是197公分，這便是一個特例。
- 資料有雜訊不僅可能導致探勘的結果不正確，也有可能會誤導探勘的結果分析。

資料不一致(Data Inconsistency)

- 資料不一致的情況有許多，主要是因為資料由不同來源整合而得所產生。
- 例如：某一跨國性的企業，其商品在台灣是以台幣計價，而在美國則是以美金計價。當兩國的銷售資料被整合在一起做探勘時，若沒有經過適當的單位轉換，便會產生完全不正確的探勘結果。

資料前置處理的主要工作

- 資料前置處理主要包含**資料整合(Data Integration)**、**資料清理(Data Cleaning)**以及**資料轉換(Data Transformation)**等三項工作。



資料整合 (Data Integration)

- 資料整合是指將多重來源的資料整合在一個貯存庫中，因此資料整合最主要的目的便是解決多重資料來源的整合問題。

- 資料整合的主要工作有二
  - 消除資料不一致
  - 消除資料重複性

資料不一致的情況 (1/3)

- 數值不一致(Data Value Conflict)
- 綱目不一致(Schema Conflict)

資料不一致的情況 (2/3)

- **數值不一致(Data Value Conflict)**
  - 例如：商品價格在某個資料來源中用台幣計價，而在另一個資料來源中卻用美金計價。這種數值單位不一致的現象，透過單位換算，使數值的計算單位統一，即可消除。
  - 例如：同一位會員在A資料表中記錄的年齡是30歲，然而在B資料表中卻是25歲。因為無法判定究竟哪一個資料表是正確的，通常會採取的作法是將該屬性的資料刪除，以空值來取代，以消除內容不一致的情況。

資料不一致的情況 (3/3)

- **綱目不一致(Schema Conflict)**
  - 多半是屬性名稱不一致所造成的。
  - 例如：有的資料來源用「會員姓名」這個屬性名稱，而另一個資料來源卻用「顧客姓名」這個屬性名稱，雖然名稱並不相同，但實際所代表的意義卻是一樣的，可以透過**屬性更名**的動作來進行統一。

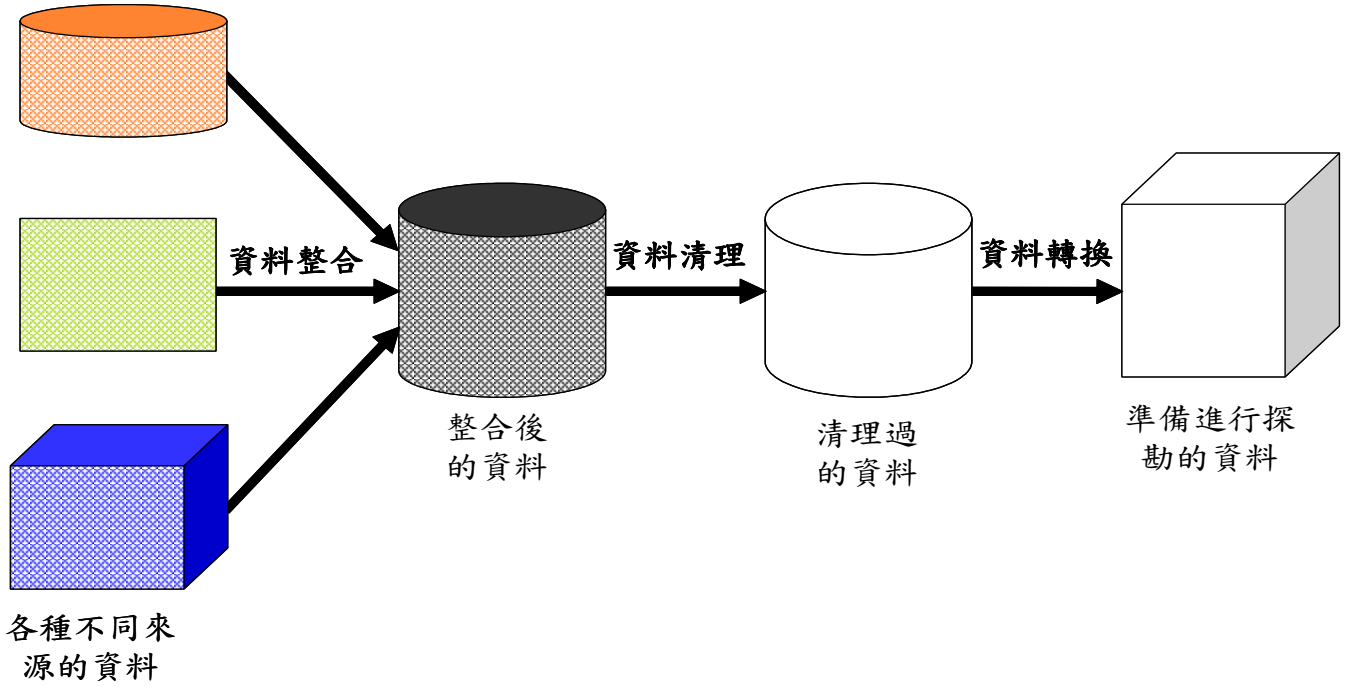
資料重複性的情況

- **數值重複**
  - 例如：整合中發現A資料表中有某會員的資料，在B資料表中也有同一位會員的資料，則可刪除其中一筆記錄，以免造成資料重複。

- **綱目重複**
  - 例如：資料經整合之後發現其中同時包含生日以及年齡這兩個屬性，因為年齡可以從生日推導出來，因此可以將年齡這個屬性刪掉以避免資料重複。

資料前置處理的主要工作

- 資料前置處理主要包含**資料整合(Data Integration)**、**資料清理(Data Cleaning)**以及**資料轉換(Data Transformation)**等三項工作。



資料清理 (Data Cleaning)

- 資料清理的主要目的是確認**資料的正確性**以及**完整性**，使得資料探勘能夠順利進行。

常見的資料正確性問題

檢查內容	說明
屬性的有效值或有效範圍	例如：性別屬性的值不是男性就是女性； 生日的月份應該介於1和12之間。
數值的唯一性	例如：身分證字號或是顧客編號不可有重複。
參考完整性	例如：存在於訂單資料表中的會員編號必須同時存在於會員資料表中。
資料的合理性驗證	例如：從會員的生日計算出該會員的年齡只有10歲，但是該會員所填寫的學歷卻是博士，顯然不合理。

常見的資料完整性問題

檢查內容	說明
是否缺少探勘所需的屬性	例如：當我們想要探勘顧客年齡與購買商品種類的關係時，卻發現資料庫中並未包含年齡這個屬性。
是否只包含統計整合過的資訊，而缺少詳細的單筆資料	例如：當我們想要分析某網站的瀏覽率以了解一天當中那一個時段最多人拜訪這個網站時，卻發現該網站每天只有記錄一筆當天的總瀏覽人次，而缺少每個小時的瀏覽人次資料。

其它相關的資料清理工作

- **遺缺填補**
  - 為了不讓屬性值有遺缺的資料影響探勘的結果，在進行資料探勘之前，應該設法把遺缺的資料填補進去。填補的方式又可分為**人工填補**或是**自動填補**。

- **雜訊消除**
  - 由於雜訊的存在有可能會使探勘的結果有相當大的偏差，因此必須將雜訊移除或是將資料做適當的**平緩化處理(Smoothing)**，以降低或是消除雜訊對於探勘結果的影響。

資料遺缺的原因

- **資料建立時未輸入**
  - 故意或不小心造成資料在建立時沒有被輸入
  - 例如：因為擔心個人資料曝光故意不填身分證號碼，或是因為疏忽漏填電話。若資料庫中的欄位未設定為不可接受空值，便有機會產生資料遺缺的情況。
- **設備故障**
  - 例如：因為收銀機故障導致顧客的消費明細無法即時輸入。
- **為了避免錯誤的資料影響分析**
  - 當資料內容不一致時，為了避免錯誤的資料影響分析結果，可能會將該項資料以空值取代，因此產生資料的遺缺。

資料遺缺的處理方法

- 直接忽略法
- 人工填補法
- 自動填補法

直接忽略法

- **直接忽略法**：直接忽略該筆內容有遺缺的資料。
  - 這種作法特別適用在進行分類探勘時。若是某筆資料的目標屬性為空值，那麼這筆資料因無法被正確分類，便可直接刪除。
  - 雖然直接忽略法相當容易，然而如果資料遺缺的比例很可觀時，此法會造成大量的資料流失，反而不利於探勘。
  - 直接忽略法較適用於所蒐集的資料量很多，但遺缺的資料只佔其中一小部分的情況。

人工填補法

- **人工填補法**：採用人工來填補遺缺的資料。
  - 為了處理資料遺缺的現象，可採用人工來填補遺缺的資料。
  - 例如：當某會員資料的生日屬性有遺缺時，可打電話詢問該會員以取得其生日加以填補。
  - 人力的負擔將會十分沉重。

自動填補法

- **自動填補法**：填入一個通用的常數值。
  - 在資料遺缺的處理方法上，較為**實際且可行**的作法便是**自動填補法**。
  - **目標屬性有遺失值**
    - 新增一類“未知”
  - **條件屬性有遺失值**
    - 數值型條件屬性：填平均值
    - 類別型條件屬性：填眾數

雜訊產生的原因

- 資料收集儀器暫時故障
- 資料輸入時的疏忽
- 資料本來就存在特例
- …

雜訊去除法

- 利用**雜訊偵測方法**將雜訊找出來並移除或處理
  - 又稱為**異常值偵測**
- 也可以利用**資料平滑化技術**將雜訊對於探勘結果的影響加以平緩。

雜訊偵測的方法

- 人工檢視
- 大於或小於平均值百分之二十以上
- 平均值加減三倍標準差
- 群集分析法
  - 以群集分析法先將資料做分群，分群之後的零散資料便可認定為雜訊。當雜訊所在的資料被辨識出來之後，便可將之移除。

資料平緩化處理的方法

- **裝箱法(Binning Method)**
  - 適用於**連續型屬性**
  - 主要可分為三個步驟
    - **(1) 排序**：將資料由小而大**排序**
    - **(2) 裝箱**：將資料分組裝入箱子中
    - **(3) 給值**：以各個箱子中所有資料的**平均值**、**中位數**或**邊界值**來取代箱子中的每一筆資料

- **常見的裝箱法**
  - **等寬分割法(或稱：等距分割法)**
  - **等深分割法(或稱：等頻分割法)**

常見的裝箱法

- **等寬分割法(或稱：等距分割法)**
  - 依照**資料的數值範圍**來劃分資料分組的區間

- **等深分割法(或稱：等頻分割法)**
  - 依照**資料的個數**來劃分資料分組的區間

等寬分割法 (1/4)

- **等寬分割法(或稱：等距分割法)**
  - (1) 將資料依數值範圍劃分為 ***N*** 個間隔相同的區間
  - (2) 假設 ***max*** 和 ***min*** 分別為該屬性中的最大與最小值，要將資料劃分成 N 個區間，則每個區間的寬度(Width)為：

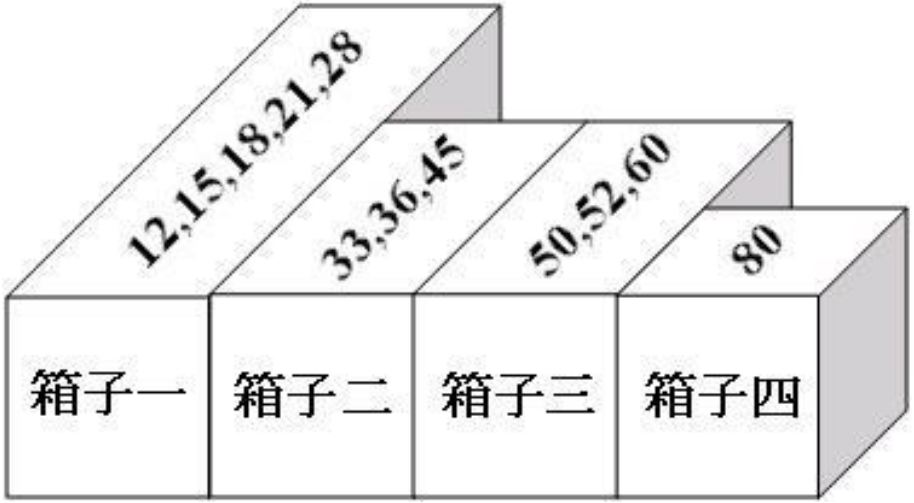
$$W = \frac{(max - min)}{N}$$

等寬分割法 (2/4)

- 假設12名顧客的年齡為：〈12, 15, 18, 21, 28, 33, 36, 45, 50, 52, 60, 80〉
- 若以**等寬分割法**進行平滑化，且將此12人的年齡分割成 $N = 4$ 個箱子來進行處理。
- 因為此年齡屬性的最大值 $max = 80$，最小值 $min = 12$，因此箱子的寬度為

$$W = \frac{(max - min)}{N} = \frac{(80 - 12)}{4} = 17$$

等寬分割法 (3/4)



- 箱子一所裝的是介在12和28之間的年齡資料，箱子二裝29~45，箱子三裝46~62，箱子四裝63~80。
- 因此箱子一會裝五個數字：12, 15, 18, 21, 28，箱子二會裝三個數字：33, 36, 45，箱子三會裝三個數字：50, 52, 60，第四個箱子只裝一個數字：80 。

等寬分割法 (4/4)

- 等寬分割法雖然頗符合直覺，然而當資料有雜訊或特例時，此種分割方法將對探勘結果有很大的影響
  - 以前圖為例，前面11個數字的分佈還算平均，然而第四個箱子卻只包含一個數字80，這一個數字很有可能是個特例。
  - 由於利用等寬分割法時，如果遇到特別高或是特別低的數值，用最大值和最小值作為區間範圍所計算出的箱子寬度，可能會造成裝入箱子裡的資料個數不平均的問題，因此等寬分割法並**不適合用在偏斜的資料(Skew Data)**上。

等深分割法 (1/3)

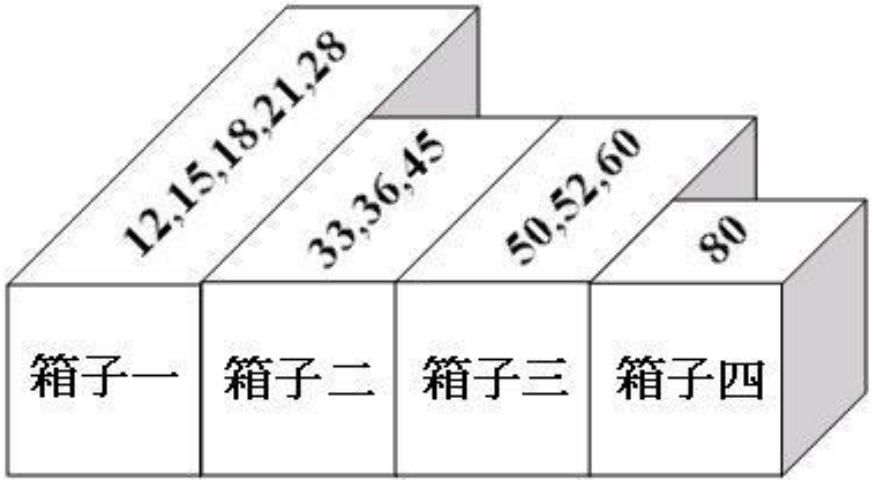
- **等深分割法(或稱：等頻分割法)**
  - 將每 ***K*** 筆資料裝入一個箱子中，每個箱子內的資料筆數相同
  - 以年齡屬性為例〈12, 15, 18, 21, 28, 33, 36, 45, 50, 52, 60, 80〉，等深分割法的結果為：

等深分割法 (2/3)

- 12個會員的年齡為⟨12, 15, 18, 21, 28, 33, 36, 45, 50, 52, 60, 80⟩，欲分割成 $N = 4$ 個箱子，12/4=3，因此每一個箱子放三個數字。箱子一放12, 15, 18，箱子二放21, 28, 33，箱子三放36, 45, 50，而箱子四放52, 60, 80。

等深分割法 (3/3)

- 等深分割法的資料分佈比等寬分割法好，箱子內含的資料量比較平均，即使資料有偏斜的情況也可以處理。



- 若採用等深分割法，每個箱子中資料的全距通常會不同

裝箱資料的平滑化方法

- 平均值法
- 邊界值法
- 中位數法

裝箱資料的平滑化方法:平均值法

- **平均值法：**以**平均值**取代個別的數字來消除雜訊。
  - 以等深分割法的例子來說，箱子一中裝著12, 15, 18, 此三數的平均值為15 (如果有小數點即四捨五入)，因此便以15, 15, 15取代12, 15, 18。箱子二裝著21, 28, 33，平均值為28，因此以28, 28, 28取代21, 28, 33。以此類推，箱子三裝的36, 45, 50以平均值取代成44, 44, 44；箱子四裝的52, 60, 80，以平均值取代成64, 64, 64。
  - 經過這樣的處理後可以發現，原本年齡80是一個特例，但被平緩化處理成64之後，其值便接近正常值了，由此可見平滑化的處理確實可消除雜訊。

裝箱資料的平滑化方法:邊界值法

- **邊界值法**：以邊界值取代個別的數字來消除雜訊。
  - 若是一數值離最小值較接近便用最小值取代，若離最大值較接近便用最大值取代。
  - 例如：12, 15, 18這個箱子中，15與最小值12和最大值18的距離相同，因此可任選12或18來取代15；假設選擇18, 則取代的結果為12, 18, 18。然而在21, 28, 33 這個箱子中，28距離33較21近，因此便用33取代28，成為21, 33, 33。
  - 用邊界值來消除雜訊時，雖然仍可看到雜訊的存在，但是雜訊的影響力已被降低。

裝箱資料的平滑化方法:中位數法

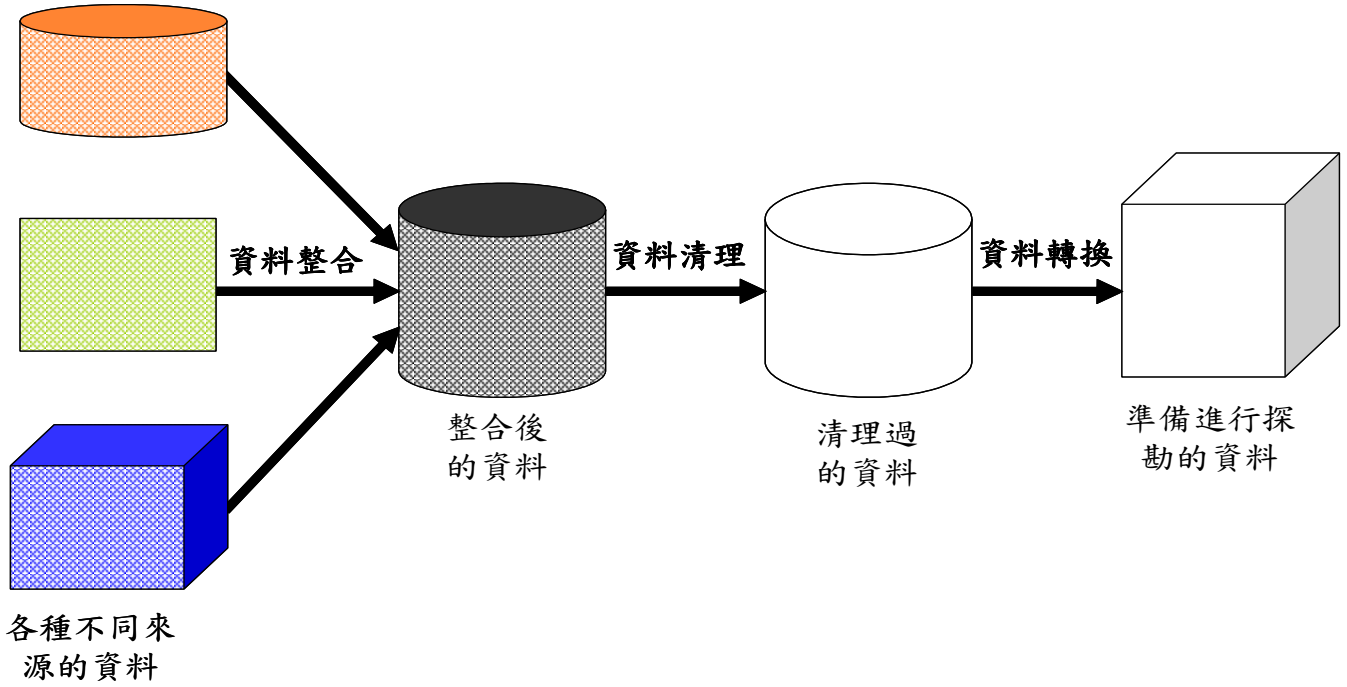
- **中位數法：**以**中位數**取代個別的數字來消除雜訊。
  - 以12, 15, 18這三個數字而言，中位數為15，因此就用15, 15, 15來取代12, 15, 18；21, 28, 33這三個數字的中位數是28，因此就用28, 28, 28取代21, 28, 33。依此類推，36, 45, 50取代成45, 45, 45；52, 60, 80則取代成60, 60, 60。

中位數平滑法的優點

- **中位數較平均值容易排除極端值的影響**
  - 在箱子四中放著52, 60, 80，其中80是一個相對較大的數字，但還不算太極端，如果80換成120，則52, 60, 120的平均值為77，對於其他數字而言，仍可視為特例
  - **平均值平滑法**無法完全避免極端的特例對平緩後的數值所產生的影響。
    - 若是採用**中位數平滑法**來處理，不管80被換成100、1,000、或是10,000，其中位數仍為60。

資料前置處理的主要工作

- 資料前置處理主要包含**資料整合(Data Integration)**、**資料清理(Data Cleaning)**以及**資料轉換(Data Transformation)**等三項工作。



資料轉換 (Data Transformation)

- 資料轉換的主要目的是將資料內容轉換成**更容易探勘**或是**探勘結果可信度更高**的狀態。
- 基礎的資料轉換工作包括：
  - 資料統整化(Data Aggregation)
  - 資料一般化(Data Generalization)
  - 建立新屬性(Attribute Construction)
- 進階的資料轉換工作包括：
  - 資料正規化(Data Normalization)
  - 資料形式轉換
  - 資料形態轉換
  - 資料模糊化(Fuzzy)

資料統整化(Data Aggregation)

- **資料統整化**是指將現有的資料做加總、統計或是建立資料方塊(Data Cube)。
  - 例如：將商品銷售資料按照銷售地區或是商品類別做加總。
- **資料統整化**的目的是將資料做初步的整理，使得資料更適合探勘的工作。
  - 例如：整合後的資料中有每一天的商品銷售紀錄，然而想要進行的探勘工作是找出銷售業績與氣候的關係。
  - 由於以每一天的數據來看，很可能因為資料變化太大而找不出規律性。
  - 可嘗試**將銷售資料按銷售月份做加總**，同時將**氣溫按照月份作平均**，用統整過後的資訊來進行探勘，有時會更容易探勘出隱藏在資料中的規律性。

資料一般化(Data Generalization)

- **資料一般化**是指將資料的概念階層(Concept Hierarchy)向上提升。
  - 例如：將會員的詳細地址用城市取代。
  - 例如：味全牛奶、林鳳營牛奶一般化成牛奶

- **資料一般化**可將某屬性中所包含的不同數值減少
  - 增加探勘結果的可用性
  - 降低探勘方法的計算成本

建立新屬性(Attribute Construction)

- 利用舊屬性將探勘所需的新屬性建立出來。
  - 例如：整合後的資料只包含會員的生日，然而探勘時要用的屬性是會員的年齡。
  - 由於年齡可以從生日推算而出，因此可在此步驟建立出所需要的年齡屬性。

資料正規化(Data Normalization)

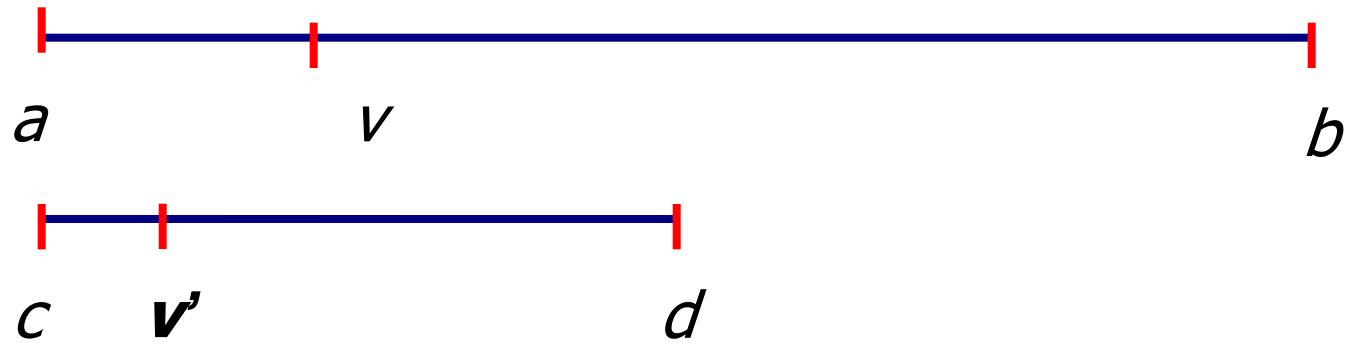
- **資料正規化**的主要目的是將不同標準之下所記錄的資料轉換到同一個標準，以便提高分析時的準確度。
- 資料的正規化會將資料重新分佈在一個較小而且特定的範圍內。
  - 例如：台灣人民的平均所得遠高於菲律賓人。一個月賺一萬八千元台幣在台灣算是中低收入。
  - 然而在菲律賓，這卻是相當於三個大學教授的薪水。
  - 若是直接拿兩國人民的收入數字來做比較，便會產生不夠客觀的問題。

極值正規化 (Cont. 1/2)

- 極值正規化的公式如下：

$$v' = c + (v - a)(d - c)/(b - a)$$

其中v為正規化前的數值，其範圍為$[a, b]$;
v'為正規化後的數值，其範圍為$[c, d]$。



極值正規化 (Exercise)

- 假設一般臺灣上班族的月收入範圍為[20000, 100000]，而一般菲律賓上班族的月收入範圍為[2000, 10000]；在台灣收入30000元相當於在菲律賓收入多少？

極值正規化 (Exercise)

- 假設一般臺灣上班族的月收入範圍為[20000, 100000]，而一般菲律賓上班族的月收入範圍為[2000, 10000]；在台灣收入30000元相當於在菲律賓收入多少？

  $v = 30000$

  $a = 20000$

  $b = 100000$

  $c = 2000$

  $d = 10000$

  將以上數字代入公式中即可求得v正規化後的數值為

$$v' = 2000 + (30000 - 20000)(10000 - 2000)/(100000 - 20000) = 3000$$

- 極值正規化適合用在需要將資料規範在某一個指定範圍內的情況。

Z-分數正規化(Z-Score Normalization)

- 公式

$$v' = \frac{\text{v} - \text{平均值}}{\text{標準差}}$$

- 範例：假設臺灣人月收入平均為35000元，標準差是10000元，利用Z-分數法將月收入30000元做正規化，將得到

$$\frac{30{,}000 - 35{,}000}{10{,}000} = -0.5$$

  - 負數表示收入低於平均，正數表示高於平均；結果之絕對值越小，表示偏離平均值程度越小，反之越高
  - Z-分數正規化適合用在需要了解數值與平均分佈之間的關係時

十進位正規化

- 十進位正規化之公式如下：

  $v' = v/10^i$，其中 i 是使得 Max($|v'|$)≤1的最小整數。

  - 假設台灣上班族**最高月收入為100,000元**，因此使得正規化後的結果小於或等於1的最小整數 i 為5。則月收入30000元經由十進位正規化之後將會得到：

$$v' = 30000/10^5 = 0.3$$

  - 十進位正規化適合用在要將數字壓縮到區間[0,1]的情況。

總結

- 資料前置處理的目的
  - 提高資料探勘的品質
  - 有高品質的資料，才有高品質的探勘結果
- 介紹資料前置處理的基本概念
  - 資料清理
  - 資料整合
  - 資料轉換