

DATAMINE 軟件時，又需要輸入數據才能啓動工序，而所得到的結果又沒有甚麼導致這個原因多數是用家不明白每個工序和參數的用途。其實輸入參數有助引導工在有限的時間裏尋找有用的資料。自動化的 DATAMINE 在於你能否了解工序，配合發揮 DATAMINE 的能力，從而得到真正的資料。



DATAMINE 至今讀者看到一些字眼，例如分類、拼法條等。它們都是在 DATAMINE 技術中出來的，可以技術的應用是非常普遍，一半 DATAMINE 軟件開發公司，技術加入其系統中，這技術就樹 (DECISION TREE)。由應用廣泛，所以清楚掌握決策作、優點和缺點，不但可以為擇合用的軟件，而且在應用這時候，便可避重就輕，大大發 DATAMINE 的功力。

請決策樹，顧名思義就是將整的流程，把其中可能牽涉到的件，有系統地以樹狀的結構呈。決策樹是由樹根、樹節和樹葉成 (見圖一)。樹根是分析的樹葉是終點、樹節是分差路 (決策點)。

馬作例設定決策樹

現在用一個數據庫共三百三十二條，資料取材於香港賽馬記錄，用其中一場參賽馬 (共十四個過往賽蹟 (見表一)。現在來決策樹的運作。首先決策樹演式放在起點的樹根，當用家確定分析的特徵後，演式把其餘的特

一輪的分裂中被選出。小組還有能力的特徵。環，直到樹葉出現，樹狀的決策樹就會形成。

數值種類可分為文字和數字。文字種類比數字種類簡單，特徵值是一個個數，按先後次序之分，便可以分。文字種類要分為「不連續型」和「連續型」，馬匹的重量是連續型的數字種類，但馬匹出賽編號是不連續型的數字種類。而連續型的特徵值需要找一個分割點，統計學的標準差可以作為分割基礎。分割後的每一小組，所得的標準差是最小。決定數字種類是非常重要的，若果把馬匹排位特徵作為分析對象，連續型的數字所得到的答案是：少於五數字為一組，大過和等於五數字為另一組。而不連續型的數字所得到的答案是：一、二、四、十一、十三、十四為一組，三、五、六、九為另一組，七、八、十為最後一組。可見分析時，筆者覺得連續型的數字種類比較適合馬匹排位特徵。

如何選出分裂目標

完成選擇數據種類，可以開始進行分裂計算。以「馬匹出賽號碼」作為「公式」上一般常用的公

Introduction to Maxeler Computing



Veljko Milutinovic, vm@etf.rs

<http://home.etf.rs/~vm>

有選，所以在分裂時，「不連續型」特徵是不合用的話，請刪掉它，以把引導向錯誤方向分析。除了「公式」外，就是樹的演式是用二非分裂，是用多枝節分裂，多能力的特徵完全發。每一層面把所有特徵值立的組別。例如十四個組別，但採用演式，每一次分裂只有兩枝節，在例中第一層面有兩組馬匹名稱。二進分裂後的特徵還有能力進行。在三個層面進行分裂。

當分裂開始了，怎樣把它停止下來？方法是有很多，比較簡單的方法是用分裂後的數據數量。當分出來的組別越多，每一個組別的數據數量越少。如果在分析前可訂下一個「

在「公式」外，就是樹的演式是用二非分裂，是用多枝節分裂，多能力的特徵完全發。每一層面把所有特徵值立的組別。例如十四個組別，但採用演式，每一次分裂只有兩枝節，在例中第一層面有兩組馬匹名稱。二進分裂後的特徵還有能力進行。在三個層面進行分裂。

這種分析會不會「走偏」？而看樹葉編號，4、5、6、7、8、9、10、11、12、13、14 不被採用是由於這些個特徵的「起碼」的標準，但仔細看看馬匹編號 4 和 5 的馬匹，這時的馬匹有十個檔案，再加上有資料比分析力的

SuperComputer Types (Pros & Cons)



- Control flow:
Compiling down to MCL (Machine Code Level)
- Data flow:
Compiling down to GTL (Gate Transfer Level)
- What we get?
Better speed/watt/size/\$
- How we pay for these benefits?
More difficult programming
- For what applications is this good?
Write once, execute many: Geophysics, Banking, Google, ...

DATAMINING 至今讀者
看到一些字眼，例如分類、拼
法條等。它們都是在
MINING 技術中出來的，可以
技術的應用是非常普遍，一半
DATAMINE 軟件開發公司，
技術加入其系統中，這技術就
樹 (DECISION TREE)。由
應用廣泛，所以清楚掌握決策
作、優點和缺點，不但可以為
擇合用的軟件，而且在應用這
時候，便可避重就輕，大大發
AMINING 的功力。

請決策樹，顧名思義就是將整
的流程，把其中可能牽涉到的
事件，有系統地以樹狀的結構呈
。決策樹是由樹根、樹節和樹
成 (見圖一)。樹根是分析的
樹葉是終點、樹節是分差路
 (決策點)。

馬作例設定決策樹
現在用一個數據庫共三百三十二
，資料取材於香港賽馬記錄。
，用其中一場參賽馬 (共十四
) 過往賽績 (見表一)。現在來
決策樹的運作。首先決策樹演式
放在起點的樹根，當用家確定
分析的特徵後，演式把其餘的特
分析，選出一個具影響力的特

An Example of WORM: Oil Drilling

一輪的分裂中被選用。當分裂後的數
組還有能力的特徵時，分裂會不斷重
環，直到樹葉出現，樹狀的決策樹就
會形成。

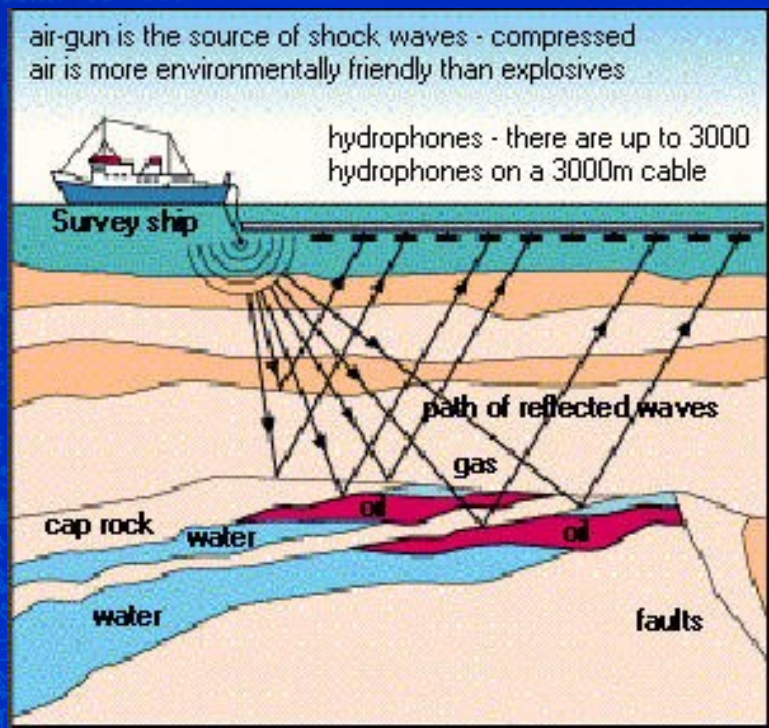
數值
字種類比
個個體，
以分開或
的名稱就
要分為「
馬匹的重
馬匹出賽
頻。而運
割點，就
基礎。分
準差是最
要，若果
對象，滿
是：少於
五數字為
所得到的
一、十二
六、九為
一組。

的是分析結果的可用性，筆者覺得連
續型的數字種類比較適合馬匹排位特
徵。

如何選出分裂目標

完成選擇數據種類，可以開始進
行分裂。利用「二元化」演式 (公式
和算式) 或「二元化」演式的公

GATE) 有
相類的數
資料提升能力的功



在三個層面進行分裂。

分裂開始後如何停下來?

當分裂開始了，怎樣把它停止下
來？方法是有很多，比較簡單的方法
是用分裂後的數據數量。當分出來的
組別越多，每一個組別的數據數量越
少，如果在分析前可訂下一個「

的資料，在
，只選出最
出。雖然這
要能從出產
次在各小區
果累積到的
計算已能出
這種分析會
能選出最
們不被採用
比低。原因
個「起碼」
編號 4 和 5
十個檔案，
被用，則

DATAMINE 軟件時，又需要輸入數據，才能啓動工序，而所得到的結果又沒有甚麼導致這個原因多數是用家不明白每個工序和參數的設定，其實數據分析是有限度的。数据挖掘 (DATA MINING) 在於你能否了解工序，配合發揮 DATAMINE 的能力，從而得到真正的實效。



An Example of WORM: Credit Transactions

DATAMINE 至今讀者看到一些字眼，例如分類、拼法條等。它們都是在 MINING 技術中出來的，可以技術的應用是非常普遍，一半 DATAMINE 軟件開發公司，技術加入其系統中，這技術就樹 (DECISION TREE)。由應用廣泛，所以清楚掌握決策作、優點和缺點，不但可以為擇合用的軟件，而且在應用這時候，便可避重就輕，大大發揮 MINING 的功力。

所謂決策樹，顧名思義就是將整的流程，把其中可能牽涉到的條件，有系統地以樹狀的結構呈。決策樹是由樹根、樹節和樹葉成 (見圖一)。樹根是分析的樹葉是終點、樹節是分差路 (決策點)。

馬作例設定決策樹

現在用一個數據庫共三百三十二條，資料取材於香港賽馬記錄，用其中一場參賽馬 (共十四的過往賽蹟 (見表一)。現在來決策樹的運作。首先決策樹演式放在起點的樹根，當用家確定分析的特徵後，演式把其餘的特

一輪的分裂中被選用。當分裂後的小組還有能力的特徵時，分裂會不斷循環，直到樹葉出現，樹狀的決策樹就會形成。

數值種類字種類比數字個個體，沒有以分開或組合的名稱就是文要分為「不連馬匹的重疊是馬匹出賽編號。而連續型割點，統計學基礎。分割後準差是最小。要，若果把其對像，連續型是：少於五數五數字為另一所得到的答案一、十三、十六、九為另一組。可見分的是分析結果續型的數字種徵。

如何選出分裂目標

完成選擇數據種類，可以開始進行分裂計算，以下地分裂計算程式稱稱為「選擇公式」，一般採用的公

GAIN)。資料變異是計算變異的相別的數值，然後求變異的資料提升能力的特徵，以決定最佳



當分裂開始了，怎樣把它停止下來？方法是有很多，比較簡單的方法是用分裂後的數據數量。當分出來的組別越多，每一個組別的數據數量越少。如果在分析前可訂下一個「起

點」，當變異的資料提升能力的特徵，以決定最佳。這種分析會不會有走質？

而看樹葉編號3、4、5的馬匹們不被採用是由於牠全個檔案的變異比低。原因是賽者在分前訂下了一個「起碼」的檔案，但仔細查看馬匹編號4和5的檔案，這時便大約有十個檔案，再加上有資料提升能力的特徵，有變異的資料提升能力的特徵，以決定最佳。

這種分析會不會有走質？

而看樹葉編號3、4、5的馬匹們不被採用是由於牠全個檔案的變異比低。原因是賽者在分前訂下了一個「起碼」的檔案，但仔細查看馬匹編號4和5的檔案，這時便大約有十個檔案，再加上有資料提升能力的特徵，有變異的資料提升能力的特徵，以決定最佳。

An Example of WORM: Social Mining



一輪的分裂中被選用。當分裂後的小組還有能力的特徵時，分裂會不斷循環，直到樹葉出現，樹狀的決策樹就會形成。

數值種類字種類比數字個個體，沒有以分開或組合的名稱就是又要分爲「不適馬匹的重量馬匹出賽頻類。而連續型分割點，統計學基礎。分割標準差是最小，若果把對像，選擇是：少於五五數字爲另一所得到的答一、十三、六、九爲另一組。可見分析結果完全不同，還有的是分析結果的可用性，筆者覺得連續型的數字種類比較適合馬匹排位特徵。

如何選出分裂目標

完成選擇數據種類，可以開始進行分裂計畫。以下地分裂計畫方程式簡稱爲「選擇公式」，一般採用的公

GAIN)。資料型別是計量型別，相別的數值。然後，數值型別資料提升能力的特徵，其特徵的價值是最大，當其比過任何特徵時。



下次分裂。（見圖一）馬匹名稱總共在三個層面進行分裂。

分裂開始要如何停下來？

當分裂開始了，怎樣把它停止下來？方法是有很多，比較簡單的方法是用分裂後的數據數量。當分出來的組別越多，每一個組別的數據數量越少。如果在分析前可訂下一個「最

佳」的數值。當其比過任何特徵時，其價值是最大，當其比過任何特徵時。

這種分析會不會有走偏？筆者想說，在分析前訂下「最佳」的數值，當其比過任何特徵時，其價值是最大，當其比過任何特徵時。這種分析會不會有走偏？筆者想說，在分析前訂下「最佳」的數值，當其比過任何特徵時，其價值是最大，當其比過任何特徵時。

筆者想說，在分析前訂下「最佳」的數值，當其比過任何特徵時，其價值是最大，當其比過任何特徵時。這種分析會不會有走偏？筆者想說，在分析前訂下「最佳」的數值，當其比過任何特徵時，其價值是最大，當其比過任何特徵時。

History



- Serbia
- Vienna
- Technion
- Stanford
- AT&T
- Imperial College London
- Maxeler
- J.P. Morgan
- Schlumberger, Exxon, British Petrol, ENI, ...
- China, Japan, ...



Future



- Fujii Labs
- China Exascale
- NASA
- NIST
- Centers of Excellence
- Programs for Universities

一輪的分裂中被選用。當分裂後的小組還有能力的特徵時，分裂會不斷循環，直到樹葉出現，樹狀的決策樹就會形成。

數值種類可分為文字和數字。文字種類簡單，特徵值是一組，而數字種類則有先後次序之分，隨便可以分開或組合。在本文例子中，馬匹的體重是不連續型的數字種類，而馬匹的跑速是連續型的數字種類，但馬匹的跑速是不連續型的數字種類。而連續型的特徵值需要找一個分割點，統計學的標準差可以作為分割基礎。分割後的每一小組，所得的標準差是最小。決定數字種類是非常重要的，否則馬匹排位的分析結果是：少於五數字為一組，大過則等於五數字為一組。而連續型的數字種類，一、十三、十四為一組，三、五、六、九為另一組，七、八、十為最後一組。可見分析結果完全不同，還有的是分析結果的可用性，筆者覺得連續型的數字種類比較適合馬匹排位特徵。

如何選出分裂目標

完成選擇數據種類，可以開始進行分裂計算。用「選擇公式」或「選擇公式」。

除了「選擇公式」外，還有另一種的演式是用多枝樹分裂會把有能力的特徵，就是說在同一層面把分出來，成為獨立的組別。馬匹的名稱就有十四個，每一枝樹，在例中第一層面有稱。二進分裂後的特徵還下次分裂，（見圖一）馬匹在三個層面進行分裂。

當分裂開始了，怎樣來？方法是有很多，比較是用分裂後的數據數量，組別越多，每一個組別的數據數量就少。如果在分析前訂下一個「

分裂開始後如何停下來

當分裂開始了，怎樣來？方法是有很多，比較是用分裂後的數據數量，組別越多，每一個組別的數據數量就少。如果在分析前訂下一個「

