総啓動工序・面所得到的結果又並考慮 動立個原因多數是用家不明自每個工序 和参數的用途。其實輸入參數有助引導工 有限的時間裏尋找有用的資料。自動化的 INING A 於你能否了解工序・配合量

> 一個的分裂中做過 超過有能力的特別 重,直到樹葉出現 會形成。

的應用だ非常資源。一十 「AMENE 軟件開發公司, 加入其系統中,這技術就 「DECISION TREE )。由 實證,所以清楚掌握決策 優點和缺點,不但可以為 用的軟件,而且在應用這 ,便可避電就輕,大大號

AMUNITING 的历历了 請決策樹, 原名思發就是將整 的流程,把其中可能牽涉到的 件,有系統地以期狀的結構呈 。決策樹是由樹根、樹苗和樹 成(見壓一)。樹根是分析的 樹葉是終點、樹苗是分差路 決策點)。

事作例設定決策被

在用一個數據庫共三百三十二 ,資料取材於香港裏馬記錄, 用其中一場參賽馬(共十四 )過住賽蹟(見表一)。現在來 )過位賽蹟(見表一)。現在來 發樹的運作。首先決策園面式 繁樹的運作。首先決策園面式 對在起點的樹根,當用家庭的 術的特徵後,讓式把其餘的特 Data Mining Versus Semantic Web

電話的重量是速模型的動字碰頻,但 馬匹出賽輛裝是不運積型的動字。 與。而還積型的特徵僅需要找一個分 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、所得的調 要是是非常人匹從位特徵包含 對像,連環型的數字所得到的容易 是:少於五數字路一組、大遍和等的 五數字路另一組。而不更模型的數字 所得到的容案是:一,二,因



Veljko Milutinovic, <u>vm@etf.bg.ac.yu</u>

http://galeb.etf.bg.ac.yu/vm

This material was developed with financial help of the WUSA fund of Austria.

Page 1/65

### DataMining versus SemanticWeb

- Two different avenues leading to the same goal!
  - The goal:
    Efficient retrieval of knowledge,
    from large compact or distributed databases,
    or the Internet
- What is the knowledge: Synergistic interaction of information (data) and its relationships (correlations).
- The major difference: Placement of complexity!

Page 2/65

## Essence of DataMining

- Data and knowledge represented with simple mechanisms (typically, HTML) and without metadata (data about data).
- Consequently, relatively complex algorithms have to be used (complexity migrated into the retrieval request time).
- In return, low complexity at system design time!

Page 3/65

## Essence of SemanticWeb

Data and knowledge represented • with complex mechanisms (typically XML) and with plenty of metadata (a byte of data may be accompanied with a megabyte of metadata). Consequently, relatively simple algorithms can be used (low complexity at the retrieval request time). However, large metadata design and maintenance complexity at system design time.

Page 4/65

## Major Knowledge Retrieval Algorithms (for DataMining)

- Neural Networks
- Decision Trees
- Rule Induction

etc...

• Memory Based Reasoning,

### Consequently, the stress is on algorithms!

Page 5/65

## Major Metadata Handling Tools (for SemanticWeb)

- XML
- RDF
- Ontology Languages
- Verification (Logic +Trust) Efforts in Progress

Consequently, the stress is on tools!

Page 6/65

## **Issues in Data Mining** Infrastructure

### Authors:



Veljko Milutinovic, <u>vm@etf.bg.ac.yu</u>

http://galeb.etf.bg.ac.yu/vm Page 7/65

影響動工序。而所得到的結果又做有量度 素致這個原因多數是用素不明自每個工序 相參數的用途。其實輸入參數有助引導工 有限的時間裏尋找有用的資料。自動化的 INING 分於你能否了解工序。配合學 E DATAN ING 的能力。從而得到高正

# Semantic Web

Ivana Vujovic (ile@eunet.yu) Erich Neuhold (neuhold@ipsi.fhg.de) Peter Fankhauser (fankhaus@ipsi.fhg.de) Claudia Niederée (niederee@ipsi.fhg.de) Veljko Milutinovic (vm@etf.bg.ac.yu)

#### 。作例股定决策视

加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 )過住賽蹟(見表一)。現在來 除樹的運作。首先決策附廣式 除樹的運作。首先決策附廣式 除樹的特徵後, 預式把其餘的等 一份好, 戰地一個且影響力的等

### http://galeb.etf.bg.ac.yu/vm

的是分析結果的可用性,等者量得) 被擊的數字種類比較適合馬匹排位。 数。

#### 如何是出分裂目標

李成演播數據電源 · 可以開始過

#### Page 8/65

分规局由复加同样下非

當分裂開始了,怎樣把它停止 來?方法是有很多,比較簡單的方為 是用分裂機的數據數量,當分出來自

## Data Mining in the Nutshell

超過有能力的特徵時。分裂會不斷值 相 環、直到樹葉出現、相就的決策固創 質

### Uncovering the hidden knowledge

## Huge n-p complete search space

### **Multidimensional interface**

對像,連模型的數字所得到的答案 是:少於五數字為一組,大通和符防 五數字為另一組,而不逆模型的數字 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、十為最待 一組,可見分析結果完全不同,還等 的是分析結果的可用性,筆者覺得的 機學的數字種類比較適合馬匹排位者 考。

#### 如何是出分裂目標

完成選擇數據範圍,可以開始譜

Page 9/65

定起來,或為來立的証则,而 匹罵的名稱就有十四倍起则。 日二進分裂的族式,每一次分裂, 技節,在刻中強一層面有簡結。 第。二進分裂後的傳還還有能。 下次分裂,(更屬一)馬匹名。 在三個層面就行分裂。

#### 分裂開始長如同停下市

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 出別越多,這一個關係的數據國際是



### You are a marketing manager for a cellular phone company

- Problem: Churn is too high
- **Turnover (after contract expires) is 40%**
- Customers receive free phone (cost 125\$) with contract
  - You pay a sales commission of 250\$ per contract
- Giving a new telephone to everyone whose contract is expiring is very expensive (as well as wasteful)
- Bringing back a customer after quitting is both difficult and expensive

Page 10/65



- Three months before a contract expires, predict which customers will leave
- If you want to keep a customer that is predicted to churn, offer them a new phone



- The ones that are not predicted to churn need no attention
- If you don't want to keep the customer, do nothing
- How can you predict future behavior?
- 例設定決策機 用一個數據庫共三百三十二 着料取材於香港實所記錄 其中一場參賽馬(共十四 生實蹟(見表一)。現在 定 超的運作。首先決策樹廣式 生起點的樹根,當用家確定 的特徵後,摘式把其餘的特 好, 業份一個具影響力的幹
- Tarot Cards? Magic Ball? Data Mining?
  - Page 11/65





The newest Computer Model, 16 Colours, on Harddisk, including Mouse......



11月夏山27年1日年 今山湾爆動線新研(13)

Page 12/65

You should wait, Your Majesty - In 6 months it costs only half.....

來拿方法是有很多,比較簡單的方面 是用分裂後的數據數量。當分出來自 這別想多,這一個說外的欺個計量。

hartrit - an file da



## The Definition

### The automated extraction of predictive information from (large) databases

AMENENG 的功力。 》 請決策樹, 顧名思義就是將 的流程, 把其中可能牽涉到的 件, 有系統地以期狀的單構呈 。決策樹是由樹根、樹錦和雪 成(見壓一)。根根是分析的 樹葉是発點、樹齡是分差路 決策點)。

,作例設定決策被

在用一個數據庫共三百三十二 ,資料取材於香港實所記錄 用其中一場套要馬(共十四 過估要蹟(見表一)。現在來 激樹的運作。首先決策樹廣式 激在起點的樹根,當用家吃定 析的特徵後,積式把其餘的等 Automated Extraction

Predictive

### Databases

實現的數字種類比較適合馬匹排位約 数。

#### 如何是出分裂目的

Page 13/65

#### 分数据由最如何停下局部

當分裂開始了,忽然把它停止了 來?方法是有俱多,比較简單的方位 是用分裂後的數律軟量。當分出來自

型名胡葉編紀2 - 4 - 5 10 費希該採用是由于估合何信念(1653) 助任。國因是來看在於1000000 個「起既」的編載, 回F書面看以 個時, 新人物曲篇, 如時完美(1653)

## History of Data Mining

TO I DOWNER IN A MANAGENERU, Martin may proce only ailes ago to montopo Kirower The main and the parties and an an gue ver and in optiming ky artistal an tity con - antion End a main and a for an and the fer The forse in and sincher in (OUXIO) Jan of ailight at any and said a the inter the property RAMPORT IPWATE MEDINA TO OPIN TOMPTINE AND ייטור לעינו גיווטרעי ביידי איראי איין איים איין איים איין איים michinal mulagric stadenic river Arcula miniar iponnylar ministiget to ismarticit Don was ippid finitar in spingto for ino sino 17 sight and successed min Highing HE MWILE ( MACC E)Lat Lowith Kay and a sould the first Non Von helogn Kay Dim pardialic Kais Toe



מכ

חעולס א מכנו כו אהעוכ

ישטוןא הכהמי כטבחרה עריקרשך כ

אתהיאאיטוס

Construction of the second of

81 M . . XF

为特徵後, 演式把其餘的特 所, 寬地一個具影響力的转

Page 14/65

## **Repetition in Solar Activity**

### 1613 – Galileo Galilei

### 1859 – Heinrich Schwabe

過住要讓(見表一)。為 象樹的運作。首先決策崩損入 放在起點的樹根,當用家脏定 折的特徵後,微式把其餘的特 分析,需找一個且影響力的特

Page 15/6<u>5</u>





## The Return of the Halley Comet

### Edmund Halley (1656 - 1742)

1531 1607

1682

### 239 BC

### 1986

2061 ???

### 1910

Page 16/65

## Data Mining is Not

- Data warehousing
- Ad-hoc query/reporting
- **Online Analytical Processing (OLAP)**
- **Data visualization**



Page 17/65



## Data Mining is

Automated extraction of predictive information from various data sources

Powerful technology with great potential to help users focus on the most important information stored in data warehouses or streamed through communication lines







的背面後一個具影響力的時



## Focus of this Presentation

一輪的分裂中被濡用。當分裂後的 超還有能力的結紮肉,分裂會不斷 重,直到創葉出現,如此由決策問 實務成。

### **Data Mining problem types**

### **Data Mining models and algorithms**

### **Efficient Data Mining**

### Available software

的是分析結果的可用性,筆者量得) 實現的數字種類比較適合馬匹排位 数。

#### 如何是出分裂目標

完成演漫數據範圍 (可以開始)

#### Page 19/65

分规院始长如何停下部

當分裂開始了,怎樣把它停止 來?方法是有很多,比較簡單的方。 是用分裂後的數據數量,當分出來的

DATAMINE 軟件時,又需要輸入電多等 能移動工序,而所得到的結果又沒有基层 等效這個原因多数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 有限的時間裏尋找有用的資料。自動化的 MINING 是在於你能否了解工序,配合參 MINING 是在於你能否了解工序,配合參

## Data Mining Problem Types

举合用的软件,而且在遮闭道 時候,便可避重就輕,大大發 AMENING 的功力。 請決策樹, 服名思發就是將整 的流程,把其中可能牽涉到的 件,有系統地以趨肤的培構型 。決策樹是由樹根、樹節和樹 成(見圖一)。樹根是分析的 樹葉是蒸點,她能是分差路 等決策點)。

#### 馬作例設定決策被

1在用一個數據庫共三百三十二 ,資料取材於香港賽馬記錄, 一用其中一場參賽馬(共十四 )過住賽寶(見表一)。現在來 除樹的運作。首先決策樹廣式 就在在起點的樹根,當用家庭定 )析的特徵後,領式把其餘的特 一分析,寬持一個互影響力的特 馬匹出賓續號是不達履帮的數字碼 馬匹出賓續號是不達履帮的數字碼 制. 而還積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 整是最小。決定數字面與是非常面 整整是最小。決定數字面與是非常面 對此, 也將出數字面與是一組, 所得的原 重要對的數字面與完整一組, 人遷和等於 所得到的答案是:一,二,四,十 一、十三、十四錢一組, 三,五 積到的答案是:一,二,四,十 一、九為另一組, 七、八,十篇最構 積到的數字種類比較差合馬匹排位精 微認的數字種類比較差合馬匹排位精

#### 如何建出分裂目標

**会成调爆数键重频,可以開始**進

Page 20/65

之公司内部的一位 [選擇公式]引向諸認治的分析。[ ]分析電気]

除了「累滞公式」外,就是因约 書標。本質子的病或是用二出分裂, 書標。本質子的病或是用二出分裂, 書一種是的預式是用多枝前分裂。 書。就是設在同一層面把所有特面信 引加強,或器就点的短別,可加一例, 四進分裂的預式,每一次分裂只有常 枝節,在例中鄉一層而有帶線透透者 積少裂強的傳成還者能力進行 下次分裂, (更需一), 医语名质德共 在三個層面就行分裂。

#### 分裂開始甚至同學下來

當分裂開始了,起裝把它停止下 來?方法是有很多,比較簡單的方符 是用分裂後的數據數量。當分出來的 這別感多,若一個提出的數據數量。

#### 國國發行會不會「北京」

旧看相關總統1-4-5 南亞維維用港市於佔金属維索加高 民低。原因是西國在分類目1-1-1 廣泛起時。的講員,目行#臺灣民族 職號本和5的附前。這時來天的第

## Data Mining Problem Types

一輪的分裂中被濡用。當分望後的。 超過有能力的結發時,分裂會不斷 環,直到樹葉出現,組狀的設置間 實形成。

數值建築可分為文字和數字。 字種類比數字種類簡單、結素值是一 個個體、沒有先後次序之分,時便可 以分開或組合、在本文例子中,馬匹 的名稱就是文字通短。而款字通短, 要分寫「不連續型」和「建領型」, 屬匹出實調號是不遵續型的數字通短, 局匹出實調號是不遵續型的數字通知。 素就計學的爆聲差可以非為分割 基礎。分割後的每一小組、所得的影響 要,若某絕其匹排位特徵而為例。

6 types

,作例股定决策视

用一個數據與共三百二十 資料取材於香港調馬記斯 其中一場參賽馬(共十四 住賽蹟(見表一)。現在來 樹的運作。首先決策樹廣式 在起點的樹根,當用家庭定 的特徵後,顏式把其餘的特 新,業份一個且影纖力的特

Often a combination solves the problem

(型的數字種類比較適合馬匹排位) (\*

如何是出分裂目。

·异动谓撑数律输冠 • 可以開始

Page 21/65

## Data Description and Summarization

- Aims at concise description of data characteristics
- Lower end of scale of problem types
  - Provides the user an overview of the data structure

### Typically a sub goal

#TF17100 年代東備 在用一個數據庫共三百三十二 。資料取材於香港賽馬記錄, 用其中一場套賽馬(共十四 過往賽蹟(見表一)。現在落 繁樹的運作。首先決策加預式 数在起點的樹根,當用家吃完 析的特徵後,演式把其餘的潮 - 各杯,要找一個且影響力的潮

六、九為另一組,七、八,十萬最後 一起。可見分析結果完全不同,還有 的是分析結果的可用性,筆者覺得速 積型的數字種類比較適合馬匹排位精 微。

#### 如何是出分裂目標

李武涛爆動擁着題,可以開始進

Page 22/65





Separates the data into interesting and meaningful subgroups or classes

Manual or (semi)automatic

A problem for itself or just a step in solving a problem

> 理的數字種類比較適合馬匹排位約 14

如何是出分裂目。

字**此满撑数操意源**,可以開始

Page 23/65



總容動工序。而所得到的結果又做有基本 ■致這個原因多數是用素不明白每個工序 和參數的用途。其實輸入參數有助引導工 有限的時間裏尋找有用的資料。自動化的 IINING ← 於你能否了解工序。配合學

## Classification

#### 通過2000次下級結構 通過有能力的特徵時,分裂

Assumption: existence of objects with characteristics that belong to different classes

Building classification models which assign correct labels in advance

Exists in wide range of various application

Segmentation can provide labels or restrict data sets

的问题出分级目标

Page 24/65

## **Concept Description**

Understandable description of concepts or classes

# Close connection to both segmentation and classification

# Similarity and differences to classification

馬作例設定決策被

在用一個數據庫共三百三十二 ,資料取材於香港雲馬記錄, 用其中一場 參賽馬(共十四 )過住賽蹟(見表一)。現在來 )過住賽蹟(見表一)。現在來 )激在起點的樹根,當用家庭定 術的特徵後,嶺式把其餘的等 一份好,寬均一個且影響力的等 一、十二、十四、一、二、五、 六、九為另一組、七、八、十四最後 一起。可見分析結果完全不同,還有 的是分析結果的可用性,筆者覺得達 續發的數字道類比較適合馬匹排位精 徵。

#### 如何是出分裂目標

**李成演爆數據意識**,可以開始這

Page 25/65

技師,在例中第一層而有 帽。二進分型從的時度遭 下次分裂,(見屬一)馬 在三個層面積行分裂。

分股開始長如同停下部?

當分裂開始了,忽候用 來?方法是有很多,比較简 是用分裂後的數據散量。當



## **Prediction (Regression)**

一輪的分裂中被濡用。當分望後的 超是有能力的結發時,分裂會不斷 環,直到創葉出現,組狀的法律關 實形成。

數值種類可分為文字和數字。 字種類比數字種類簡單,特徵值是 個個體,沒有先後次序之分,簡似

Finds the numerical value of the target attribute for unseen objects



### Similar to classification - difference: discrete becomes continuous

的是分析結果的可用性,筆者覺得達 實學的數字證题比較適合馬匹排位結 費。

#### 如何是出分裂目

李武涛爆動儘續線(可以開始)

Page 26/65

在三個層面就行分裂。

分数期始最如何停下呢?

當分裂開始了,忽禄把它停止了 來?方法是有很多,比較簡單的方位 是用分裂機的數據數量,當分出來的

## **Dependency Analysis**

### Finding the model that describes significant dependences between data items or events

### Prediction of value of a data item

#### 数字含另一组。而

### **Special case: associations**

一起。可見分析結果完全不同,還有 的是分析結果的可用性,筆者覺得速 續型的數字種類比較適合馬匹排位結 後。

#### 如何是出分裂目標

· 字成選擇數據範圍 · 可以開始過

Page 27/65

在三個層面進行分裂。

分级院始长如同样下部?

當分裂開始了,怎樣把它停止 來?方法是有很多,比較簡單的方向 是用分裂機的數據數量,當分出來自

DATAMINE 軟件時,又需要輸入很多更 能移動工序,而所得到的結果又沒有基层 導致這個原因多數是用數不明白每個工序 和參數的用途。其實輸入參數有助引導工 有限的時間裏尋找有用的資料。自動化的 MINING 是在於你能否了解工序,配合参 和DATAMINING 的能力,從而得到偏正

## **Data Mining Models**

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。) 請決策細,原名思義就是將整 的流程,把其中可能牽涉到的 件,有系統地以期狀的結構型 。決策樹是由樹恨、樹節和樹 成(見壓一)。樹根是分析的 樹底是感點、地能是分差路 \$決策點)。

#### 馬作例設定決策被

1在用一個數據庫共三百三十二 ,資料取材於香港賽馬記錄, 一用其中一場參賽馬(共十四)過往賽寶(見表一)。現在來 、策樹的運作。首先決策樹廣式 就在在起點的樹根,當用家庭定 計析的特徵後,讀式把其餘的特 一分析,寬持一個互影響力的特

#### 如何還出分裂目標

会成選擇數據酸類,可以開始進

Page 28/65

除了「劉滞公式」外,就是即約 書標。本質子的病或是用二因分裂, 書標。本質子的病或是用二因分裂, 書標。就是說在同一層面把所有特面的 引加強,或器就立的冠別,可加一例 二進分裂的病式,每一次分裂所有常 技節,在例中鄉一層面省閉根透透者 稱。二進分型整的特成還有能力進行 下次分裂,但具需一刻, 而近代裂。

#### 分裂開始甚至同學下來

當分裂開始了,想裝把它停止下 來?方法是有很多,比較簡單的方在 是用分裂後的數據數量。當分出來的 這別感多,若一個視測的數律對量為

#### 國國發行會不會「北京」

但他相關關鍵。 「小和和國語」 此任。原因是王書在外如用11日 個派起時。的講著。目行由主書目 關鍵。約5.約批測。這時有大的

## Neural Networks

Characterizes processed data with single numeric value

Efficient modeling of large and complex problems

**Based on biological structures Neurons** 

Network consists of neurons grouped into layers

的是分析結果的可用性,筆者覺得運 資源的戲字道與比較適合馬匹排色特 發。

#### 如何是出分裂目的

李甫馮擇數據範領(可以開始)



## **Neuron Functionality**

,作例設定決策權

# II WI I2 W2 I3 W3 In Wn

### Output = f (W1\*I1, W2\*I2, ..., Wn\*In)

后出, 策樹的運作。首先決策間面式 放在起點的樹根,當用家庭定 析的特徵後,面式把其餘的特 發好,需找一個且影纖力的差

如何是出分裂目標

**完成選擇數據範圍**,可以開始

Page 30/65



的特面後,領式化共脉的符。 44杯,開始——值且影響力的將 **卓山溝爆動儘能類**(可)

Page 31/65





A way of representing a series of rules that lead to a class or value

Iterative splitting of data into discrete groups maximizing distance between them at each split

**Classification trees and regression trees** 

**Univariate splits and multivariate splits** 

Unlimited growth and stopping rules CHAID, CHART, Quest, C5.0

Page 32/65

※啓動工序・面所得到的結果又沒有並属 致這個原因多數是用素不明白每個工序 自參數的用途。其實輸入參數有助引導工 可服的時間裏尋找有用的資料。自動化的 INING ご於你能否了解工序,配合錄 (NING)

## **Decision Trees**

一輪的分裂中被濫用。當分望後的 超還有能力的結散時,分裂會不斷 環,直到樹葉出現,離狀的決策對 會形成。

版 加速效可分离义子 伊根 字 遗嘱比散字 禮頭 簡單,特贵 僧個體,没有先後次序之分, 以分崩或領合,在本文例子中

Balance>10

ULE DATE WE WE HE WE

在用一個數據廣共三百三十二 ,資料取材於香港賽馬記錄, 用其中一場參賽馬(共十四 過住賽蹟(見表一)。現在來 策樹的運作,首先決策樹廣式 放在起點的樹根,當用家庭定 析的特徵後,搞式把其餘的特 )音楽是:一、二、 二、十四條一規・7

Age<=32

### Married=NO

的是分析結果的 複型的數字種2 卷。

的何處出分裂目標

**註說這擇數據範疇**,可以!

Page 33/65

SINCLAIP SINCLAIP CONTRACTOR

Age>32

Married=YES

Balance<=10

2啓動工序,而所得到的結果又進去重要 致這個原因多數是用家不明白每個工序 0多數的用途。其實輸入參數有助引導工 1股的時間裏导找有用的資料。自動化的 INING 二 於你能否了解工序,配合參

## **Decision Trees**

一輪的分裂中被適用。當分望後回。 超還有能力的特徵問。分裂會不斷改 環,直到刺菜出現,相狀的表類樹創

用其中一场要求時、大十口 過估實踐(見表一)。現在來 能樹的運作。首先決策關菌式 做放在起點的樹根,當用家脏定 析的特徵後,銜式把其餘的特 品好,需找一個具影響力的差

實現的數字證單比較適合馬匹排位。 對。

如何是出分裂目的

**会成演播數據範題**(可以開始過

Page 34/65

分裂运出更如同得下派

當分裂開始了,怎樣把它停止 來?方法是有很多,比較簡單的方 是用分裂後的數據數量。當分出來

総容動工序・面所得到的結果又做有益属 致這個原因多數是用家不明自每個工序 和參數的用途。其實輸入參數有動引導工 有限的時間裏导找有用的資料。自動化的 IINING 分於你能否了解工序・配合學 IINING の能力、從面得到電源

## **Rule Induction**

一輪的分裂中被選用。當分望後的 結選有能力的結發時,分裂會不開 環、直到樹葉出現,組成的決測量

### Method of deriving a set of rules to classify cases

**Creates independent rules that are unlikely to form a tree** 

**Rules may not cover all possible situations** 

### **Rules may sometimes conflict in a prediction**

如何是出分裂目。

Page 35/65







If balance>100.000 then confidence=HIGH & weight=1.7

If balance>25.000 and status=married then confidence=HIGH & weight=2.3

If balance<40.000 then confidence=LOW & weight=1.9



好, 元州——值且影墨力的韩

Page 36/65
## K-nearest Neighbor and Memory-Based Reasoning (MBR)

Usage of knowledge

of previously solved similar problems in solving the new problem

Assigning the class to the group where most of the k-"neighbors" belong

First step – finding the suitable measure for distance between attributes in the data

How far is black from green?

+ Easy handling of non-standard data types

- Huge models

Page 37/65



## Data Mining Models and Algorithms

### Many other available models and algorithms

- Logistic regression
- **Discriminant analysis**
- Generalized Adaptive Models (GAM)
- Genetic algorithms
  - Etc...

Many application specific variations of known models

Final implementation usually involves several techniques

Selection of solution that match best results

Page 39/65

DATAMINE 軟件時,又需要輸入電子學 能移動工序,而所得到的結果又沒有基层 導致這個原因多數是用數不明白每個工序 和參數的用途。其實輸入參數有助引導工 有限的時間裏尋找有用的資料。自動化的 MINING 是在於你能否了解工序,配合參 MINING 是在於你能否了解工序,配合參

## **Efficient Data Mining**

擇合用的軟件,而且在應用這 時候,便可避重就輕,大大發 TAMINING 的功力。 請決策細,原名思義就是將整 的流程,把其中可能牽涉到的 件,有系統地以趨肤的培構型 。決策樹是由樹恨、樹踏和樹 成(見壓一)。樹根是分析的 執葉是感點,她能是分差路 \$決策點)。

#### 馬作例設定決策被

1在用一個數據庫共三百三十二 ,資料取材於香港賽馬記錄。 一用其中一場拿賽馬(共十四)過住賽寶(見表一)。現在來 、策樹的運作。首先決策樹廣式 就在在起點的樹根,當用家庭定 計析的特徵後,領式把其餘的特 一分析,寬持一備且影響力的待 馬匹出宴調號是不遵職帮的數字看 個。而運職型的特徵值需要找一個分 朝點,統計學的標準差可以作為分割 總處。分割後的每一小組、所得的標 準差是最小。決定數字風氣是非常重 要。若果绝馬匹拌位特徵何為一個 一、一個人一個、大選和等於 所得到的答案是:一、二、四、十 二、九賞另一組、而不定積型的數字 所得到的答案是:一、二、四、十 二、九賞另一組、七、八、十篇最 補 一、九賞另一組、七、八、十篇最 補 一、九賞另一組、七、八、十篇最 補 一、九賞另一組、七、八、十篇 最 一、九賞另一組、七、八、十 篇 最 一、九賞另一組、七、八、十 篇 最 一 後 一

#### 如何還出分裂目標

李成選擇數據睡識,可以開始進 "

Page 40/65

這個公司/1920年 這個公式」引向錯進77百元相一世 1分析實現1

除了「劉滞公式」外,就是即約 8件 集合子的演式是用二周分裂, 8 件 是 能 開 式 是 用 多枝前分裂, 5 積 分裂 會 把 有 極 力 的 修 微 只 全 动 1 章 就是 號 在 同 一 層 面 把 所 有 析 面 信 2 加 弦, 成 路 房 立 的 冠 所, 可 加 十 阿 2 進 分裂的 演式, 每 一 次 分裂 所 有 朗 2 進 分裂的 演式, 每 一 次 分裂 所 有 朗 2 進 分 裂 检 前 特 成 型 有 相 机 馬 匹 名 制 。 二 進 分 裂 检 的 特 成 型 有 相 机 馬 匹 名 制 。 二 進 分 裂 检 的 特 成 型 有 超 力 進 行 第 次 分裂 素 何 更 留 一 》 馬 匹 名 斯 想 并 在 三 做 册 面 就 行 裂 。

#### 分裂開始長安同停下來

當分裂開始了,起裝把它停止了 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 這別感多,若一個這些的數據數量。

#### 国際分析論系領に対応した

に当相関部党1-4-5 病存在採用建由公共全国社会中 1966年頃歴史新学区分析10000000 商行起時。的講者、目行相主者目的 開始本和5.的批問、課時本大約30



Page 41/65

# **DM Process Model**

一輪的分裂中被選用。當分望後即今 超還有能力的結發時,分裂會不斷 環,直到創葉出現,關於的決策困難 實務成。

5A – used by SPSS Clementine (Assess, Access, Analyze, Act and Automate)

SEMMA – used by SAS Enterprise Miner (Sample, Explore, Modify, Model and Assess)

### **CRISP-DM – tends to become a standard**

·六、九賞另一組,七、八、十寫最後 一趟。可見分析結果完全不同,還有 的是分析結果的可用性,筆者覺得通 被型的數字種類比較適合馬匹排位結 徵。

#### 如何是出分裂目的

- 字成溝場數據範圍,可以開始書

Page 42/65

著。二進が韓國的号の連有権の項目 下次分裂、(見圖一)馬匹名解題共 主面層面進行分裂。

分裂開始長如同學下來?

當分裂開始了,忽禄把它停止一 來?方法是有俱多,比較簡單的方法 是用分裂後的數據數量,當分出來自



## **CRoss-Industry Standard for DM Conceived in 1996 by three companies:**

## DAIMLERCHRYSLER

Page 43/65



# CRISP – DM methodology

### Four level breakdown of the CRISP-DM methodology:



初的通作。首先从凤间两天 在起點的樹根,當用家脏定 的特徵後,演式把其餘的等 將,還投一個且影響力的將

何**建出分裂目標** 完成清掃軟據範續,可以開

Page 44/65

## Mapping generic models to specialized models

- Analyze the specific context
- Remove any details not applicable to the context
- Add any details specific to the context
- Specialize generic context according to concrete characteristic of the context
- Possibly rename generic contents to provide more explicit meanings

E用一個數據庫共三百三十二 資料取材於香港宴馬記録。 用其中一場參賽馬(共十四 個住賽蹟(見表一)。現在來 除樹的運作。首先決策側面式 放在起點的樹根,當用家庭定 所的特徵後, 演式把其餘的特 补析, 需找一個且影響力的特

六、九貧另一祖,七、八,十為反領 一超。可見分析結果完全不同,還有 的是分析結果的可用性,筆者帶得運 實態的數字邏輯比較適合馬匹排色特 數。

#### 如何是出分裂目標

<u>会成演播數據範圍</u>,可以開始港

Page 45/65

每一一地方或後的中心進行之。 下次分裂。(夏陽一)馬匹名制 在三面層面進行分裂。

分数期始最如何停下部注

當分裂開始了,忽候把它停止 來?方法是有很多,比較簡單的方。 是用分裂機的數據數量,當分出來自

## **Generalized and Specialized** Cooking

- Preparing food on your own
  - Ratinstake with wegetables?

    - Find the recipe for that meal Check the Cookbook or call mom Gather the ingredients

  - Defrost the meat (if you had it in the fridge) Prepare the meal Buy missing ingredients

    - - EPipyr Kowrtie from the neighbors
  - Clean,

ater)

dishes

Page 46/65

# CRISP - DM model

- Business understanding
- Data understanding
- Data preparation
- Modeling
- Evaluation
- Deployment

用具甲一吻 新要為(共下四 過住要職(見表一)。現在來 策樹的運作。首先決策關廣式 放在起點的樹根,當用家吃定 析的特徵後,讀式把其餘的特 分析, 戰快一個且影響力的特 - 虹。可見分析結果完全不同、更有 为是分析結果的可用性,筆者覺得的 動學的數字種類比較適合馬匹排位有 素。

#### 如何是出分裂目標

Page 47/65



# **Business Understanding**

一輪的分裂中被還用。當分望後的小 這還有能力的特徵時,分裂會不斷備 重,直到樹葉出現,如於由美筆樹創

## **Determine business objectives**

Assess situation

## **Determine data mining goals**

## Produce project plan

E用一個數庫庫共三百三十二 資料取材於香港宴馬記録。 用其中一場參賽馬(共十四 動住實蹟(見表一)。現在來 做個的運作。首先決策側面式 做在起點的樹根,當用家庭定 新的特徵後,領式把其餘的特 新術,飄粉一個且影響力的特

六、九宫另一祖、七、八、十萬慶後 一題。可見分析結果完全不同、還有 的是分析結果的可用性,筆者覺得速 續發的數字種類比較適合馬匹排位結 發。

#### 如何是出分裂目的

**李成满楼数據和版**(可以開始港

Page 48/65

留。二進分裂後的特容還有能力 下次分裂。(見圖一)馬匹名制 在三個層面幾行分裂。

分级院始版如何停下时程

當分裂開始了,想樣把它停止下 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



# Data Understanding

Collect initial data

### Describe data

## Explore data

## Verify data quality

十二 六、九鹙另一組,七、八、十 線, 一短。可見分析結果完全不同 十四 的是分析結果的可用性,筆者 在來 積型的數字種類比較適合馬匹 廣式 微。

#### 如何是出分裂目標

完成演播數據意識,可以開始港

Page 49/65

在三個屬面就行分裂。

分裂開始長如同停下於

當分裂開始了,想錄把它停止下 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

# **Data Preparation**

## Select data

Clean data

Construct data

## Integrate data

Format data

一起。可見分析結果完全不同、處有 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 微。

#### 如何是出分裂目的

皇武演爆數據範圍,可以開始港

Page 50/65

技部分 裂 着把有些2000 %3 揮, 就是設在周一層而把所 分选來, 成為集立的組所。 三進分裂的訴式, 每一次分 較高, 在刘中旗一層而有兩 不力裂。二進分裂後的特徵還有 下方量, 〔其斷一〕馬匹 在二個層面積行分裂。

分级影曲复如同停下来已

當分裂開始了,怎樣把它伴! 來?方法是有很多,比較簡單的? 是用分裂後的數據數量。當分出;



総督動工序・南所得到的結果又做各面展 素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 有限的時間裏尋找有用的資料。自動化的 IINING → 於你能否了解工序・配合參



## Select modeling technique

## Generate test design

## Build model

## **Assess model**

#### 馬作例股定決策被

在用一個數據庫共三百三十二 , 資料取材於香港裏馬記錄, 用其中一場參賽馬(共十四 )過住賽蹟(見表一)。現在來 案樹的運作。首先決策固廣式 数在起點的樹根,當用家脏完 析的特徵後,讓式把其餘的等 - 各種, 戰役一個且影響力的等 五數字錄另一組。而不還讓型的數字 所得到的答案是:一、二、四、十 一、十三、十四條一組,三,五, 六、九為另一組,七、八、十為最補 一起。可見分析結果完全不同,還者 的是分析結果的可用性,筆者看得到 續型的數字種類比較適合馬匹排位結 徵。

#### 如何是出分裂目標

**会成演漫數據範圍**,可以開始港

Page 51/65

很新分裂者把早期的市场的市场。 算,就是跟在同一層面把作 分出來,就是跟立的起所。 四應的名稱就有十四倍級別 工業分裂物辦式,每一次分 較節,在例中第一層而有兩 層。二進分裂機的轉要還有 下次分裂,(其屬一)馬匹 在三個層面還行分裂。

分积累结长如何停下时

當分裂開始了,怎樣把它徑) 來?方法是有很多,比較簡單的; 是用分裂後的數據數量。當分出;



影響動工序。而所得到的結果又做容量展 ■致這個原因多數是用家不明自每個工序 和參數的用途。其實輸入參數有助引導工 有限的時間裏尋找有用的資料。自動化的 IINING ← 於你能否了解工序。配合參



## results = models + findings

### **Evaluate results**

### **Review process**

## **Determine next steps**

馬作例設定決策被

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 過住賽蹟(見表一)。現在來 策樹的運作,首先決策划演式 放在起點的樹根,當用家庭定 析的特徵後,演式把其餘的特 分析,要投一個且影響力的券 一、十三、十四為一絕,三,五, 六,九當另一組,七、八,十萬最後 一起。可見分析結果完全不同,還有 的是分析結果的可用性,筆者量得速 複型的數字種類比較適合馬匹排位特 徵。

#### 如何是出分裂目標

李武涛爆動據藏羅,可以開始港

Page 52/65

校町、在時干燥一層両 瞬。二進分裂後的特徴。 下次分裂、(見編一)」 在三個層面幾行分裂。

當分裂開始丁,忽然 來?方法是有很多,比影 是用分裂後的數據數量。 品別總多,若一個的方法







一幅的分裂中很适用。 當分裂破即少 結過有能力的特徵時,分裂會不斷適 環,直到樹葉出現,相狀的決策固創 電形成。

### Plan deployment

### Plan monitoring and maintenance

### **Produce final report**

### **Review project**

#### 事作例設定決策被

在用一個數據庫共三百三十二 , 資料取材於香港裏馬記錄, 用其中一場參賽馬(共十四 )過住賽蹟(見表一)。現在來 案樹的運作。首先決策固廣式 数在起點的樹根,當用家脏完 析的特徵後,讓式把其餘的等 - 各種, 戰役一個且影響力的等 一、十三、十四時一起,三、五、 六、九賞另一起,七、八、十萬泉德 一起。可見分析結果完全不同,還有 的是分析結果的可用性,筆者覺得速 續型的數字種類比較適合馬匹排位特 徵。

#### 如何處出分裂目標

· 字成演漫數據範題 · 可以開始港

Page 53/65

匹應的站場就看十四時起初。但有 二進分裂的演式,每一次分裂因利 技師,在例中第一層而有用相思也。 磨。二進分裂後的特徵還有能力地 下次分裂,(更屬一)馬匹名稱進 在三個層面處行分裂。

分裂属出版如同律下示

當分裂開始了,忽候把它停止 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



E啓動工序。而所得到的結果又沒有当所 要改這個原因多數是用素不明自每個工序 因多數的用途。其實輸入多數有助引導工 有限的時間裏尋找有用的資料。自動化的 INING ALL 於你能否了解工序。配合學 IDATAN ING 的能力。從而得到集正



ò

MODE

看到一些手服,例如分别、贯 畫 條 等 它 們 都 是 在 INING 技術中出來的,可以 该術的應用是非常替通,一半 DATAMENE 軟件開發公司, 该術加入其系統中,這技術就 例(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 作。優點和缺點,不但可以為 時候,便可遵電就輕,大大號 AMINING 的功力。) 請決策樹,履名思發就是將整 的流程,把其中可能牽涉到的 件,有系統地以樹狀的結構堅 。決策樹是由樹慢、樹銷和約

#### 馬作例股定決策被

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 )過住賽蹟(見表一)。現在來 )過住賽蹟(見表一)。現在來 )微放在起點的樹根,當用家庭定 )析的特徵後,讓式把其餘的特 ,各種,寬持一個具影響力的特



要,若果把馬 對像,違履型 是:少陰另一, 所得到的答案 一、九賞見分 的是分析結果, 微型的數字種 微型。

如何是出分率





基础分析像不像「主要」》 更有思致相比3、4、3、6%。 費希該形用法由主任金術性工作由 比任3、原因是来考社会也同样目 但「起既」的准備。即時編去費用 編發4、和5、的准備。即時編天均的人

Page 54/65

DATAMINE 軟件時,又需要輸入電子手 能移動工序,而所得到的結果又沒有基层 等效這個原因多数是用款不明白每個工序 和參數的用途。其實輸入參數有助引導工 有限的時間裏尋找有用的資料。自動化的 MINING 是在於你能否了解工序,配合参

## **Available Software**

擇合用的軟件。而且在應用這 時候,便可避重就輕,大大發 TAMENENG 的功力。》 請決策樹,原名思義就是將整 的流程,把其中可能牽涉到的 件,有系統地以趨肤的培構型 。決策樹是由樹恨、樹節和樹 成(見屬一)。樹根是分前的 樹底是感點、她能是分差路 \$P>樂點)。

#### 馬作例設定決策被

1在用一個數據庫共三百三十二 ,資料取材於香港賽馬記錄, 一用其中一場拿賽馬(共十四)過住賽寶(見表一)。現在來 ,除樹的運作。首先決策樹領式 ,就在起點的樹根,當用家庭定 ,析的特徵後,領式把其餘的特 ,分析,寬持一備且影響力的待 國伍的重數地域模型的聲字磁模, 與 回出實續號是不進價聲的數字確模, 明 副 前連續型的特徵值需要找一個分 副點,統計準的標準差可以作為分割 影燈。分割後的每一小組,所得的原 準差是最小,決定數字風類是非常強 要,若果把馬匹排位特徵(許為所 對像,連續型的數字所得到)的答案 是: 心於面數字徑一組,大邊和等於

五數字貨另 组。而不逆接到 所得到的答案是:一、二、 一、十三、十四爲一組、三、五、 六、九貨另一組、七、八、十萬最後 一組。可見分析結果完全不同、還有 的是分析結果的可用性。筆者醫得連 複型的數字種類比較適合周匹排位特 卷。

#### 如何是出分裂目鼻

**会成選擇數據確認**,可以開始進

Page 55/65

海洋台口中的中国。 「選擇公式」引向錯進方向分析。但 日分析要與1

除了「關係公式」外,就是此的 將標。本面子的演式是用三次分裂, 另一種是的演式是用多校符分裂, 各都分裂會把有部力的特别只全分 課,就是說在同一層面把所有許面前 分加底,或器就立的组別,可如一切 定進分裂的演式,每一次分裂只有再 校館,在例中落一層面有閉線透透常 稱。二違分裂後的傳送還有能力進行 下次分裂, (更屬一) 馬匹名個處非 在三個層面成行分裂。

#### 分裂開始長如同停下來

當分裂開始了,忽樂光在停止下 來?方法是有很多,比較簡單的方法 是用分裂後的數據數量。當分出來自 這別這多,若一個這比的數據數量。

#### 11月月3日,11日不會「北京」1

に通過問題設計 - 病不被採用送店並佔金高級。 - 第五、前國建選者在外前前面目目 - 高近起點。的語彙 - 但行程至何因 - 韓林本的本的時間, 這時本来的

# **Comparison of forteen DM tools**

The Decision Tree products were

- CART
- Scenario
  - See5
- S-Plus
- The Rule Induction tools were
  - WizWhy DataMind

  - **DMSK**

Neural Networks were built from three programs

- NeuroShell2
- **PcOLPARS**
- PRW
- The Polynomial Network tools were
  - ModelQuest Expert
  - Gnosis
  - a module of NeuroShell2
  - KnowledgeMiner

Page 56/65

# Criteria for evaluating DM tools

A list of 20 criteria for evaluating DM tools, put into 4 categories:

- *Capability* measures what a desktop tool can do, and how well it does it
  - Handless missing data
  - Considers misclassification costs
  - Allows data transformations
  - Quality of tesing options

  - Has programming language Provides useful output reports
  - Visualisation

Page 57/65



# Criteria for evaluating DM tools

*Learnability/Usability* shows how easy a tool is to learn and use:

Tutorials
Wizards
Easy to learn
User's manual
Online help
Interface

新律、權權(Mana) (A = 50,000) (A =

#### 如何是出分裂目標

字成溝場數據範題,可以開始港

Page 58/65

(2.1) 的最短短期间的 为外菌,在本例子中、局面的 高端保全式」者把作的资源 所能在分析前。如果知道运行。 一般的新闻型的方向力排。 一般的新闻型方向力排。 一般就有一般的新闻之后。 一般的新闻是用多枝前外别。 不可引导的新闻是用多枝前外别。 是是有能力的特别是全别的 是是有能力的特别是全别的 是是有能力的特别是全别的 是是有能力的特别是全别的 是是有能力的特别是全别的 是是有些力的特别是全别的 是是有些力的特别是全别的 是是有些力的特别是全别的 是是有些力的特别是全别的 是是有些力的特别是全别的 是是有些力的特别是一种 是是有些力的特别是一种 是是有些力量的。 是是一种的新闻是一种。 是是一种的新闻是一种。 是是一种的新闻是一种。 是一种的新闻是一种。 是一种的"是一种"。 是一种。 是一种"是一种"。 是一种"是一种"

#### 分裂透出复如同种下部?

當分裂開始了,想錄記它停止了 報2方法是有俱多,比較簡單的方位 是用分裂後的數據數量。當分送來自



原有相関制制3、4、3、6% 備存該採用是由計估金価償本的面3 比值。原因是至常な分析の11 值「起鉄、動温素、口仔細企業10

# Criteria for evaluating DM tools

*Interoperability* shows a tool's ability to interface with other computer applications

- Importing data
- Exporting data
- Links to other applications

### • Flexibility

- Model adjustment flexibilityCustomizable work environment
- Ability to write or change code

#### 作例設定決策権

生用一個數據庫共三百三十二 · 資料取材於香港宴馬記録。 用其中一場參裏馬(共十四 過住實蹟(見表一)。現在來 發樹的運作。首先決策崩面式 放在起點的樹根,當用家庭定 所的特徵後,讓式把其餘的特 等杯,寬投一個且影響力的等 一、十三、十四為一組,三,五, 六,九為另一祖,七、八,十萬最後 一起。可見分析結果完全不同,還有 的是分析結果的可用性,筆者量得速 續型的數字種類比較適合馬匹排位精 像。

#### 如何是出分裂目的

· 完成調擇數據種類 · 可以開始港

Page 59/65

二地历秋时间代。5000年1月前有用相助医常 第《二進分型後的特徵還有能力進行 新次分型、(更屬一)馬匹名稱進出 生三個層面還行分裂。

#### 分期開始長如同傳下影響

當分裂開始了,忽候把它停止了 來?方法是有俱多,比較简重的方法 是用分裂後的數據數量。當分出來自



# A classification of data sets

### Pima Indians Diabetes data set

- 768 cases of Native American women from the Pima tribe some of whom are diabetic, most of whom are not
- 8 attributes plus the binary class variable for diabetes per instance

### Wisconsin Breast Cancer data set

- 699 instances of breast tumors some of which are malignant, most of which are benign
- 10 attributes plus the binary malignancy variable per case
- The Forensic Glass Identification data set
  - 214 instances of glass collected during crime investigations
  - 10 attributes plus the multi-class output variable per instance
- Moon Cannon data set
  - 300 solutions to the equation:
    - $x = 2v 2 \sin(g) \cos(g)/g$
    - the data were generated without adding noise

地分析結果的可用性,筆著看 例的數字種類比較適合馬匹亞 -

#### 如何是出分裂目的

完成演繹數據藏題,可以開始進

### Page 60/65

# Evaluation of forteen DM tools

Technology	Tool	Capability	Learnability/ Usability	Interoperability	Flexibility	Accuracy	Overall (equal weights)	Price (US\$)
Tree	CART	+	$\checkmark$	-	✓ +	+	✓+	995
	Scenario	√-	+	+	-		$\checkmark$	695
	See5	<	√-	$\checkmark$	√-	+	$\checkmark$	440
	S-Plus	+	√-	++	+	+	+	1,795
	Tree Average*	~	$\checkmark$	✓+	$\checkmark$	✓+	✓+	Median = 845
Rule	WizWhy	~	✓ +	$\checkmark$	√-	-	$\checkmark$	4,000
	DataMind	✓ +	++	+	√-	~	✓+	25,000
	DMSK	-		√-	-	+	-	75
	Rule Average*	$\checkmark$	$\checkmark$	✓+	-	$\checkmark$	$\checkmark$	Median = 4,000
Neural	NeuroShell 2	-	$\checkmark$	-	-	++	√-	395
	PcOLPARS	√-	-	-	√-	$\checkmark$	√-	495
	PRW	✓ +	+	++	$\checkmark$	++	+	10,000
	Neural Average*	√-	$\checkmark$	$\checkmark$	√-	+	√-	Median = 495
Poly Net	MQ Expert	+	$\checkmark$	$\checkmark$	✓ +	+	✓+	5,950
	NeuroShell 2	√-	~	$\checkmark$	~	+	$\checkmark$	495
	Gnosis	√-	~		√-	++	√-	4,900
	KnowledgeMiner	-	-	-	√-	+	-	100
	Poly Net Average*	√-	√-	-	$\checkmark$	+	√-	Median = 2,698
	Overall Average*	$\checkmark$	$\checkmark$	$\checkmark$	√-	✓ +	$\checkmark$	Median = 845

は世界後期になっていた。



DATAMINE 軟件時,又需要輸入電子師 能移動工序,而所得到的結果又沒有基层 等效這個原因多數是用數不明白每個工序 和參數的用途。其實輸入參數有助引導工 有限的時間裏尋找有用的資料。自動化的 MINING 是在於你能否了解工序,配合参 MINING 是在於你能否了解工序,配合参

## Conclusions

爆合用的软件。而且在應用這 時候,便可變電就輕,大大發 AMINING 的功力。 請決策樹,原名思發就是將整 的流程,把其中可能牽涉到的 件,有系統地以期狀的結構呈 。決策樹是由樹根、樹節和樹 成(見圖一)。樹根是分析的 樹底是感點、樹前是公美路 \$決策點)。

#### 馬作例設定決策被

1在用一個數據庫共三百三十二 ,資料取材於香港賽馬記錄, 一用其中一場拿賽馬(共十四)過住賽寶(見表一)。現在來 、策樹的運作。首先決策樹廣式 就在在起點的樹根,當用家庭定 計析的特徵後,領式把其餘的特 一分析,寬持一個互影響力的特 馬匹出賣誦號是不連續證的數字麵 圖。而運續證的特徵值需要找一個分 割點,統計學的標準差可以非為分割 基礎。分割後的售一小組、所得的標 書差是最小。決定數字風融是非常領 書差是最小。決定數字風融是非常領 基礎。若果絕馬匹將位特徵何為分類 的數字路男、組。而不遭積型的數字 所得到的答案是:一、二、四、十 一、九為另一組、七、八、十屬最有 的是分析結果的可用性、筆者覺得通 積型的數字種類比較適合馬匹排位特 素。

#### 如何建出分裂目標

会成演播數據種類,可以開始進

Page 62/65

海洋市市地址。 「選擇公式」引向錯誤市向計批一個 日分析要與1

除了「累滞公式」外, 乾隆時的 結構。本質子的國或是用二組分裂, 各一種是的關或是用多枝前分裂。 各種分裂會把有些力的特徵只全導 素。就是說在同一層面把所有特面的 引加強, 或器就立的超近, 可如一切 正進分裂的演式, 每一次分裂因指異 枝節, 在個中鄉一層面有用根馬匹名 稱。二進分型後的傳成還有能力進行 下次分裂, 可提冊一製。

#### 分裂開始甚至同學下來

當分裂開始了,起裝把它停止下 來?方法是有很多,比較簡單的方符 是用分裂後的數據數量。當分出來的 這別感多,若一個這男的數據數量。

#### 國國分析會不會「北北」」

に当相関語設計 4-5 病不振祥用是由於語金属操作的言う 其近。原因是要常在外前的1000日 廣泛起動。的講者 目行相差常用所 開始來和5 的批冊。這時來大約55



字成演播數據載題 · 可以開始

啓動工序。面所得到的結果又就有面层 致這個原因多數是用家不明白每個工序 多數的用途。其實輸入參數有助引專工 限的時間裏尋找有用的資料。自動化的 NING 分於你能否了解工序。配合夢 DATAM ING 的能力。從而得到真正



GREED SLOTH ENVY WRATH PRIDE LUST

GLUTTONY

Seven deadly sins. Seven ways to die.

Page 64/65



MINING 技術中出來的,可 技術的應用是非常審測,一 DATAMINE 軟件階發公司 技術加入其系統中,還技術 個(DECISION TREE)。 應用履泛,所以清楚掌握與 作、優點和缺點,而且在應用 作、優點和缺點,而且在應用 構成。便可避重就輕,大 TAMINING 的功力。 謂決策個,頗名思幾就是將 的流程,把其中可能牽強與 代,策個是由期假、樹錦是 以 相葉是紊點,個語是分差 (決策點)。

事件例設定決策視

|注用一個數據庫共三百三十 ,資料取材於香港賽馬配劃 用其中一場參賽馬(共十 )過住賽蹟(見表一)。現在 )邊裡的運作。首先決策關係 做在起點的樹根,當用家願 術的特徵後,顏式把其餘的 一番杯,飄炒一個且影響力的



Page 65/65