# 白血病醫學文件分類研究

報告人: 蘇佳益

指導老師: 陳聰毅

國立高雄科技大學電子工程系

# Agenda

- Introduction
- Introduction to Database
- Introduction to System
- Overview of Dataset
- Model Evaluation
- Classification Algorithm
- Reference

# Introduction to Bioinformation

- 生物資訊所包含的範圍
  - 病理影像
  - 病例
  - 醫學資料
- 本研究主要利用文本資料分類出白血病的類別

# Introduction to the Categories in this Research

- AML
- ALL
- CML
- CLL
- AML&ALL
- AML&CML
- ALL&CML
- ALL&CLL，其中有四個為有相同的基因表示名稱，也就是說兩個類別有交集，所以將兩個結合成一個類別。在此研究中，分別比較三種演算法，Neural Network、Naïve Bayes、Decision Tree，並探討其效能。

# Overview of Leukemia

- 什麼是白血病?
- 白血病種類
  - 急性淋巴性白血病(ALL)
    - 好發於小孩，成人約占20%，此類別有可以分成三型。
  - 急性骨隨性白血病(AML)
    - 好發於成年人，此疾病又可分為七型。
  - 慢性淋巴性白血病(CLL)
    - 常發生於50~55歲以上的老年人，此疾病又分為三型。
  - 慢性骨隨性白血病(CML)
    - 源自於造血細胞的一種惡性腫瘤，由於染色體的病變。

# Introduction to Database

# PubMed

- PubMed 系統簡介
- PubMed 查詢法則
  1. 自然語言搜尋
  2. MeSH(Medical Subject Headings)
- MeSH搜尋結果(以鳥(Bird)為例)
  - 搜尋結果
- 現今有許多醫學研究都是採用此資料庫

# Entrez Gene

- What is Entrez Gene?
- 可以找出與PubMed相對應的ID

| *gene2pubmed* | *attribute* | *description* |
|---|---|---|
| | *tax_id* | The unique identifier provided by NCBI Taxonomy for the species or strain/isolate. |
| | *GeneID* | The unique identifier for a gene. |
| | *PubMed ID* | The unique identifier in PubMed for a citation. |

gene2pubmed | gene_info

```
#tax_id GeneID  PubMed_ID
9       1246500 9873079
9       1246501 9873079
9       1246502 9873079
9       1246503 9873079
9       1246504 9873079
9       1246505 9873079
9       1246509 10984505
9       1246510 10984505
```

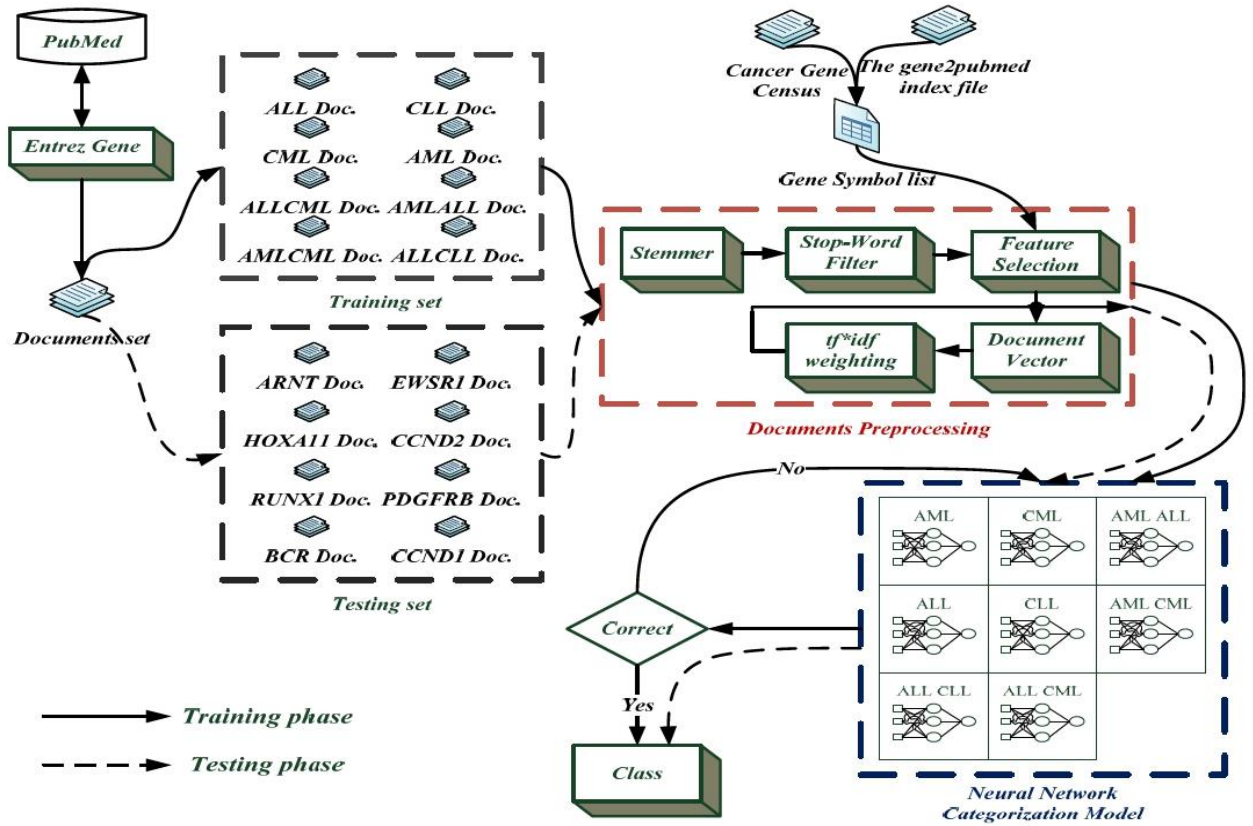
```
#tax_id GeneID  Symbol  LocusTag    Synonyms    dbXrefs chromosome  map_location    description type_of_gene    Symbol_from_nomenclature_authority  Full_name_from_nomenclature_authority   Nomenclature_status Other_designations  Modification_date   Feature_type
7   5692769 NEWENTRY    -   -   -   -   Record to support submission of GeneRIFs for a gene not in Gene (Azotirhizobium caulinodans.  Use when strain, subtype, isolate, etc. is unspecified, or when different from all specified ones in Gene.).  other   -   -   -   -   20190202    -
9   1246500 repA1   pLeuDn_01   -   -   -   -   putative replication-associated protein protein-coding  -   -   -   -   20180129    -
9   1246501 repA2   pLeuDn_03   -   -   -   -   putative replication-associated protein protein-coding  -   -   -   -   20180129    -
9   1246502 leuA    pLeuDn_04   -   -   -   -   2-isopropylmalate synthase  protein-coding  -   -   -   -   20180129    -
9   1246503 leuB    pLeuDn_05   -   -   -   -   3-isopropylmalate dehydrogenase protein-coding  -   -   -   -   20180129    -
```

# Cancer Gene Census

- What is Cancer Gene Census?
- 此研究採用此資料集及Entrez Gene 所提供的gene2pubmed來做特徵選取

# Introduction to System

# System Architecture



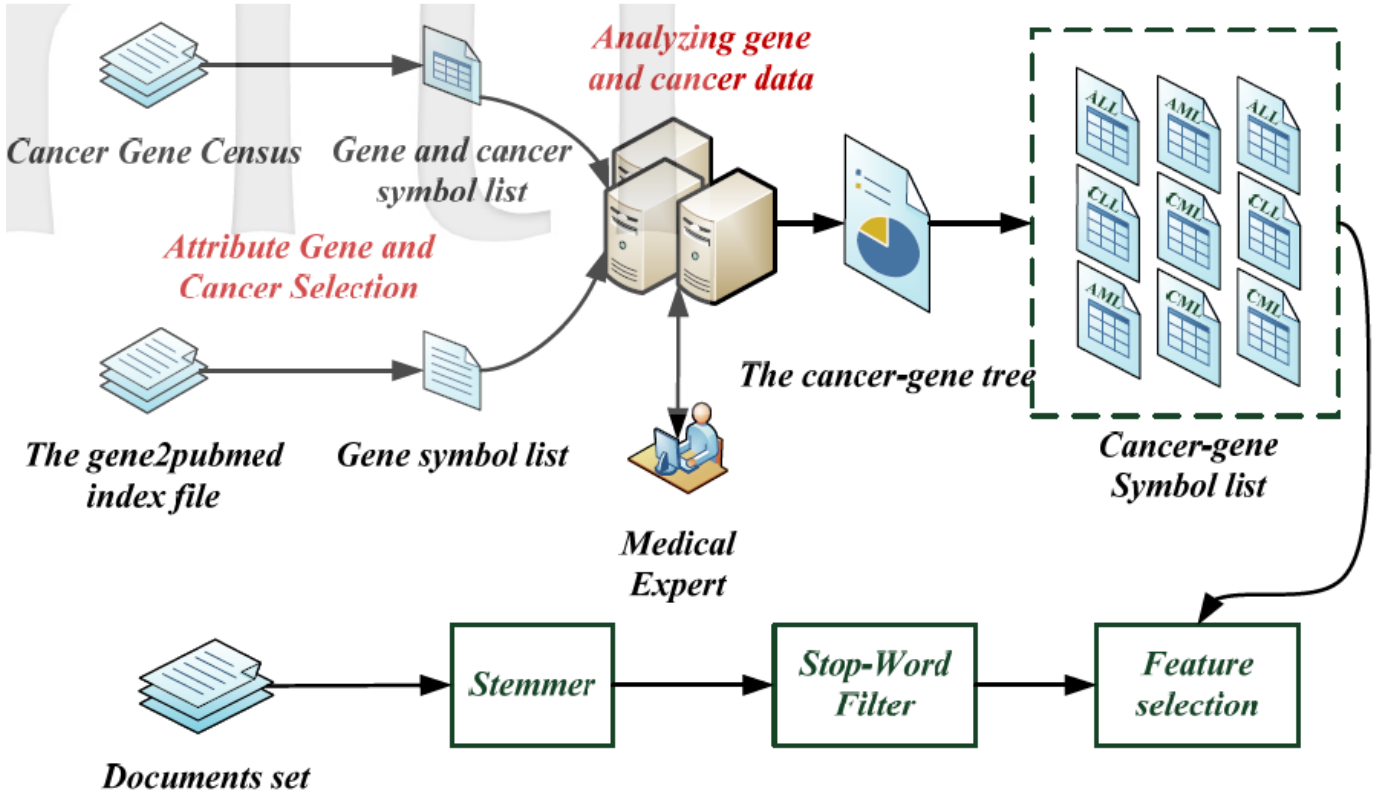
# Stemmer

- What is Stemmer?
- Problem?
- The difference between Stemmer and Lemmatization?

# Feature Selection

1. 從gene2pubmed 及 Cancer Gene Census 中建立基因表示名稱 並各自建立一個list。
2. 將文件進行編碼並進行統計分析
3. 將基因名稱的list做成參考依據。

# Architecture of Feature Selection



# Overview of Dataset

# Genetic Type List – Single Type

| Symbol | Gene ID | Tumor |
|---|---|---|
| ABL2 | 27 | AML |
| AF15Q14 | 57082 | AML |
| ARHGEF12 | 23365 | AML |
| ARNT | 405 | AML |
| CEBPA | 1050 | AML |
| CREBBP | 1387 | AML |

| Symbol | Gene ID | Tumor |
|---|---|---|
| AF1Q | 10962 | ALL |
| AF3p21 | 51517 | ALL |
| AF5q31 | 27125 | ALL |
| CDK6 | 1021 | ALL |
| EWSR1 | 2130 | ALL |
| FBXW7 | 55294 | ALL |

| Symbol | Gene ID | Tumor |
|---|---|---|
| D10S170 | 8030 | CML |
| HOXA11 | 3207 | CML |
| MSI2 | 124540 | CML |

| Symbol | Gene ID | Tumor |
|---|---|---|
| BCL3 | 602 | CLL |
| BTG1 | 694 | CLL |
| CCND2 | 894 | CLL |
| FSTL3 | 10272 | CLL |
| MYC | 4609 | CLL |
| TCL1A | 8115 | CLL |

# Genetic Type List – Combination

| Symbol | Gene ID | Tumor |
|---|---|---|
| FLT3 | 2322 | AML, ALL |
| JAK2 | 3717 | AML, ALL |
| MLL | 4297 | AML, ALL |
| NUP214 | 8021 | AML, ALL |
| PICALM | 8301 | AML, ALL |
| RUNX1 | 861 | AML, ALL |

| Symbol | Gene ID | Tumor |
|---|---|---|
| BCL11B | 64919 | ALL,CLL |
| BCL9 | 607 | ALL,CLL |
| CCND1 | 595 | ALL,CLL |
| IGH@ | 3492 | ALL,CLL |

| Symbol | Gene ID | Tumor |
|---|---|---|
| EVI1 | 2122 | AML, CML |
| PDGFRB | 5159 | AML, CML |
| RPL22 | 6146 | AML, CML |

| Symbol | Gene ID | Tumor |
|---|---|---|
| ABL1 | 25` | ALL, CML |
| BCR | 613 | ALL, CML |

# Crude Dataset

- quantity of label Documents

| Tumor Types | ALL | AML | CLL | CML | ALL,CLL | ALL, CML | AML, ALL | AML, CML |
|---|---|---|---|---|---|---|---|---|
| No. of doc. | 532 | 512 | 538 | 128 | 475 | 500 | 521 | 468 |

- quantify of genetic documents

How to deal with this data?

| Gene | EWSR1 | ARNT | CCND2 | HOXA11 | CCND1 | BCR | RUNX1 | PDGFRB |
|---|---|---|---|---|---|---|---|---|
| No. of doc. | 161 | 175 | 200 | 101 | 180 | 223 | 181 | 195 |
| Tumor Types | ALL | AML | CLL | CML | ALL,CLL | ALL, CML | AML, ALL | AML, CML |

# Normalized Dataset

- quantity of label documents

| Tumor Types | ALL | AML | CLL | CML | ALL,CLL | ALL, CML | AML, ALL | AML, CML |
|---|---|---|---|---|---|---|---|---|
| No. of doc. | 500 | 500 | 500 | 500 | 500 | 500 | 500 | 500 |

- quantity of genetic documents

| Gene | EWSR1 | ARNT | CCND2 | HOXA11 | CCND1 | BCR | RUNX1 | PDGFRB |
|---|---|---|---|---|---|---|---|---|
| No. of doc. | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| Tumor Types | ALL | AML | CLL | CML | ALL,CLL | ALL, CML | AML, ALL | AML, CML |

# Model Evaluation

# Parameters for Evaluation

- True Positive(TP) –The ground-truth value is true and the predicted value is true as well.
- False Positive(FP) – The ground-truth value is false and the predicted value is true as well.
- True Negative(TN) – The ground-truth value is false and the predicted value is false as well.
- False Negative(FN) – The ground-truth value is true and the predicted value is false as well.

# Evaluation Methods

- Definition of Precision, Recall and F1-Score
- Formulae

1. Precision $\frac{TP}{TP+FP}$

2. Recall $\frac{TP}{TP+FN}$

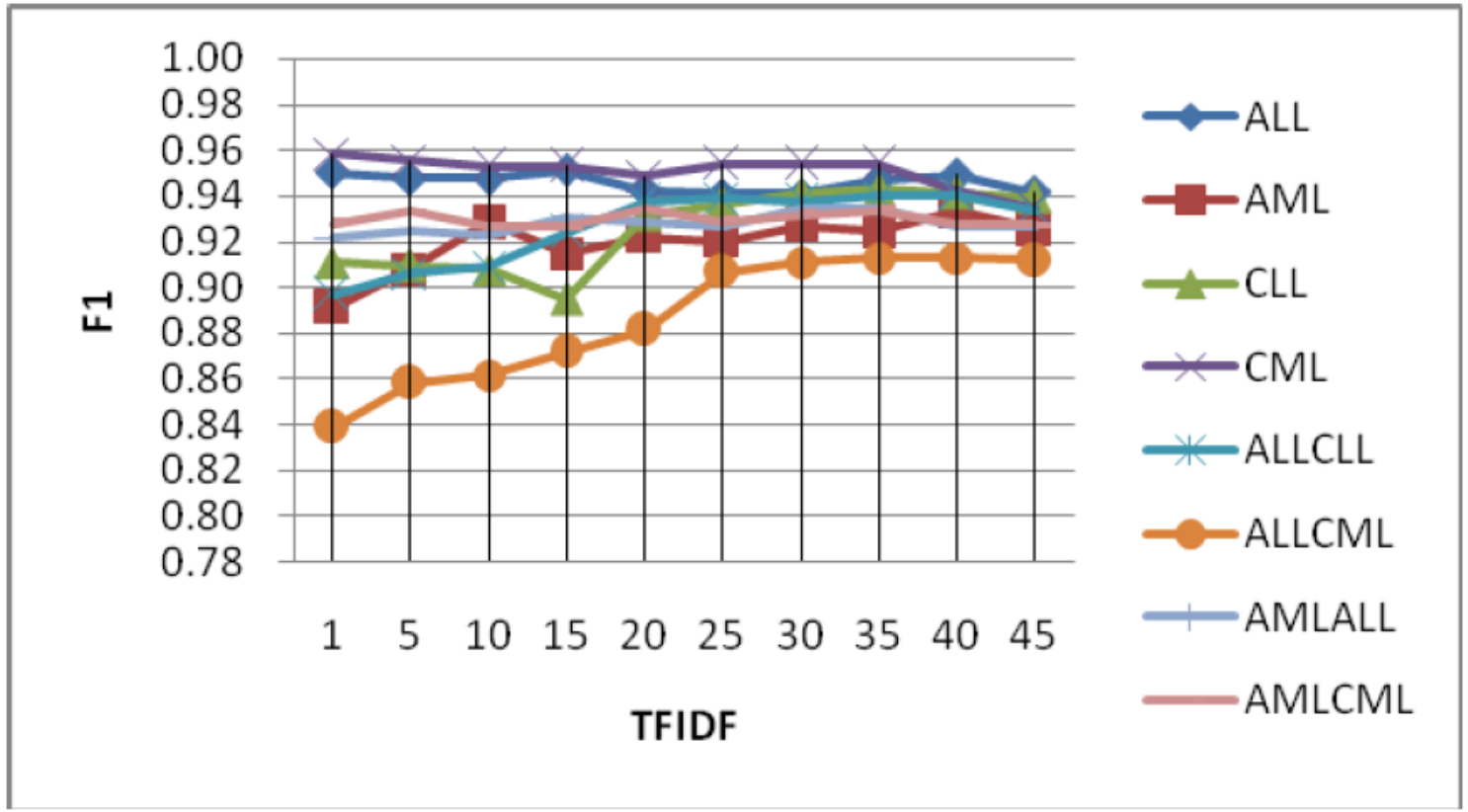
3. F1-Score $\frac{2*Precision*Recall}{Precision+Recall}$

- Why do we need these ways to evaluate model?

# Result of the Best Performance

| Category | TFIDF | dim | precision | recall | F1 |
|---|---|---|---|---|---|
| **ALL** | 40 | 256 | 0.9151 | 0.9857 | 0.9491 |
| **AML** | 40 | 223 | 0.8937 | 0.9762 | 0.9331 |
| **CLL** | 35 | 253 | 0.9011 | 0.9883 | 0.9427 |
| **CML** | 35 | 355 | 0.9213 | 0.9891 | 0.9540 |
| **ALLCLL** | 40 | 225 | 0.8949 | 0.9910 | 0.9405 |
| **ALLCML** | 40 | 224 | 0.8607 | 0.9719 | 0.9129 |
| **AMLALL** | 35 | 284 | 0.8873 | 0.9864 | 0.9342 |
| **AMLCML** | 35 | 271 | 0.8872 | 0.9981 | 0.9338 |

# F1-Score in Different TF-IDF



# Classification Algorithms

# Decision Tree

# Information Theory

- Entropy
  - Entropy defines the amount of information provided by an event. The lower the probability is, the more information load is.
  - Formula of Entropy

  $Entropy(p) = -\sum_{i=1}^{n} p_i * log_2(p_i)$, n = the number of sample
- Gain Information
  - Gain tells us how much information about the specific class we get from the feature.
  - Formula of Gain Information

  $Gain(p,T) = Entropy(p) - \sum_{j=1}^{n}(p_j * Entropy(p_j))$
- Gain Ratio
  - Gain Ratio is the normalized term of Gain Information
  - Formula of Gain Ratio

Gain Ratio$= \frac{Gain(p,T)}{splitinfo(p,T)}$

# ID3 Algorithm - Introduction

- ID3 is used to construct tree based on information gain.
- Suppose we want to classify “whether we should play ball or not”

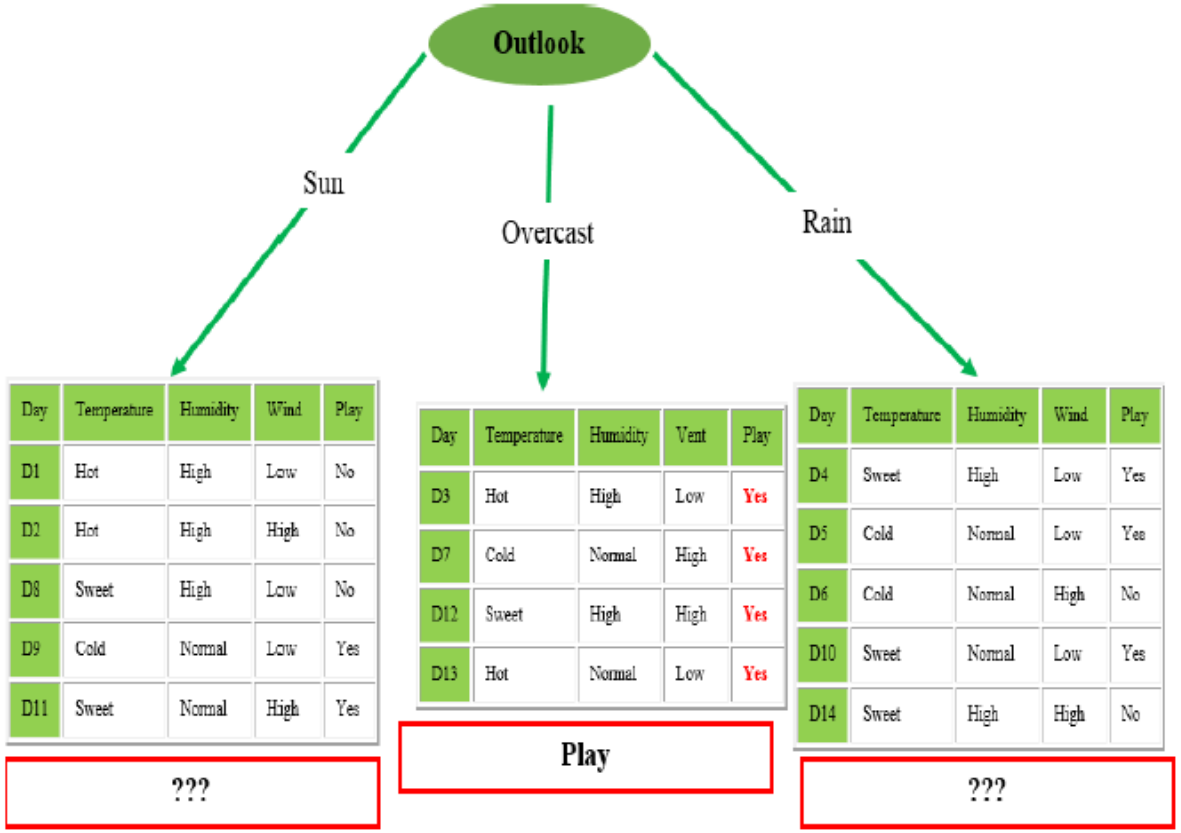
| Day | Outlook | Temperature | Humidity | Wind | Play |
|---|---|---|---|---|---|
| **D1** | Sun | Hot | High | Low | No |
| **D2** | Sun | Hot | High | High | No |
| **D3** | Overcast | Hot | High | Low | Yes |
| **D4** | Rain | Sweet | High | Low | Yes |
| **D5** | Rain | Cold | Normal | Low | Yes |
| **D6** | Rain | Cold | Normal | High | No |
| **D7** | Overcast | Cold | Normal | High | Yes |
| **D8** | Sun | Sweet | High | Low | No |
| **D9** | Sun | Cold | Normal | Low | Yes |
| **D10** | Rain | Sweet | Normal | Low | Yes |
| **D11** | Sun | Sweet | Normal | High | Yes |
| **D12** | Overcast | Sweet | High | High | Yes |
| **D13** | Overcast | Hot | Normal | Low | Yes |
| **D14** | Rain | Sweet | High | High | No |

# ID3 Algorithm – Example

- Example of wind attribute
  - Entropy(S) = $-\frac{9}{14} * log_2\left(\frac{9}{14}\right) - \frac{5}{14} * log_2\left(\frac{5}{14}\right) = 0.94$
  - Gain(S, Wind)=Entropy(S)$-\frac{8}{14} * Entropy(Low) - \frac{6}{14} * Entropy(High) = 0.048$
  - $Entropy(Low) = -\frac{2}{8} * log_2\left(\frac{2}{8}\right) - \frac{6}{8} * log_2\left(\frac{6}{8}\right)$
  - $Entropy$(High)=$-3/6 * log_2\left(\frac{3}{6}\right) - 3/6 * log_2\left(\frac{3}{6}\right)$
- Gain of all attributes
  - Gain(S, Wind) =0.048
  - Gain(S, Temperature) =0.0289
  - Gain(S, Humidity) =0.1515
  - Gain(S, Outlook) =0.246

the most largest one

# ID3 Algorithm – Result



# C4.5 Algorithm - Introduction

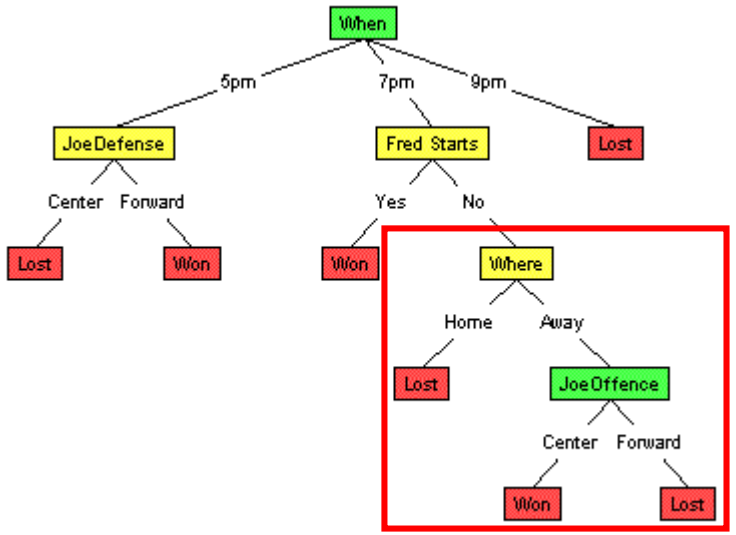
- C4.5 has some improvements on ID3. C4.5 also employs Gain Ratio instead of Gain.
- Advantages of C4.5 Algorithm
  - unknown value
  - values on continuous interval
  - reducing overfitting
  - more accurate and efficient while comparing with ID3

| Size of Data Set | Algorithm | |
|---|---|---|
| | ID3 (%) | C4.5 (%) |
| 14 | 94.15 | 96.2 |
| 24 | 78.47 | 83.52 |
| 35 | 82.2 | 84.12 |

| Size of Data Set | Algorithm | |
|---|---|---|
| | ID3 (%) | C4.5 (%) |
| 14 | 0.215 | 0.0015 |
| 24 | 0.32 | 0.17 |
| 35 | 0.39 | 0.23 |

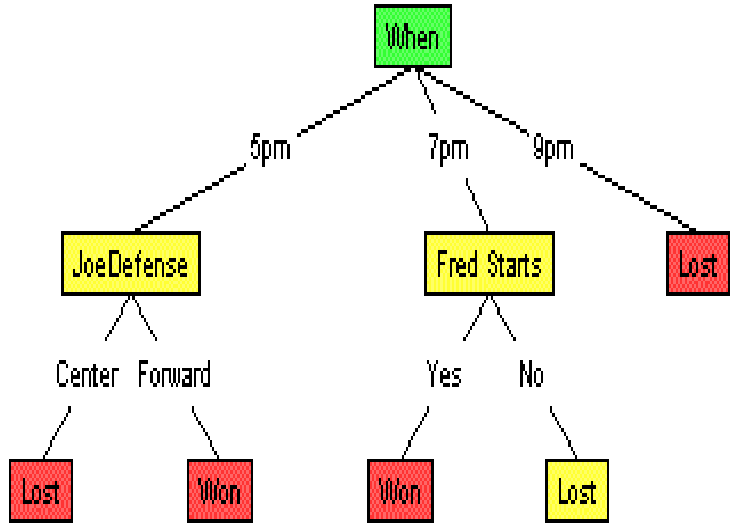
# C4.5 Algorithm - Pruning

- Pruning is a technique to reduce our tree for avoiding “overfitting”.
- Reduced Error Pruning




suppose we have higher accuracy than unpruned tree

it still has many ways to prune our tree. We just easily introduce one way.



# Application on Text Classification

1. Treating every keywords as our attribute
2. Quantilizing our data e.g. TF, TF-IDF
3. Calculating the values that we need for decision tree

# Naïve Bayes

- Based on Bayes Theory
- Having i.i.d. assumption
- However: It is hard to satisfy this assumption in the real data.
- Naïve Bayes still performs very well.
- We will discuss Naïve Bayes by using text classification problem.

# Posterior Probability

- Determining the probability that the sentence belongs which class given a set of words
- Formula

$$posterior\ probability = \frac{(conditional\ probability)*(prior\ probability)}{evidence}\ \alpha\ (conditional\ prob.) * (prior\ prob.)$$

# Prior Probability

- Determining the probability of each class
- Usually uniformly distributed
- The posterior probability will depend on conditional probability when the prior probability is an uniform distribution
- Formula of Prior Probability
  - P($c_j$) = $\frac{N_j}{N}$, j = 1 , 2…, (Number of Words)
  - N $_j$: the number of words in the specific class
  - N: the number of all words

# Conditional Probability

- Determining the probability of the specific word in the specific class.
- Formula of Conditional Probability
  - $P(w_j \mid c_i) = \frac{N_{c_i,w_j}}{N_{c_i}}$ , i = 1,…., Number of Categories, j = 1, …, Number of Words
  - $N_{c_i,w_j}$ : the number of word j in the class i
  - $N_{c_i}$: the number of word in the class i

# Naïve Bayes: Practical Discussion

- Problem: The posterior probability will be 0 when the specific word does not appear in the specific class.
- Resolution: In order to prevent this situation, we will add smoothing term.
- Problem: too many digits after decimal point
- Resolution: Using log probability

# Naïve Bayes: Additive Smoothing

- Smoothing term is called $\alpha$.
- We also call Laplace Smoothing when the $\alpha$ is one
- The revised formula is as follows:
- $P(w_j \mid c_i) = \frac{N_{X_i, w_j} + \alpha}{N_{c_i} + \alpha D}$ , i = (1,…., Number of Categories), j = (1, …, Number of Words)
- D: the number of all words

# K-nearest Neighbor – Distance Metric

- There are two distance metrics for K-nearest Neighbor. Those are illustrated as follows.
- Distance Metric

| L1 Distance | L2 Distance |
| --- | --- |
| $D_1 = \sum_p \left\lvert I_1^p - I_2^p \right\rvert$ | $D_2 = \sqrt{\sum_p \left( I_1^p - I_2^p \right)^2}$ |

- Algorithm Complexity
  - Training Complexity: $O(1)$
  - Test Complexity: O(N), where N is the number of Examples.

# K-nearest Neighbor (CONT.)

- Steps of KNN
  - Store Data
  - Calculate Distance Metric
  - Find k nearest neighbor
  - Find the label that appears the most

# Experimental Result

| *F1 measure* | | Classifier | | |
|---|---|---|---|---|
| | | **ANN** | **BP** | **DT** |
| **Gene doc.** | **EWSR1** | **0.9531** | 0.9258 | 0.9343 |
| | **ARNT** | **0.9323** | 0.9119 | 0.9237 |
| | **CCND2** | **0.9576** | 0.9343 | 0.9412 |
| | **HOXA11** | **0.9751** | 0.8423 | 0.9478 |
| | **CCND1** | **0.9453** | 0.9117 | 0.9132 |
| | **BCR** | 0.9053 | 0.9137 | **0.9122** |
| | **RUNX1** | **0.9257** | 0.9147 | 0.9152 |
| | **PDGFRB** | 0.8957 | 0.8973 | **0.9153** |

# References


李俊宏, 類神經網路與癌症基因統計資訊應用於醫學文件分類研究

Badr HSSINA et al., A comparative study of decision tree ID3 and C4.5

Raschka , Naive Bayes and Text Classification I Introduction and Theory

CS231n: Convolutional Neural Networks for Visual Recognition, Stanford University

Machine Learning Crash Course, Google

Decision Tree Puning

白血病簡介

PubMed

PubMed 規則

Entrez 資料