時序資料探勘與軟體操作

國立宜蘭大學資訊工程系
吳政瑋 助理教授
wucw@niu.edu.tw

教學目標

- 了解時序資料特徵擷取應用
- 了解下列時序資料特徵擷取方法
  - 時域資料特徵(Time Domain Feature)
  - 斜率特徵(Slope Feature)
  - 序列特徵(Sequence Feature)
  - 輪廓特徵(Shapelet Feature)
- 使用軟體擷取時序資料特徵
- 使用軟體分析特徵重要程度

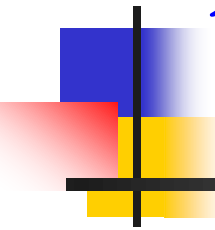
教學單元

- 時序資料特徵擷取應用
- 時序資料特徵擷取方法
- 使用軟體擷取時序資料特徵
- 使用軟體分析特徵重要程度

教學單元
時序資料特徵擷取應用

時序資料特徵擷取應用

- 時序資料探勘前處理
- 時序資料分類
- 時序資料分群
- 型樣擷取
- 資訊檢索



教學單元
時序資料特徵擷取方法

時序資料特徵擷取方法

- 時域資料特徵(Slope Feature)
- 斜率特徵(Slope Feature)
- 序列特徵(Sequence Feature)
- 輪廓特徵(Shapelet Feature)

教學單元
時域資料特徵特徵擷取方法

Basic Features of a Time Series

- Let $X = \langle x_1, x_2, ..., x_n \rangle$ be the time series data of an arbitrary axis of the accelerometer data in the segment, where x_i is the i-th sample in X.

- The length of X is defined as $|X| = n$
- The maximum of X is defined as $max(X) = max\{x_1, x_2, ..., x_n\}$
- The minimum of X is defined as $min(X) = min\{x_1, x_2, ..., x_n\}$
- The range of X is defined as $Range(X) = |max(X) - min(X)|$

平均值(Mean)

- The mean is a measure of central tendency that reflects the average of values in *A*, which is defined as

$$Mean(A) = \bar{A} = \frac{\sum_{i=1}^{n} a_i}{n}$$

舉例說明：計算平均值

- Mean
  - Given a data set

$$Mean(A) = \bar{A} = \frac{\sum_{i=1}^{n} a_i}{n}$$

	t_1	t_2	t_3
X_i	1	3	5
Y_i	-3	1	-1
Z_i	4	1	-5

- $\overline{X} = \frac{1+3+5}{3} = 3$
- $\overline{Y} = \frac{(-3)+1+(-1)}{3} = -1$
- $\overline{Z} = \frac{4+1+(-5)}{3} = 0$

變異數(Variance)

- The *variance* is a measure that reflects the average squared deviation of each value in *A* from the mean of *A*, which is defined as

$$Var(A) = \frac{\sum_{i=1}^{n}(a_i - \bar{a})^2}{n}$$

舉例說明：計算變異數

- Variance

$$Var(A) = \frac{\sum_{i=1}^{n}(a_i - \bar{a})^2}{n}$$

  - Given a data set

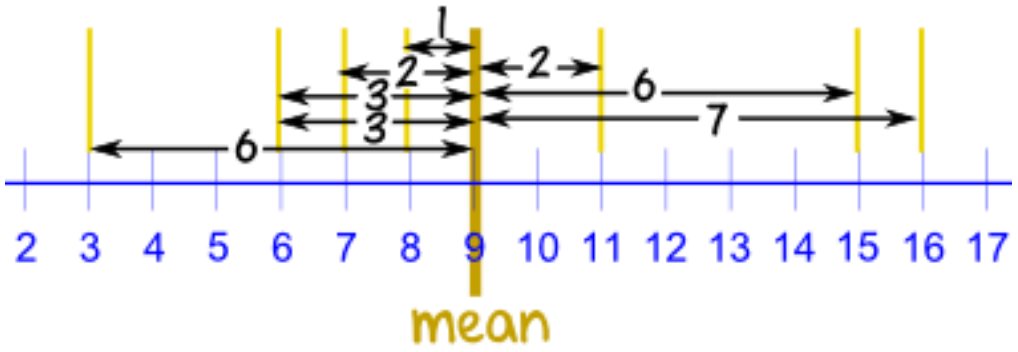
	t_1	t_2	t_3
X_i	1	3	5
Y_i	-3	1	-1
Z_i	4	1	-5

  - $var(X) = \frac{(1-3)^2+(3-3)^2+(5-3)^2}{3} = \frac{8}{3}$
  - $var(Y) = \frac{(-3-(-1))^2+(1-(-1))^2+(-1-(-1))^2}{3} = \frac{8}{3}$
  - $var(Z) = \frac{(4-0)^2+(1-0)^2+(-5-0)^2}{3} = 14$

標準差(Standard Deviation)

- The standard deviation of A is defined as $\sigma_A = Std(X)$
- $[Std(X)]^2 = \sigma_A^2 = Var(X)$

Mean Absolute Deviation (MAD)



Step 1: Find the mean (the average)
Step 2: Find the *deviations*: subtract each number and the mean.
Step 3: Find the average of all of the "deviations"
(make the answers positive in step 2, add them all ,then divide)

Example: Find the mean absolute deviation of:
80, 85, 81, 0, 85, 90, 87, 92

Step 1:
Find the average:

80 + 85 + 81+0 + 85 + 90 + 87 + 92 = 600
600/8 = 75

Step 2:
Find the deviations

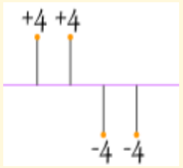
80 - 75 = **5** 85 - 75 = **10** 81 - 75 = **6** 0 - 75 = **-75**
85 - 75 = **10** 90 - 75 = **15** 87 - 75 = **12** 92- 75 = **17**

Step 3:
Find the average of the absolute value of the deviations:

5 + 10 + 6+ 75 + 10 + 15 + 12 + 17 = 150
150/8 = **18.75**

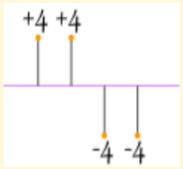
*Footnote: Why *square* the differences?

If we just add up the differences from the mean ... the negatives cancel the positives:



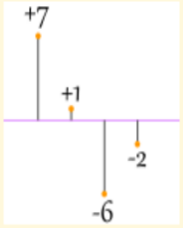
$$\frac{4 + 4 - 4 - 4}{4} = 0$$

So that won't work. How about we use absolute values ?



$$\frac{|4| + |4| + |-4| + |-4|}{4} = \frac{4 + 4 + 4 + 4}{4} = 4$$

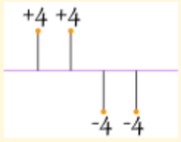
That looks good (and is the Mean Deviation), but what about this case:



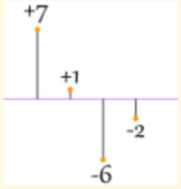
$$\frac{|7| + |1| + |-6| + |-2|}{4} = \frac{7 + 1 + 6 + 2}{4} = 4$$

Oh No! It also gives a value of 4, Even though the differences are more spread out.

So let us try squaring each difference (and taking the square root at the end):



$$\sqrt{\left(\frac{4^2 + 4^2 + 4^2 + 4^2}{4}\right)} = \sqrt{\left(\frac{64}{4}\right)} = 4$$



$$\sqrt{\left(\frac{7^2 + 1^2 + 6^2 + 2^2}{4}\right)} = \sqrt{\left(\frac{90}{4}\right)} = 4.74...$$

That is nice! The Standard Deviation is bigger when the differences are more spread out ... just what we want.

舉例說明：偏度(Skewness)

$$M_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k$$

$$\text{skewness } g_1 = \frac{\sqrt{n}M_3}{M_2^{\frac{3}{2}}}$$

- Skewness
  - Given a data set

	t_1	t_2	t_3
X_i	1	3	5
Y_i	-3	1	-1
Z_i	4	1	-5

- $skewness\ X =$

$$\sqrt{3}\left(\frac{(1-3)^3+(3-3)^3+(5-3)^3}{3}\right)/\left(\frac{(1-3)^2+(3-3)^2+(5-3)^2}{3}\right)^{\frac{3}{2}} =$$

$$\sqrt{3}\left(\frac{(-8)+0+8}{3}\right)/\left(\frac{4+0+4}{3}\right)^{\frac{3}{2}} = 0$$

舉例說明：峰度(Kurtisis)

- Kurtosis
  - Given a data set

$$\text{kurtosis } g_2 = \frac{nM_4}{M_2^2} - 3$$

	t_1	t_2	t_3
X_i	1	3	5
Y_i	-3	1	-1
Z_i	4	1	-5

- $kurtosis\ X =$

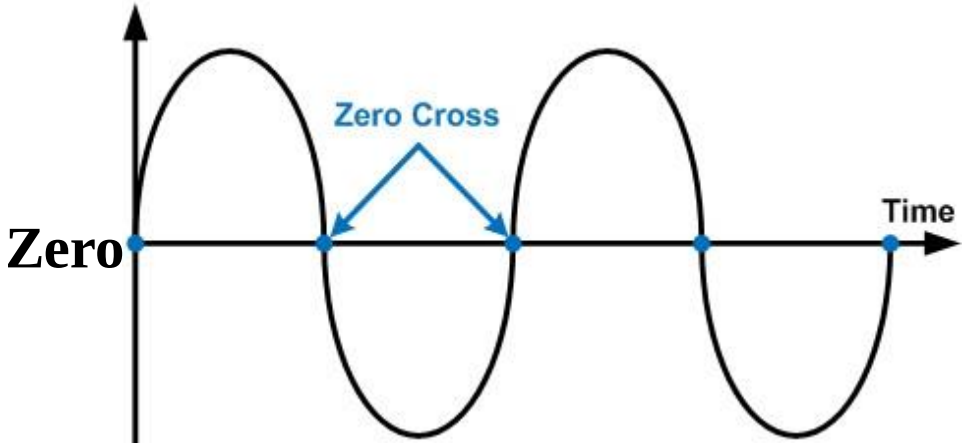
$$3\left(\frac{(1-3)^4+(3-3)^4+(5-3)^4}{3}\right)/\left(\frac{(1-3)^2+(3-3)^2+(5-3)^2}{3}\right)^2 - 3 =$$

$$3\left(\frac{16+0+16}{3}\right)/\left(\frac{4+0+4}{3}\right)^2 - 3 = 4.5 - 3 = 1.5$$

過零率 (Zero Crossing Rate；ZCR)

- The *zero cross rate* is a measure that reflects how many times the sign of two adjacent values in *A* changes from positive to negative or vice versa, which is defined as

$$ZCR(A) = \frac{\sum_{i=2}^{n} |sign(a_i) - sign(a_{i-1})|}{2}$$



$sign(a_i)$	$sign(a_{i-1})$	$\|sign(a_i) - sign(a_{i-1})\|$
1	1	|1 − 1| = 0
1	-1	|1 − (-1)| = |2| = 2
-1	1	|(-1) − 1| = |-2| = 2
-1	-1	|(-1) − (-1)| = 0

舉例說明：計算過零率

- ZCR
  - Given a data set

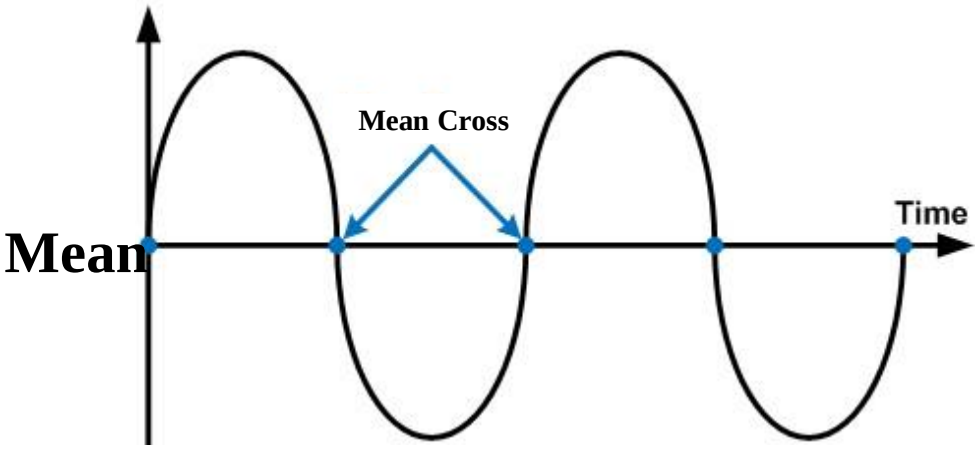
	t_1	t_2	t_3
X_i	1	3	5
Y_i	-3	1	-1
Z_i	4	1	-5

  - $ZCR\ X = 0$
  - $ZCR\ Y = 2$
  - $ZCR\ Z = 1$

Mean Crossing Rate (MCR)

- The *mean cross rate* is a measure that reflects how many times the sign of two adjacent values in *A* cross mean, which is defined as

$$MCR(A) = \frac{\sum_{i=2}^{n} |sign(a_i - \bar{a}) - sign(a_{i-1} - \bar{a})|}{2}$$



舉例說明：計算 MCR

- MCR
  - Given a data set

	t_1	t_2	t_3
X_i	1	3	5
Y_i	-3	1	-1
Z_i	4	1	-5

$\bar{X} = 3$
$\bar{Y} = -1$
$\bar{Z} = 0$

  - $MCR\ X = 0$
  - $MCR\ Y = 1$
  - $MCR\ Z = 1$

共變異數及相關係數 (Covariance and Correlation)

- Let $A = \langle a_1, a_2, ..., a_n \rangle$ and $B = \langle b_1, b_2, ..., b_n \rangle$ be the time series data of any two distinct axis of the accelerometer data in the segment, where a_i and b_i are the i-th samples in A and B, respectively.

- **Covariance**

$$Cov(A, B) = \frac{\sum_{i=i}^{n} |(a_i - \bar{a})(b_i - \bar{b})|}{n}$$

- **Correlation**

$$Corr(A, B) = \frac{Cov(A,B)}{\sigma_A \sigma_B}$$

舉例說明：計算共變異數

- Covariance
  - Given a data set

	t_1	t_2	t_3
X_i	1	3	5
Y_i	-3	1	-1
Z_i	4	1	-5

  - $cov(x,y) = \frac{(1-3)(-3-(-1))+(3-3)(1-(-1))+(5-3)(-1-(-1))}{2} = \frac{4+0+0}{2} = 2$
  - $cov(y,z) = \frac{(-3-(-1))(4-0)+(1-(-1))(1-0)+(-1-(-1))(-5-0)}{2} = \frac{-8+2}{2} = -3$
  - $cov(x,z) = \frac{(1-3)(4-0)+(3-3)(1-0)+(5-3)(-5-0)}{2} = \frac{-8+0+(-10)}{2} = -9$

Average Resultant Acceleration (ARA)

- ARA
  - Given a data set

$$\mathrm{ARA} = \frac{\sum_{i=1}^{N} \sqrt{x_i^2 + y_i^2 + z_i^2}}{N}$$

	t_1	t_2	t_3
X_i	1	3	5
Y_i	-3	1	-1
Z_i	4	1	-5

- $ARA = \frac{\sqrt{1+9+16}+\sqrt{9+1+1}+\sqrt{25+1+25}}{3} = \frac{\sqrt{26}+\sqrt{11}+\sqrt{51}}{3} \approx 5.18$

Magnitude

- Magnitude
  - Given a data set

$$\text{magnitude} = \sqrt{x^2 + y^2 + z^2}$$

	t_1	t_2	t_3
X_i	1	3	5
Y_i	-3	1	-1
Z_i	4	1	-5

Average Absolute Difference (AAD)

- AAD

$$AAD = \frac{\sum_{i=2}^{N} |x_i - x_{i-1}|}{N}$$

  - Given a data set

	t_1	t_2	t_3
X_i	1	3	5
Y_i	-3	1	-1
Z_i	4	1	-5

  - $AAD\ X = \frac{|3-1|+|5-3|}{3} = \frac{4}{3} = 1.3\overline{3}$
  - $AAD\ Y = \frac{|1-(-3)|+|(-1)-1|}{3} = \frac{6}{3} = 2$
  - $AAD\ Z = \frac{|1-4|+|-5-1|}{3} = \frac{9}{3} = 3$

Average Absolute Value (AAV)

$$AAV = \frac{\sum_{i=1}^{N} |x_i|}{N}$$

- AAV
  - Given a data set

	t_1	t_2	t_3
X_i	1	3	5
Y_i	-3	1	-1
Z_i	4	1	-5

- $AAV\ X = \frac{|1|+|3|+|5|}{3} = \frac{9}{3} = 3$
- $AAV\ Y = \frac{|-3|+|1|+|-1|}{3} = \frac{3+1+1}{3} = \frac{5}{3} = 1.6\overline{6}$
- $AAV\ Z = \frac{|4|+|1|+|-5|}{3} = \frac{4+1+5}{3} = \frac{10}{3} = 3.3\overline{3}$

Root Mean Square (RMS)

- RMS
  - Given a data set

$$RMS = \sqrt{\frac{\sum_{i=1}^{N}(x_i^2)}{N}}$$

	t_1	t_2	t_3
X_i	1	3	5
Y_i	-3	1	-1
Z_i	4	1	-5

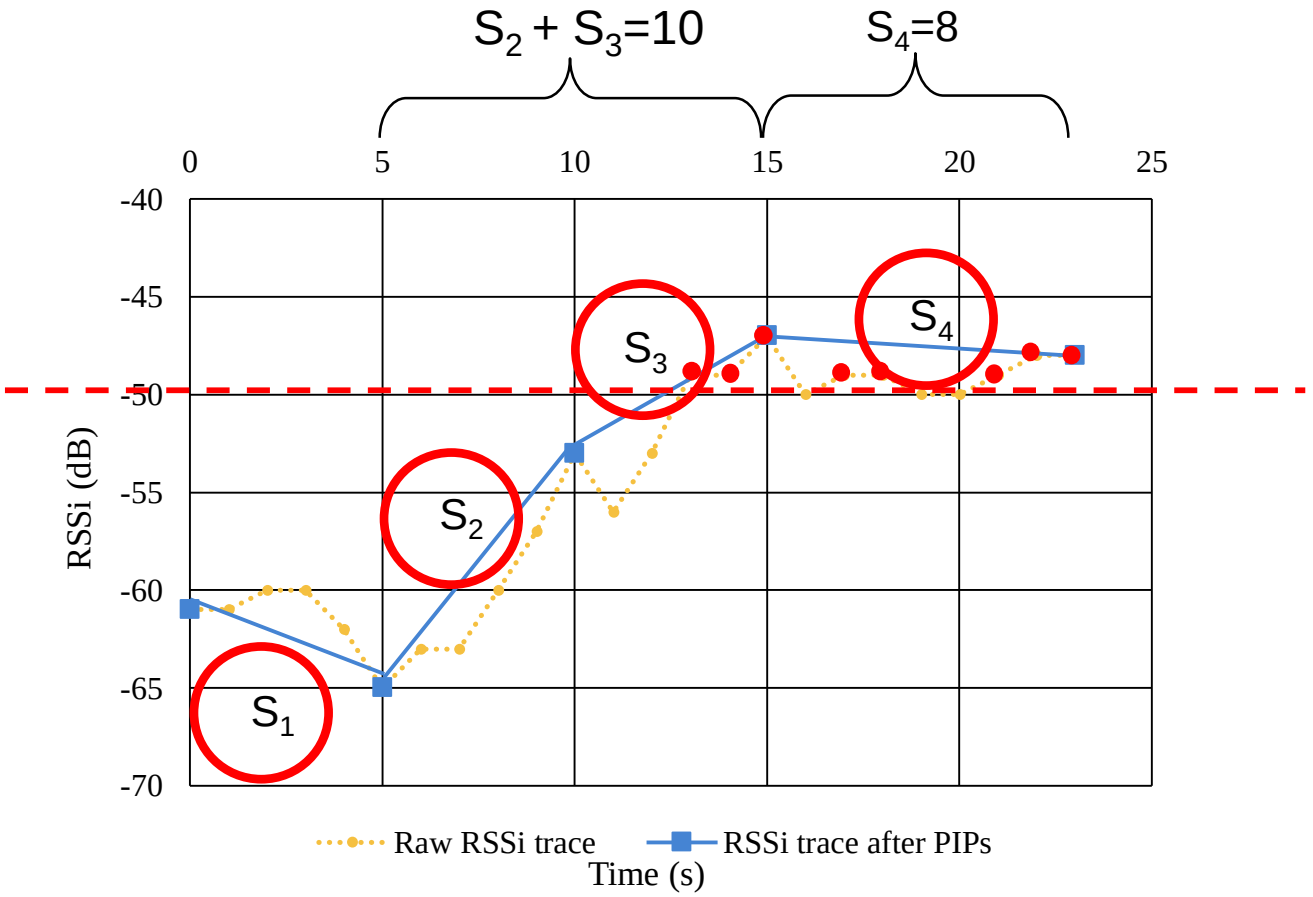
- $RMS\ X = \sqrt{\frac{1+9+25}{3}} = \sqrt{\frac{35}{3}} \approx 3.42$
- $RMS\ Y = \sqrt{\frac{9+1+1}{3}} = \sqrt{\frac{11}{3}} \approx 1.91$
- $RMS\ Z = \sqrt{\frac{16+1+25}{3}} = \sqrt{\frac{42}{3}} \approx 3.74$

教學單元
斜率特徵擷取方法

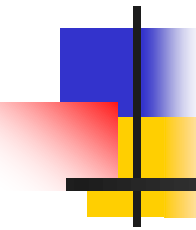
斜率特徵擷取
(Slope Feature Extraction)

Slope Features	Definition
R_{mean}	The mean of TS values
$M_{positive}$	The number of slopes which are positive
$M_{negative}$	The number of slopes which are negative
S_{max}	The maximum slope in the segment
S_{min}	The minimum slope in the segment
$D_{positive}$	The maximum period of the consecutive positive slope
$D_{negative}$	The maximum period of the consecutive negative slope
R_{diff}	The subtraction between maximum slope and minimum slope
R_{stable}	The times of TS values that are greater than a threshold

舉例說明：斜率特徵擷取



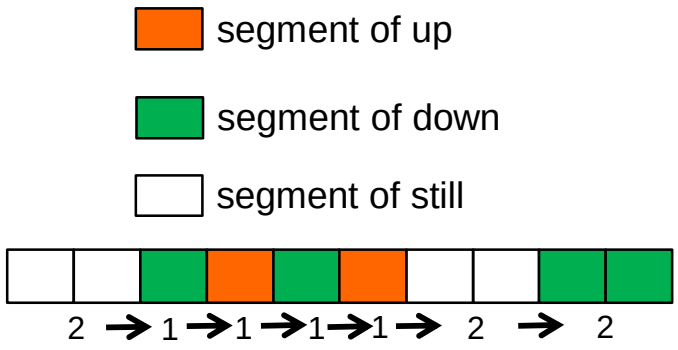
Features	Result
R_{mean}	-54.67
R_{stable}	8
R_{diff}	17
$M_{positive}$	2
$M_{negative}$	2
S_{max}	1.6
S_{min}	-0.8
$D_{positive}$	10
$D_{negative}$	8



教學單元
序列特徵特徵擷取方法

序列的特徵擷取
Feature Extraction of Sequences

Behavior Features Extraction		Total
C	Number of behavior changes in the window	3
B	Number of duration of each behavior in the window	8
D_{min}	Minimum duration of each behavior in the window	8
D_{max}	Maximum duration of each behavior in the window	8
D_{mean}	Mean duration of each behavior in the window	8



Features	Results
C	6
$B(up)$ / $B(down)$ / $B(still)$	2 / 3 / 2
$D(up)_{min}$ / $D(down)_{min}$ / $D(still)_{min}$	1 / 1 / 2
$D(up)_{max}$ / $D(down)_{max}$ / $D(still)_{max}$	1 / 2 / 2
$D(up)_{mean}$ / $D(down)_{mean}$ / $D(still)_{mean}$	1 / (4/3) / 2

教學單元
輪廓特徵擷取方法

Series Shapelet Mining

- Lexiang Ye and Eamonn Keogh, “Time Series Shapelets: A New Primitive for Data Mining, ” *in Proc. of ACM SIGKDD Int’l Conf. on Knowledge Discovery and Data Mining (KDD)*, pp. 947-956 , 2009.

  Google Citation: 634

- **Source**
  - http://dl.acm.org/citation.cfm?id=1557122

Basic Idea of Shaplet

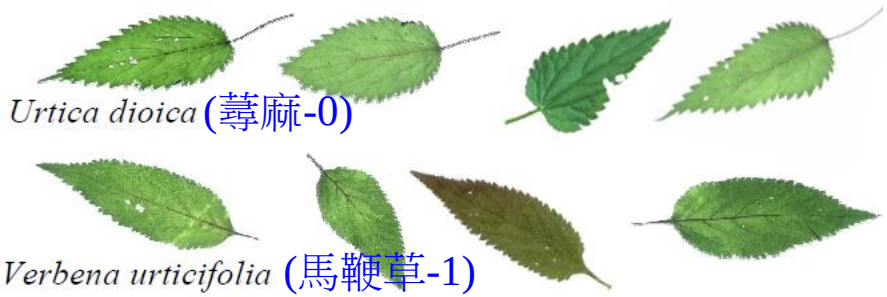


Figure 1: Samples of leaves from two species. Note that several leaves have the insect-bite damage

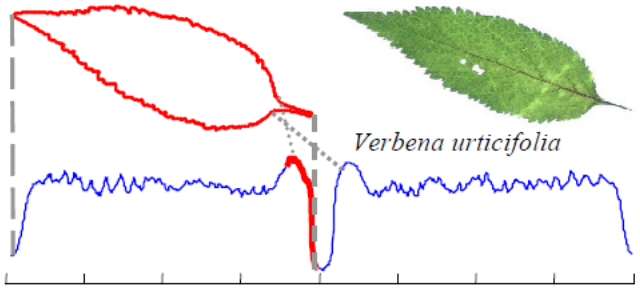


Figure 2: A shape can be converted into a one dimensional "time series" representation. The reason for the highlighted section of the time series will be made apparent shortly

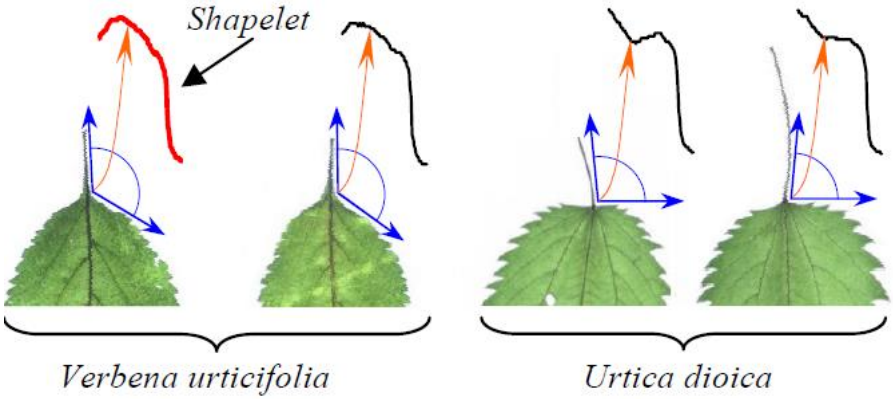


Figure 3: Here, the shapelet hinted at in Figure 2 (in both cases shown with a bold line), is the subsequence that best discriminates between the two classes

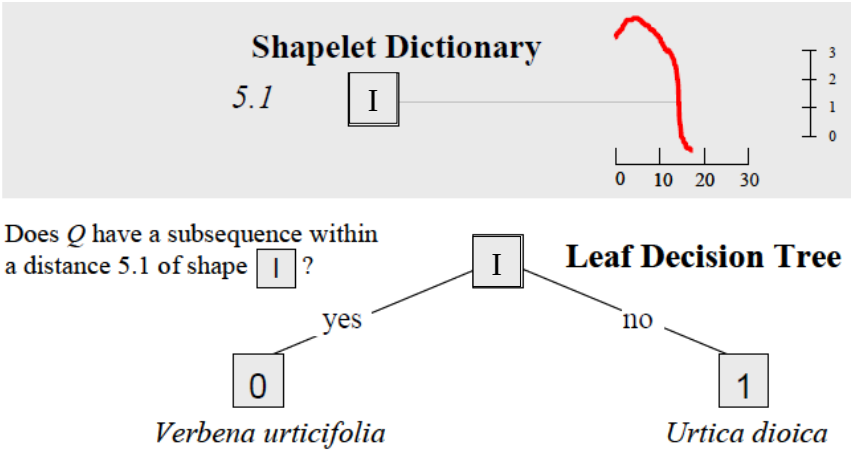
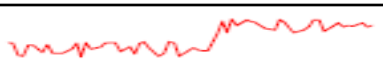


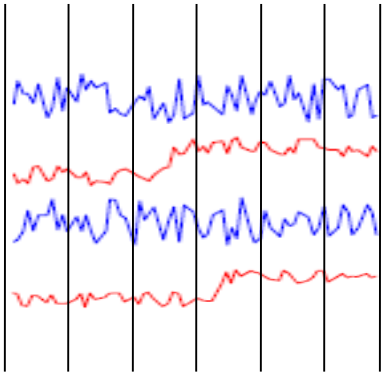
Figure 4: A decision-tree classifier for the leaf problem. The object to be classified has all of its subsequences compared to the shapelet, and if any subsequence is less than (the empirically determined value of) 5.1, it is classified as *Verbena urticifolia*

Finding Shapelets

Input Training Data

ID	Time Series	Class
T_1		A
T_2		B
T_3		A
T_4		B

Windowing



Candidate Pool

ID	Candidate	Class
S_1		A
S_2		A
…		B
S_n		B

Similarity Calculation

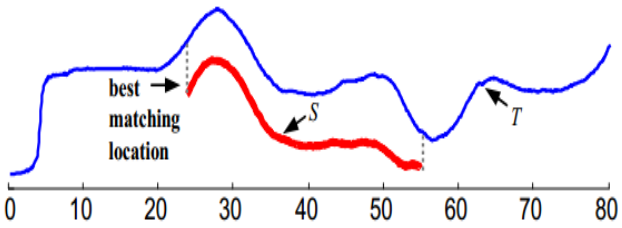


Figure 5: Illustration of best matching location in time series T for subsequence S

Distance Matrix

	T_1	T_2	T_3	T_4
S_1	0	7	2	8
S_2	0	6	4	3
…	…	…	…	…
S_n	9	1	12	0

Constructing Decision Tree with Information Gain (Cont. 1/2)

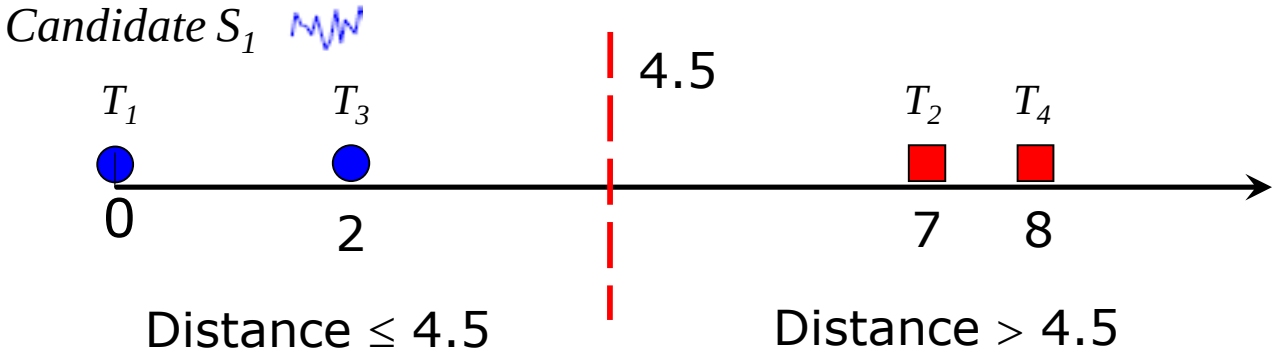
- *Information Gain* (S_i) = *Entropy*(D) – *Entropy*(S_i)

Input Training Data

ID	Time Series	Class
T_1		A
T_2		B
T_3		A
T_4		B

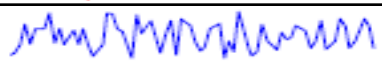
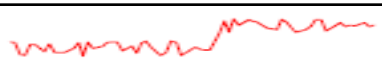
Distance Matrix

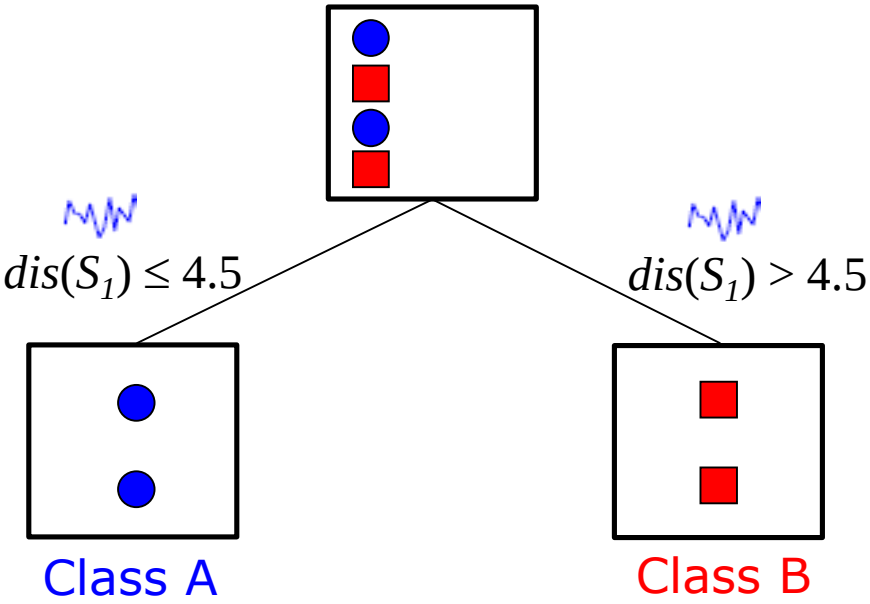
	T_1	T_2	T_3	T_4
S_1	0	7	2	8
S_2	0	6	4	3
…	…	…	…	…
S_n	9	1	12	0

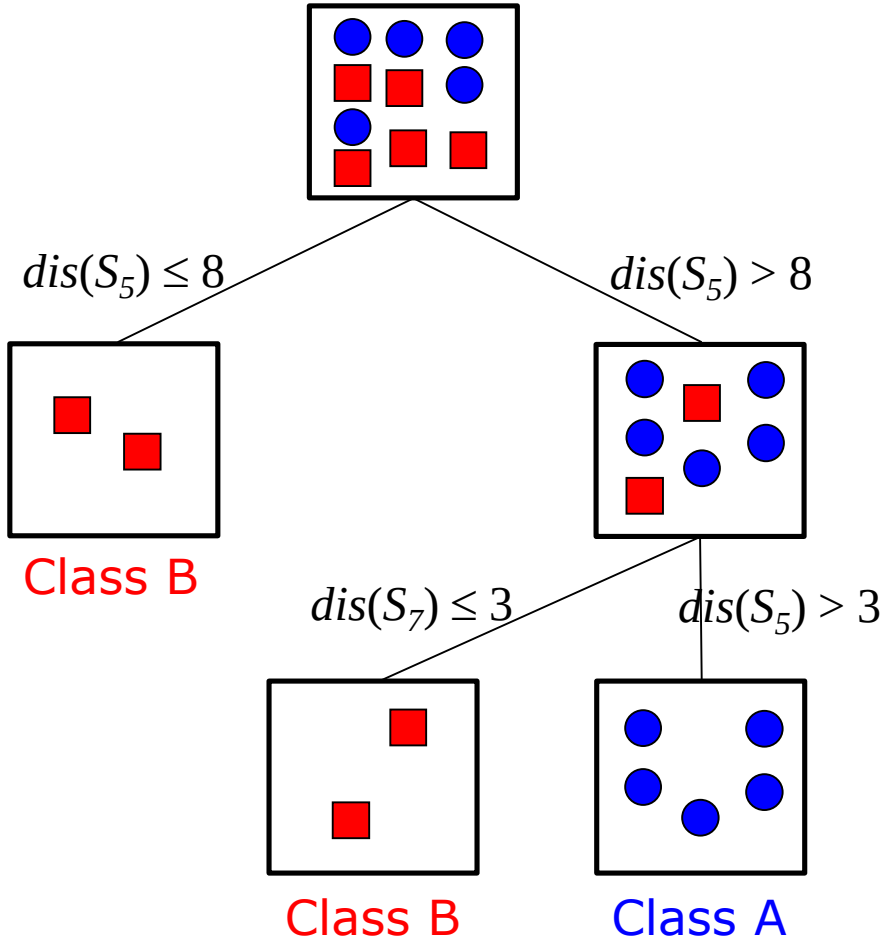


Constructing Decision Tree with Information Gain (Cont. 2/2)

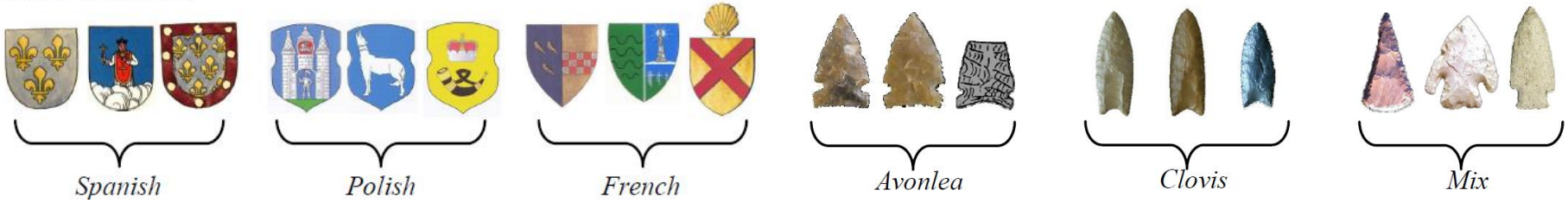
Input Training Data

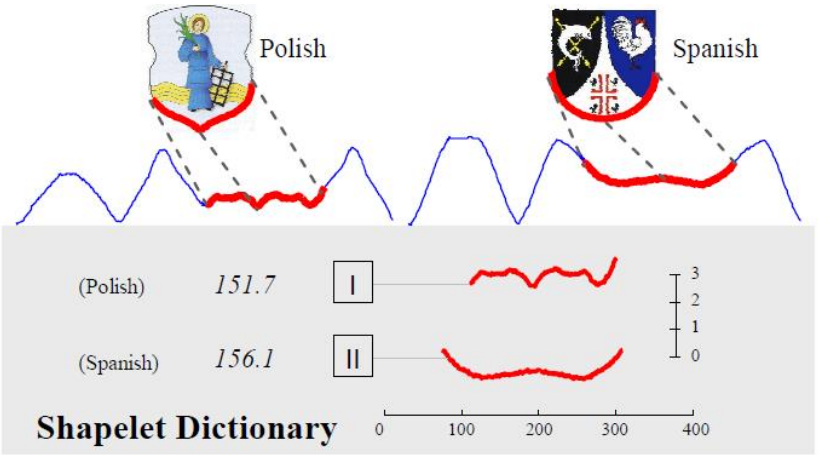
ID	Time Series	Class
T_1		A
T_2		B
T_3		A
T_4		B

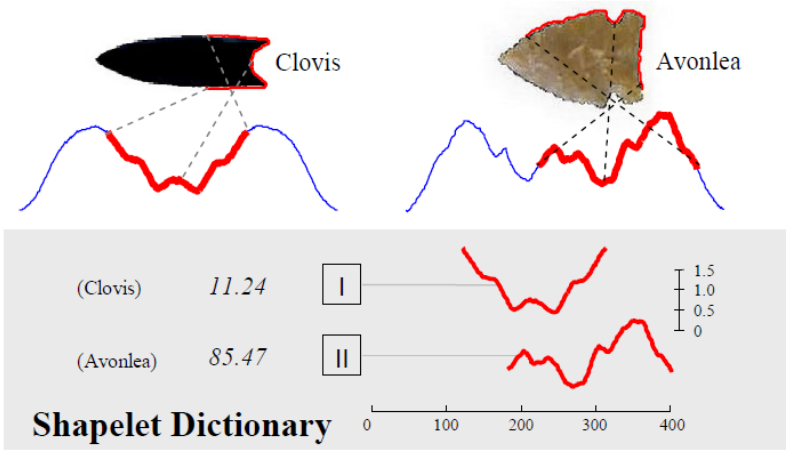


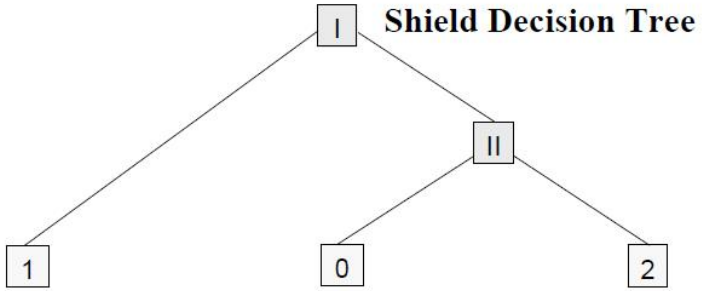


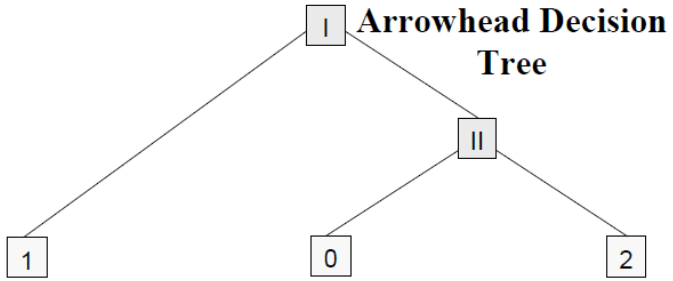
Results for Some Datasets



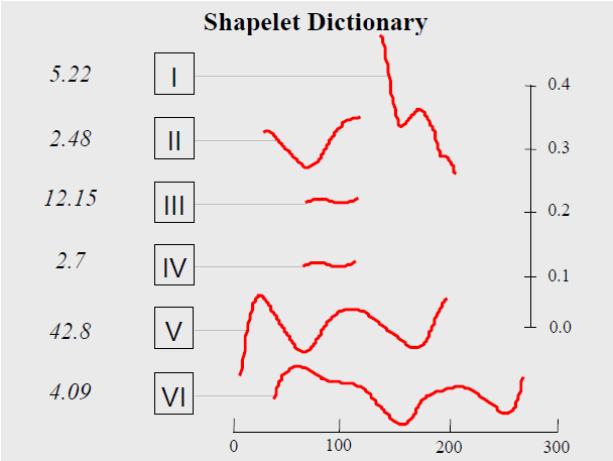






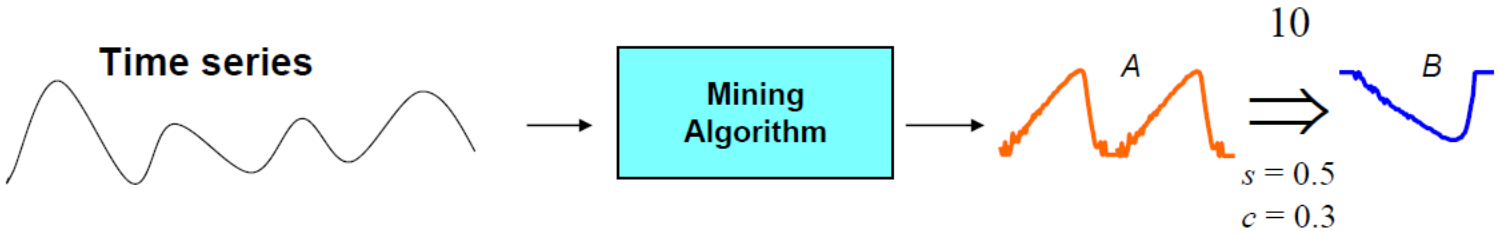


Shaplet-based Association Rules



ID	Time Series
T_1	
T_2	
T_3	
T_4	

ID	Shapelet
T_1	<IV, II, IV>
T_2	<VI, II, V>
T_3	<V, V, II, VI>
T_4	<II, II, I, V, I>



"If *A* occurs, then *B* occurs within time *10*",
where *A* and *B* are patterns and *10* is a time duration.

結論

- 時序資料特徵擷取應用
- 時序資料特徵擷取方法
- 使用軟體擷取時序資料特徵
- 使用軟體分析特徵重要程度