國立雲林科技大學工業工程與管理所
Graduate school of Industrial Engineering & Management,
National Yunlin University of Science & Technology

系統可靠度實驗室 System Reliability Lab.
http://campusweb.yuntech.edu.tw/~qre/index.htm

Asymmetrical Margin Approach to Surveillance of Nosocomial Infections Using Support Vector Classification

作　　者：Gilles Cohen and M´elanie Hilario
St´ephane Hugonnet and Hugo Sax

出　　處：IDAMP 2003
Intelligent Data Analysis in Medicine and Pharmacology

報 告 者：賴璟瑞
指導教授：童超塵　教 授

Contents

Introduction

- 院內感染監控(NIs)
  - have become a major concern not only in health care institutions but also among the general public.
  - Since 1994 the Geneva University Hospital has been undertaking yearly prevalence studies in order to monitor and detect NIs.

- 關鍵因素在預測和控制感染
  - 提供數據以評估問題的嚴重性
  - 發現疫情、辨識風險因素
  - 對高危險病人、加護病房為目標作控制
  - 評估預防方案

Introduction (con.)

- 最終目標➡減少監測感染的風險並提高病患的安全
- 在未來建立黃金目標 ；全院預測監控
  - 缺點:超出任何醫院可負擔的成本
  - 建議:選定特定的病房，如重症加護病房
  - 此替代經感染率調查，是有效的監控策略
- 確認風險因素之限制
  - 回溯性質、依賴容易得到之數據、感染率的偏誤
  - 無法查明的疫情、有限的能力

Introduction (con.)

- Support Vector Machine (SVM)
  - 提供良好的數據來衡量嚴重的問題、評估預防方案和幫助分配資源
- 手法上提供NIs的臨床快照(snapshot)活動在指定的一天，並提供有關的頻率和感染特性資訊
- 感染控制政策的功效
  - 可以容易地測量透過廣泛的重複調查
    - *French et al., 1983*

Data collection and preparation

- 院內住院調查
  - 至少48小時才能評估是否存在一個活躍的NIs

- 數據收集
  - medical records、 kardex (護理工作單)
  - Xray and microbiology reports
  - interviews with nurses and physicians in charge of the patient

Data collection and preparation

- 在院內感染調查那一天之前的六天，被Centres for Disease Control (CDC)記錄和辨識
- 收集變數存在相關的感染
  - 行政資訊、人口統計特徵、入院診斷
  - 合併症病患嚴重度分數、入院種類
    - 合併症；高血壓的合併症=>糖尿病
  - 各種危險因素的感染
    - 手術、加護病房逗留期間、設備、抗生素、免疫治療
  - clinical and paraclinical information

Data collection and preparation

- 全院監控再收集資訊就要花大約800小時，故利用資料探勘技術去收集2002年病患的研究，目的在及時發現NIs的病患
- 透過醫院專家過濾掉虛假的紀錄和不相關的多餘變數
  - 從688名病人記錄修改成683例
  - 從83種變數修改成49變數
- 而有些變數有缺失值，主要是由於測量錯誤或遺失
  - 缺失值假設為隨機缺失

Data collection and preparation

- 本研究利用類別條件均值class-conditional mean取代缺失值
- 資料探勘目的
  - 前處理的操作通常需要再作回顧性分析(retrospective analyses)
  - 而在數據收集沒有被專門設計

The imbalanced data problem

- 主要困難
  - 在固定數據中分佈存在高偏態
  - 683例中，只有75例感染而608沒有
- 應用中特別重要的議題；數據集不平衡
  - 目標；最大限度地認同少數類
    - For convenience we identify positive cases with the minority and negative cases the majority class
- Asymmetrical soft margin support vector machines
  - biasing the inductive process to boost sensitivity

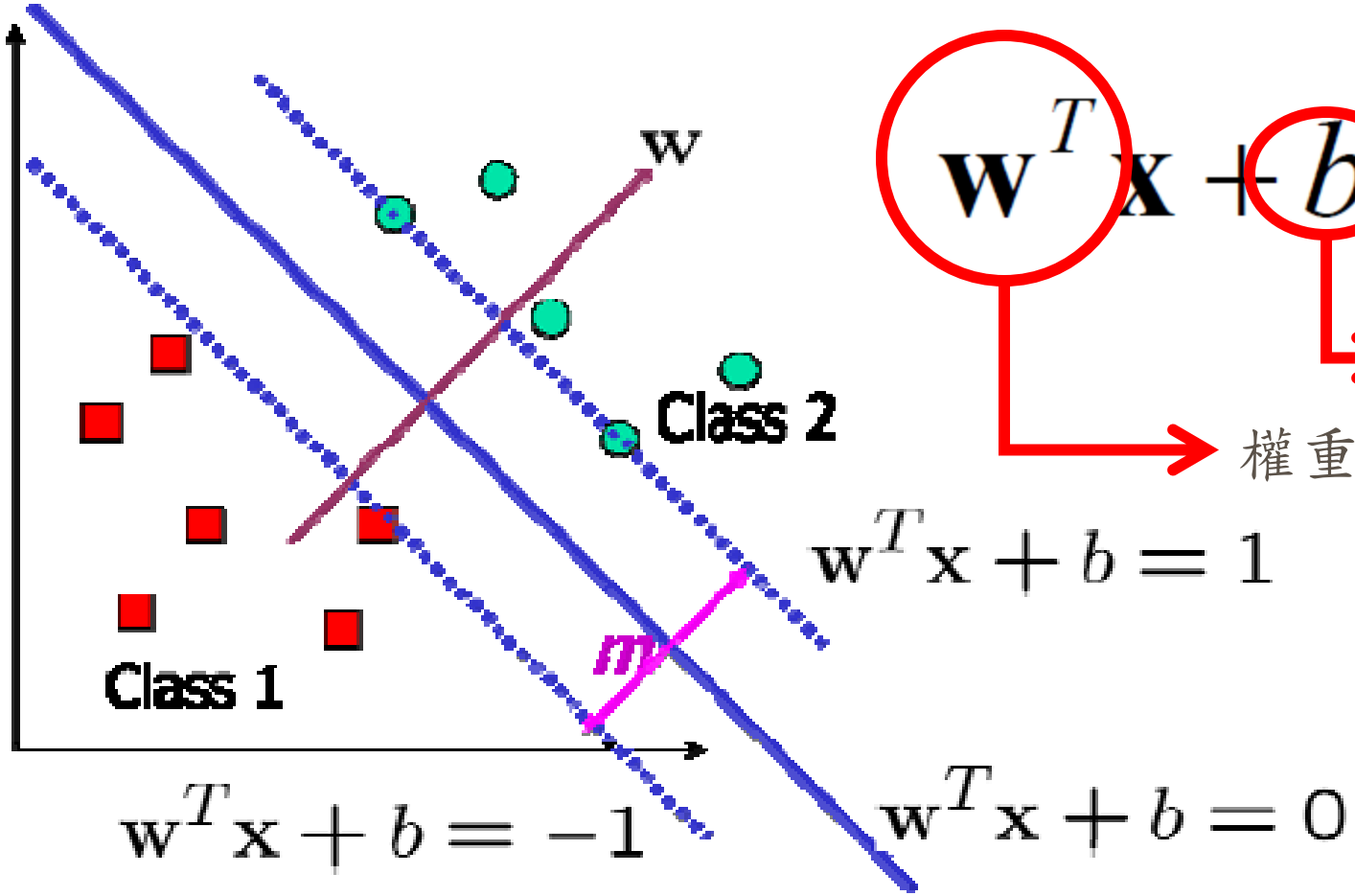
分類不平衡議題

Japkowicz, 2002	• 處理數據的方法，包含超採樣(oversampling)少數類別
Domingos, 1999	• 建構成本敏感度分類，將成本高的分配到少數類別的錯誤分群
Kubat and Matwin, 1997	• 以分層採樣去平衡在分群分配的訓練部分
Ali et al., 1997	• 以規劃為基礎的方法，嘗試以學習高信賴規則在少數分類

Support vector machine

- 支撐向量機(SVM)是由Vapnik 在1995 年和AT&T 實驗室團隊所提出的一個新方法，其主要的理論是來自統計學習理論中結構化風險最小誤差法(Structural Risk Minimization, SRM)。
  - SRM目的；希望分類器能在期望誤差中找到最小值。
- SVM目的
  - 尋求最小化預測結果之誤差上界(SRM)，而不是訓練誤差(ERM)的最小化。
  - Empirical Risk Minimization(ERM) 經驗風險最小化法則

Separating hyperplane



$$\mathbf{w}^T\mathbf{x}+b=0$$

偏差

權重向量

所以處理分類的函數為

$$f(\mathbf{x}) = sign\left(\sum_{k=1}^{s} 1\alpha_k y_k (\mathbf{x}_k \mathbf{x}) + b\right) \tag{3.9}$$

當 $f(x) > 0$ 時，代表該資料和標示為+1 的資料同一類；反之則是屬於另一類[21]。

$$\text{Minimize } \frac{1}{2}\|\mathbf{w}\|^2 \text{ subject to } y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 \ \forall i \tag{3.6}$$

利用 Lagrange 乘數方法((Lagrange Multiplier Method))可將(3.6)轉換

$$\text{Maximize } W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1, j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

$$\text{subject to } a_i \geq 0, \sum_{i=1}^{n} a_i y_i \mathbf{x}_i = 0 \tag{3.7}$$

接著藉由二次規畫(Quadratic Programming)方法可以求出滿足(3.7)的 α_i 集合，且 $\mathbf{w} = \sum_{i=1}^{n} 1\alpha_i y_i \mathbf{x}_i$，其中 α_i 不為 0 的 $\mathbf{x}_i$ 即稱為支持向量(support vector)。假設共有 s 個支持向量，因此改寫 $\mathbf{w} = \sum_{k=1}^{s} 1\alpha_k y_k \mathbf{x}_k$，可得到邊界方程式

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b = \sum_{k=1}^{s} 1\alpha_k y_k (\mathbf{x}_k \mathbf{x}) + b \tag{3.8}$$

Soft margin

在大部分的情形中，往往很難找到一條邊界可以完整的將資料群分開(圖 3.6)，因此必須容忍分類錯誤的行情發生。今ξ為分類錯誤時所需付出的代價(圖 3.7)，所以(3.5)可

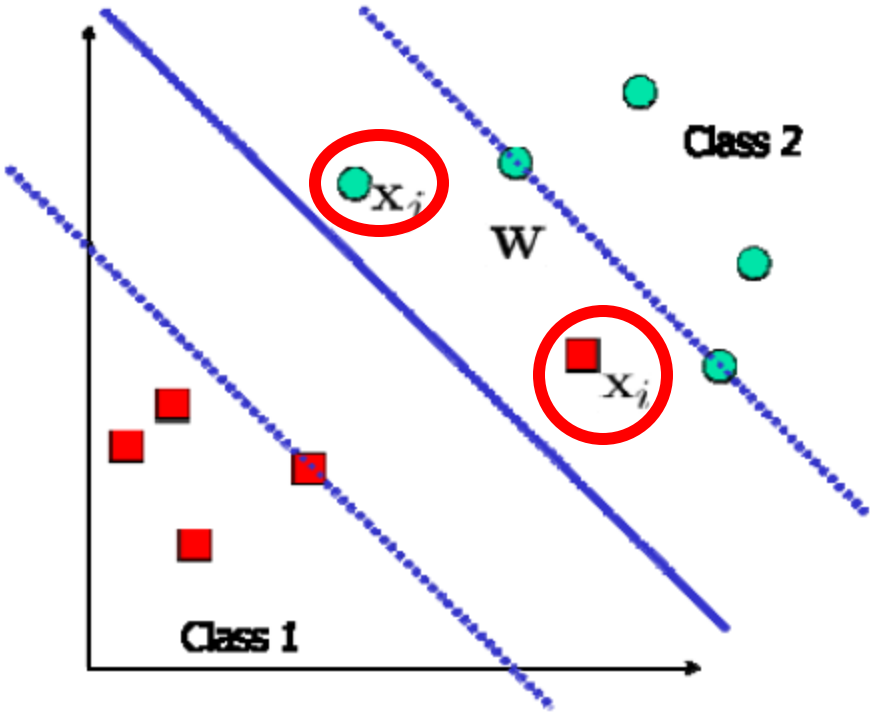


圖 3.6 分類錯誤

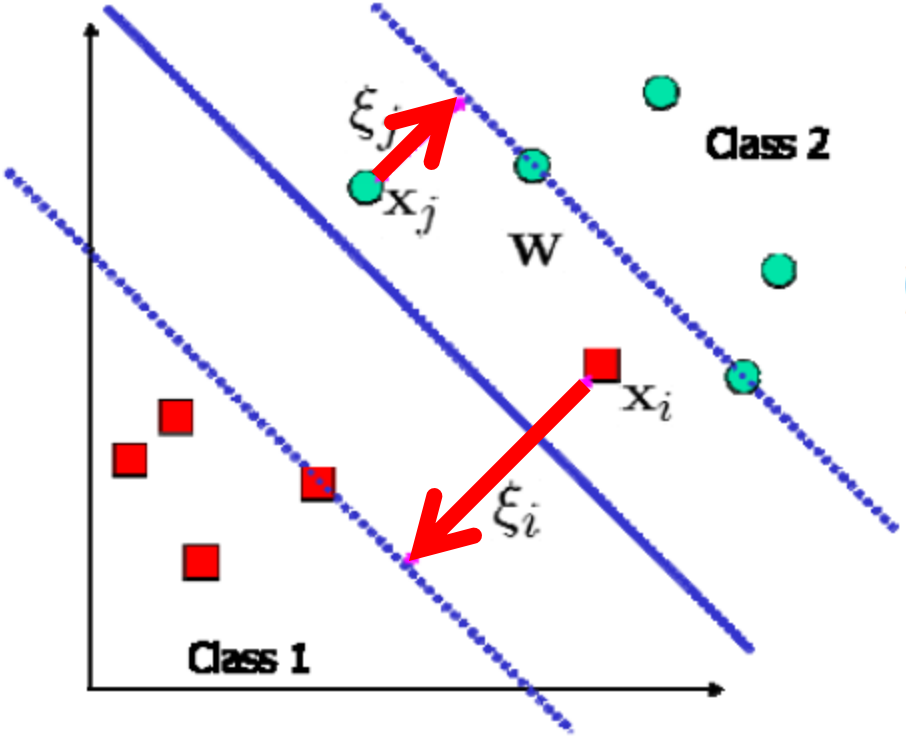


圖 3.7 加入ξ參數

目標同樣為最大化 m，所以解

$$\text{Minimize } \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \quad \text{subject to } y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1 - \xi_i,\ \xi_i \geq 0 \qquad (3.11)$$

利用 Lagrange 乘數方法將(3.11)轉換可得

$$\text{Maximize } W(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1,j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j$$

$$\text{subject to } C > \alpha_i \geq 0, \sum_{i=1}^{n} a_i y_i \mathbf{x}_i = 0 \qquad (3.12)$$

到此步驟可以發現，為了容忍分類錯誤的發生而加入錯誤代價，僅僅改變了 α_i 的上限而已。

因此，可以定義核心函數(kernel function)

$$k(\mathbf{x},\mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle \qquad (3.13)$$

如此一來，即可直接利用核心函數計算出資料在特徵空間中的內積值，不需將資料映射到特徵空間中後再進行內積運算，所以(3.12)可改寫為

$$\text{Maximize } W(\alpha)=\sum_{i=1}^{n}\alpha_i - \tfrac{1}{2}\sum_{i=1,j=1}^{n}\alpha_i\alpha_j y_i y_j k(\mathbf{x}_i,\mathbf{x}_j)$$

$$\text{subject to } C>\alpha_i\geq 0, \sum_{i=1}^{n} a_i y_i = 0 \tag{3.14}$$

同樣的，以 Quadratic Programming 方法找出滿足上式條件的α_i集合和$\mathbf{w}=\sum_{k=1}^{s}1\alpha_k y_k \mathbf{x}_k$，

可得到邊界方程式

$$f(\mathbf{x})=\mathbf{w}^T\mathbf{x}+b=\sum_{k=1}^{s}1\alpha_k y_k k(\mathbf{x}_k\mathbf{x})+b \tag{3.15}$$

最後，處理分類的函數為

$$f(\mathbf{x})=sign\left(\sum_{k=1}^{s}1\alpha_k y_k k(\mathbf{x}_k\mathbf{x})+b\right) \tag{3.16}$$

當$f(x)>0$時，代表該資料和標示為+1 的資料同一類；反之則是屬於另一類[21]。

Asymmetrical soft margin

- In order to adapt the SVM algorithm to these cases the basic idea is to introduce different error weights C_+ and C_- for the positive and the negative class, which results in a bias for larger multipliers α of the critical class.

$$
\begin{aligned}
minimize \quad & \frac{1}{2}||\mathrm{w}||^2 + C^- \sum_{i:y_i=-1}^{n} \xi_i^- + C^+ \sum_{i:y_i=+1}^{n} \xi_i^+ \\
s.t \quad & (\mathrm{w}.\mathrm{x}_i + b) \geq 1 - \xi_i^+, \qquad (9) \\
& (\mathrm{w}.\mathrm{x}_i + b) \leq -1 + \xi_i^-
\end{aligned}
$$

Performance metrics

- 在分類任務上，分類效能一般量化在預測的準確性，在測試集上誤判數據點的分數。
- 混亂矩陣

真實 \ 預測	C_1	C_2
C_1	*TP*	*FN*
C_2	*FP*	*TN*

Positive

Negative

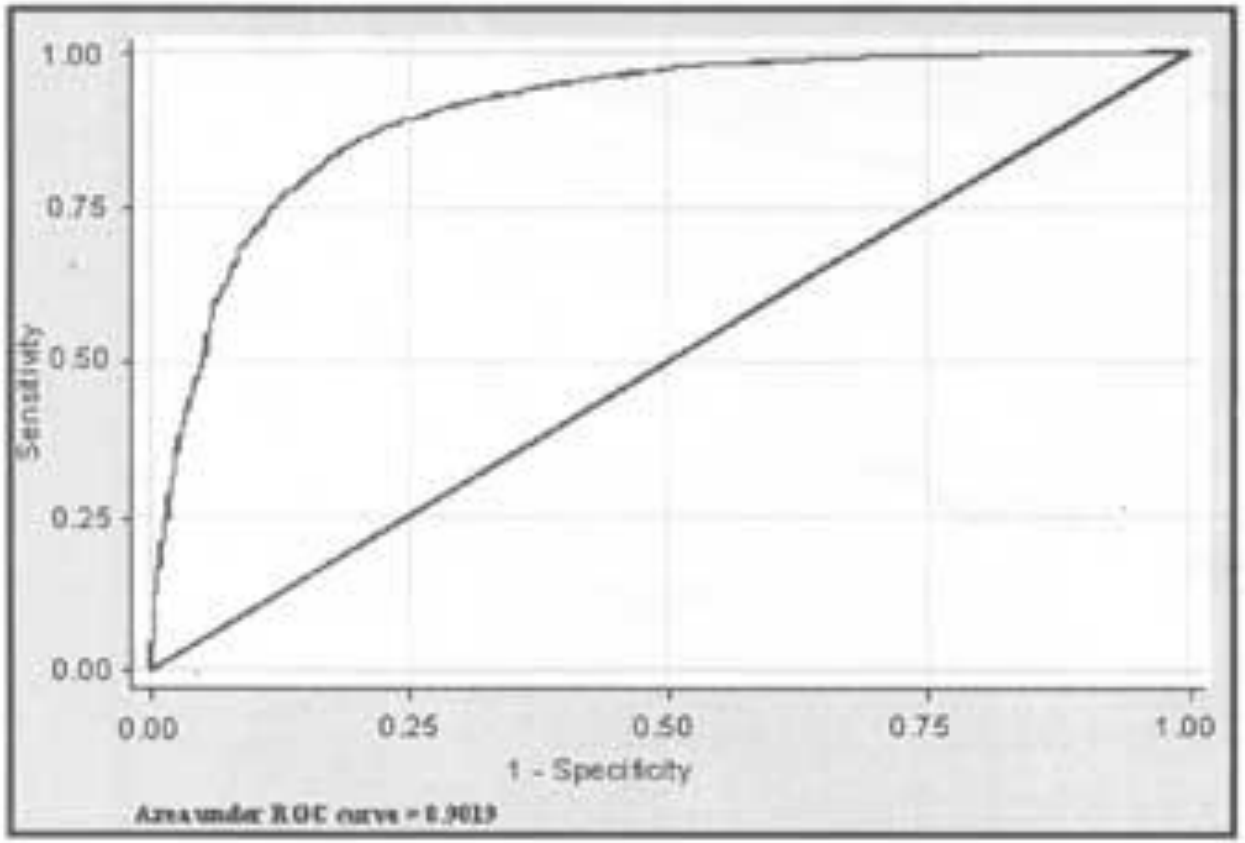
$$\text{敏感度} = \frac{TP}{TP + FN}$$

$$\text{特定性} = \frac{TN}{FP + TN}$$

$$\text{正確性} = \text{敏感度} + \text{特定性} = \frac{TP + TN}{TP + FN + FP + TN}$$

ROC curves

- 西元1980年代以來，流行採用ROC (receiver operating charact 面積」(area u 法的正確度。
- ROC 可能之截切點 感度及特異度 黃軸為（1- 牛 各座標點。



Results (sym. margin)

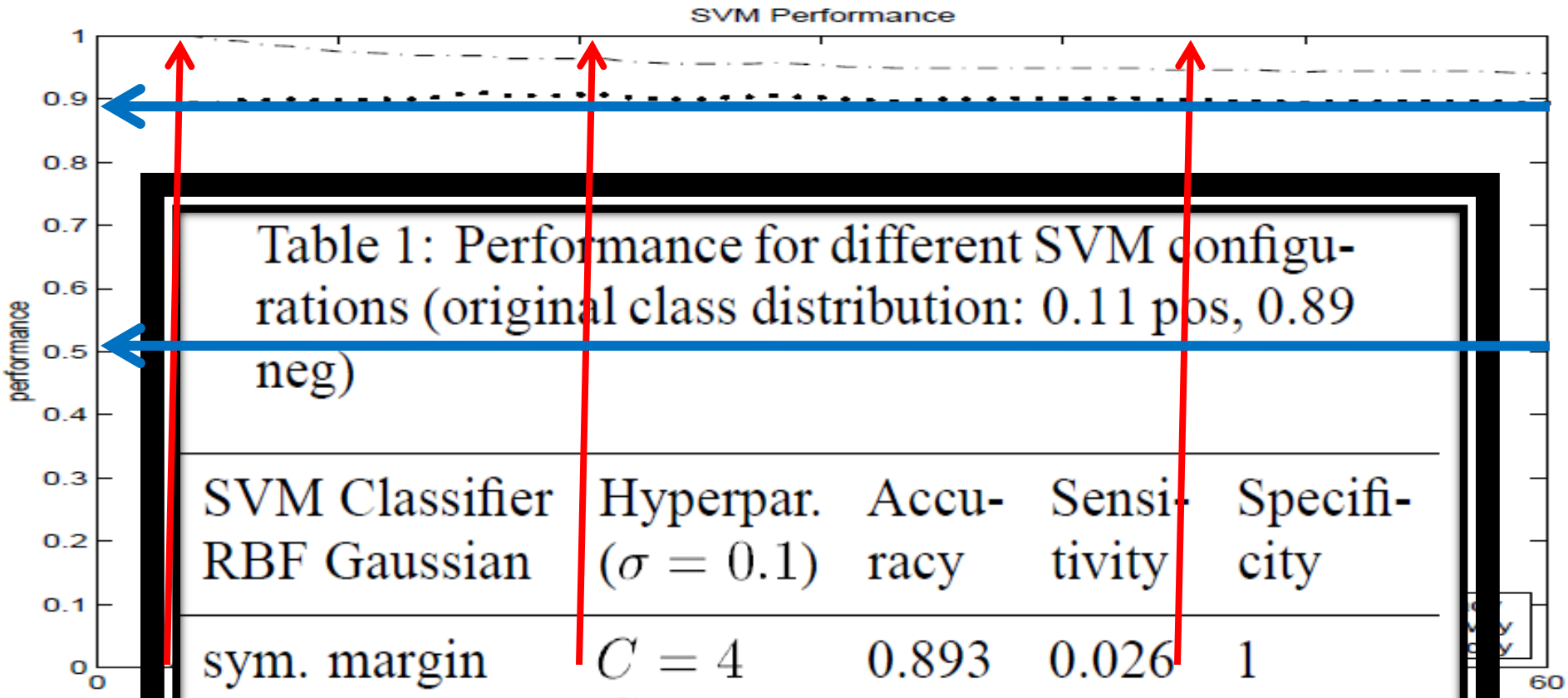


Table 1: Performance for different SVM configurations (original class distribution: 0.11 pos, 0.89 neg)

SVM Classifier RBF Gaussian	Hyperpar. ($\sigma = 0.1$)	Accuracy	Sensitivity	Specificity
sym. margin	$C = 4$	0.893	0.026	1
sym. margin	$C = 20$	0.906	0.44	0.964
sym. margin	$C = 45$	0.896	0.506	0.944

SVM classifier against different C values

Results (asym. margin)

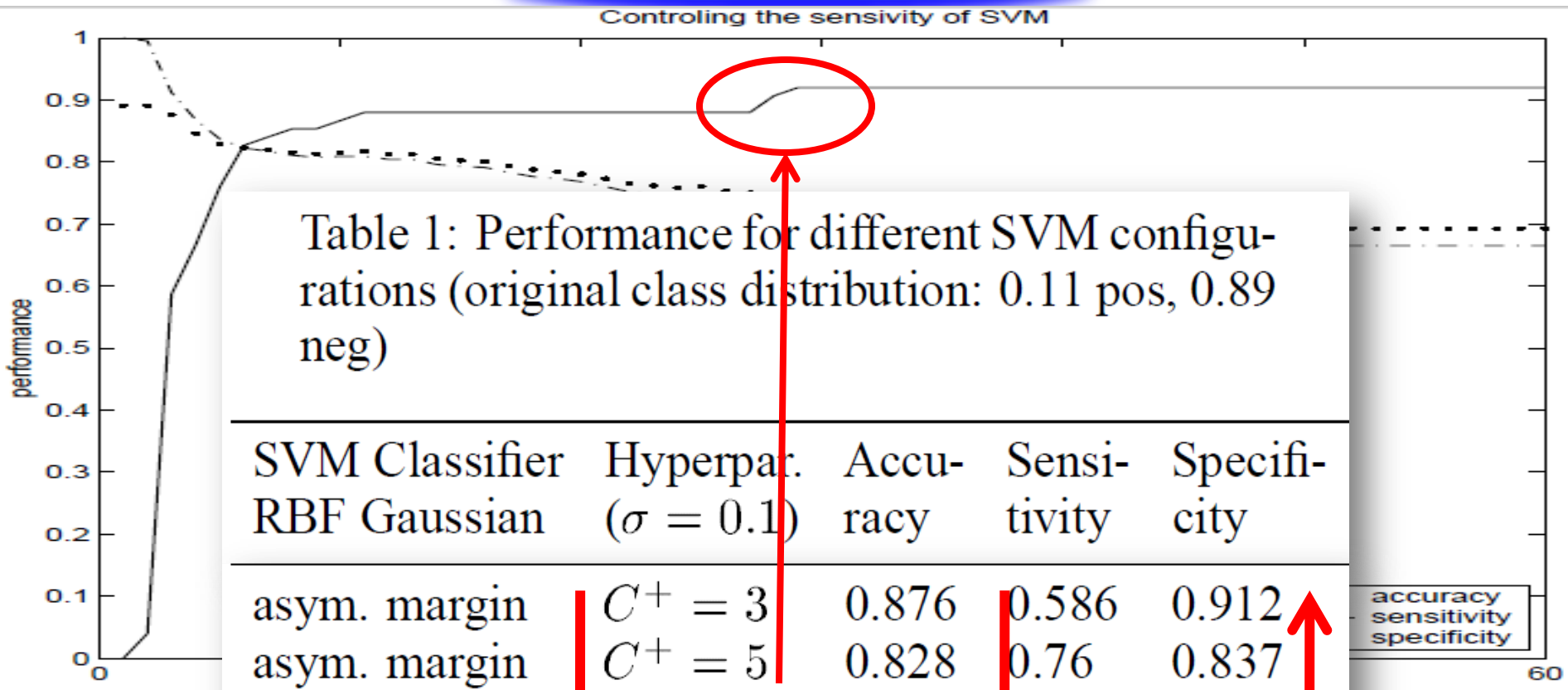


Table 1: Performance for different SVM configurations (original class distribution: 0.11 pos, 0.89 neg)

SVM Classifier RBF Gaussian	Hyperpar. ($\sigma = 0.1$)	Accuracy	Sensitivity	Specificity
asym. margin	$C^+ = 3$	0.876	0.586	0.912
asym. margin	$C^+ = 5$	0.828	0.76	0.837
asym. margin	$C^+ = 11$	0.816	0.88	0.809
asym. margin	$C^+ = 29$	0.744	0.92	0.722

mance of the asymmetrical-margin SVM classifier against different C^+ values

Results (compare)

- *Cohen et al., 2003*

Table 2: Over/undersampling via synthetic example generation (0.5 pos 0.5 neg). Bracketed figures are baseline sensitivity rates obtained prior to class balancing.

Classifier	Accu.	Sensitivity	Specif.	Method
IB1	0.84	0.56 [0.19]	0.88	KMU
NaiveBayes	0.75	0.87 [0.57]	0.74	HYB
C4.5	0.68	0.72 [0.28]	0.67	KMU
AdaBoost	0.75	0.84 [0.45]	0.74	KMU
SVM	0.75	0.83 [0.43]	0.74	KMU

貝氏判別
決策樹
推進法

K-means based undersampling

hybrid method

- The best sensitivity rate in these previous experiments was 0.87, attained by Naive Bayes coupled with hybrid over/undersampling via prototype generation.
- SVMs using asymmetrical margins and a C+ parameter of 29 perform remarkably better with a sensitivity rate of 0.92.

Results (ROC curve)

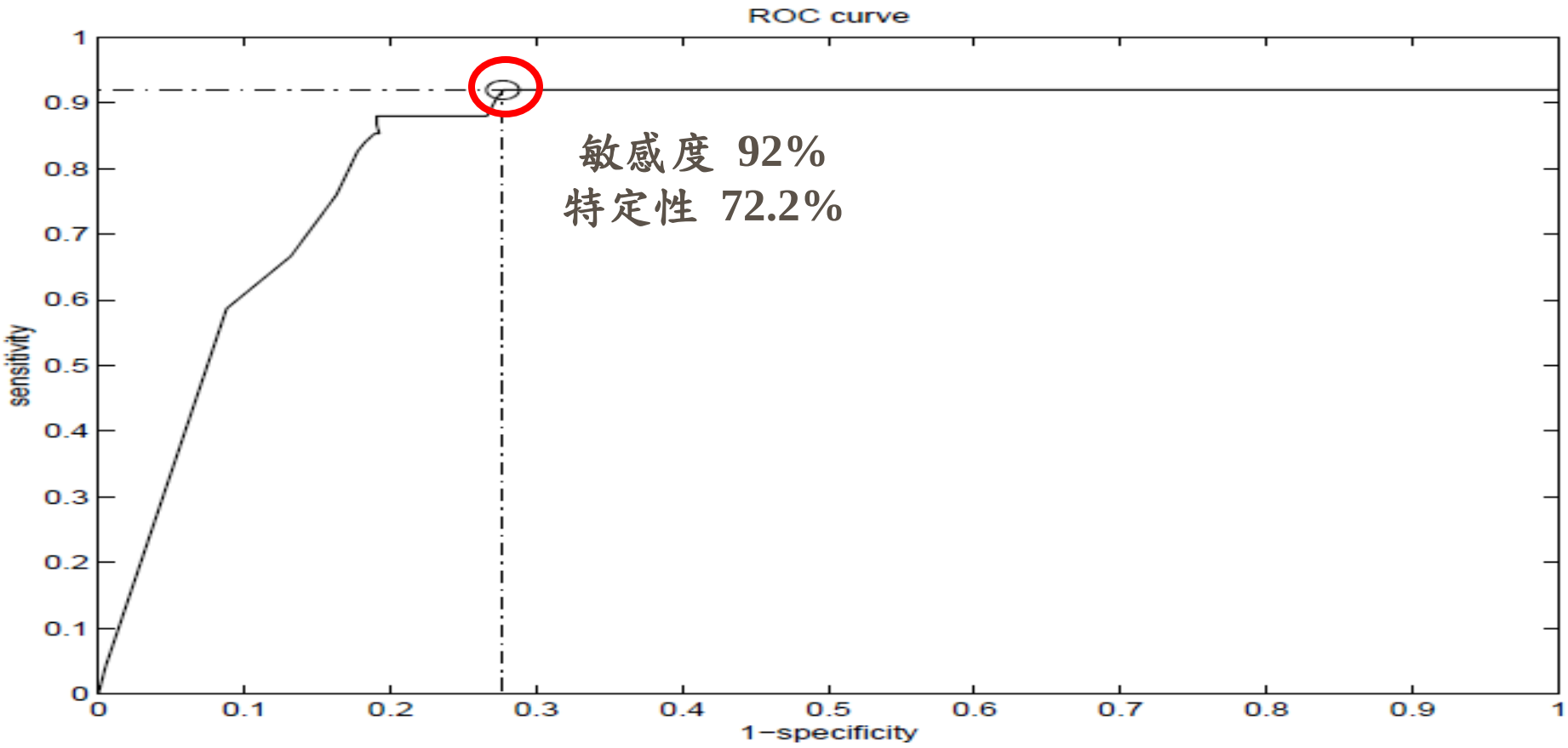


Figure 3: ROC Curve for SVM classifiers varying error weight values for the positive class C^+

Conclusion

- 根據病患去預測感染風險，其分析結果最主要障礙
  - 罕見的陽性問題
- 為了處理這問題，利用(asymmetrical soft margin)演算法，得到最大邊距在少分類的另一邊。
- Symmetrical soft margin SVMs中敏感度【2.6-50.6】%
- Asymmetrical soft margin SVMs中敏感度【58.6-92】%

類別不平衡問題

- 在分類問題中，類別不平衡問題(Class Imbalance Problems)會使分類器在訓練時產生偏誤，導致其對少數類別(Minority Class Examples)有相當低的預測正確率。

- 這個問題是因為不平衡的資料所造成，在此種型態資料中，一個類別的樣本數會遠超過其它類別的樣本數，使類別樣本的分佈呈現偏斜狀況(Skewed Class Distribution)，而相較於多數類別樣本，少數樣本通常是較感興趣的類別。
  - 例如，醫學診斷資料的少見疾病

- 當從不平衡資料萃取數據時，傳統的資料探勘方法會對多數類別樣本追求高的分類正確率，但對少數類別有極差的預測正確率，所以它們並不適合用來處理類別不平衡的資料。

- 利用對稱與不對稱SVMs進行比較，再以混亂矩陣分析其敏感度、特定性、正確性，最後再以ROC curve呈現圖表。

系統可靠度實驗室 System Reliability Lab.
http://campusweb.yuntech.edu.tw/~qre/index.htm

THE END