資料探勘於胸痛分類模型之建構

成員:黃咨翰,黃逸展,陳偉倫

執導教授:薛幼苓 Yuling hsueh

一、摘要

在一般看診或是急診時，醫師沒有足夠的時間去完整的閱讀病人的病歷資料。通常都是靠醫師詳細詢問病患的病史與觸診，才能正確地診斷，有時候還需要更進一步的儀器檢查，需要等到檢查報告完成，醫師才能根據檢查報告來診斷並告知病患檢查過後的結果，如此一來就得花費更多的時間。本系統利用病人之病歷資料，透過求出SU值及FAST演算法的方法來降低資料的複雜度，最後再把資料放入模型中判斷病人是否患有危險性高之胸痛。

二、系統架構

如圖 1，使用者在系統上輸入病人之病歷號碼，系統利用生成出來的預測模型進行預測，最後結果會顯示出是否為危險及四個主要症狀的數值，使用者可以在此留下建議並儲存此結果。

三、開發環境工具和Data set介紹

- 開發環境：eclipse, weka 3.6
- 開發語言：java, java GUI
- Server環境：Ubuntu 12.04.4 LTS (GNU/Linux 3.11.0-15-generic i686)
- Mysql server架設環境： mysql.cs.ccu.edu.tw(FreeBSD 6.2 Stable, MySQL版本:MySQL 5.1.36)
- Eclipse：JDBC Connector/J 5.1.36
- Data set 介紹:
  嘉義基督教醫院101、102年急診室病歷資料,資料筆數約10萬筆,胸痛患者約1萬5千筆，特徵為病歷資料上之各項檢查，項目共為183項

四、研究方法及步驟

實作出預測模型分為三個步驟：Data preprocessing, Feature selection, Building model

1. Data preprocessing : 將資料轉換成符合模組的格式，並將不合適的資料去除。

2. Feature selection 我們使用FAST演算法去除關聯性低的特徵，以高準確度。

3. Building model : 將第2步選出來的特徵，利用**J48 、 Naive bayes**、**RandomForest**等演算法建立模型模型來完成我們最後的預測。

FAST演算法 (A Fast Clustering-Based Feature Subset Selection Algorithm)

此演算法是一種以圖論為基礎的特徵分群演算法,分成去除**相關性低的特徵**及去除**同質性特徵**兩個部份

第一部份透過**特徵**對分類**類別**的SU值來做一次特徵的刪除，再把剩下的特徵取互相的SU值來建一棵**MST**。

第二部分則是針對每個邊做刪減，再把每個群組中關聯性最高的**特徵**留下，最後形成我們需要**特徵子集合**。

***SU值為WEKA里內建用來計算相關性的指標**
***原本183項經過刪減變為16項，包含呼吸速率，脈搏，高低血壓等等**

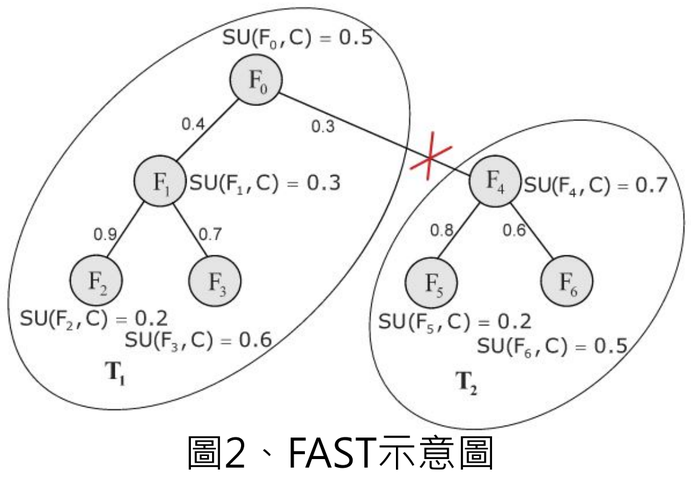

圖2、FAST示意圖

五、系統流程圖

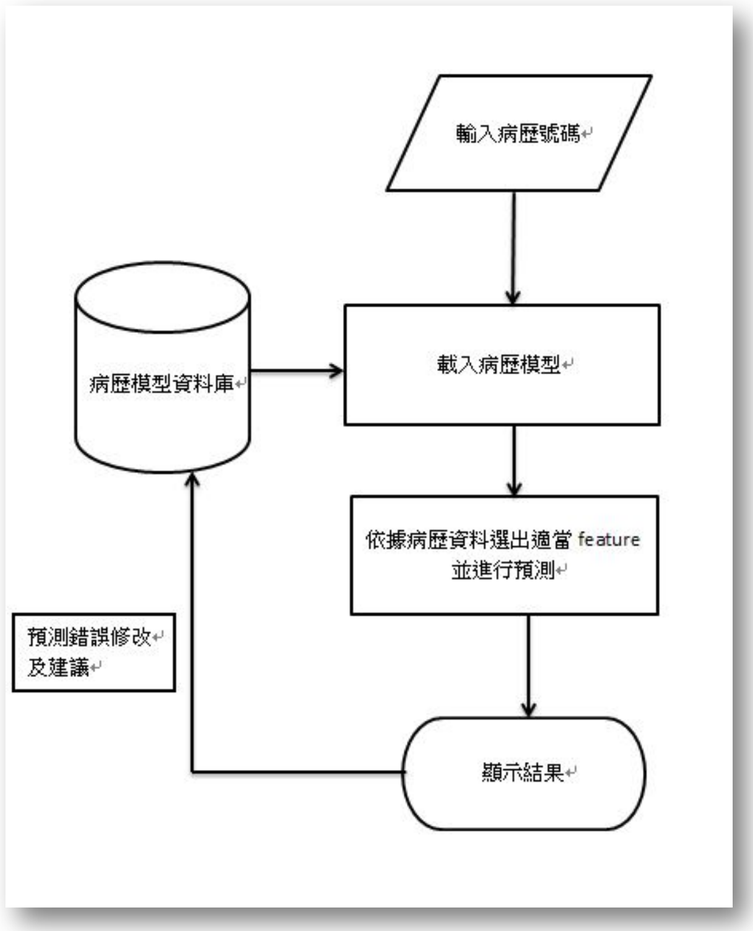

圖1 、系統流程圖

六、研究結果

模型準確度分析

Model:weka.classifiers.trees.RandomForest

=== Confusion Matrix ===

a	b	<-- classified as
982	9	a = danger
87	559	b = safe

TP Rate	FP Rate	Precision	Class
0.991	0.135	0.919	danger
0.865	0.009	0.984	safe

TP Rate=TP / (TP+FN)
FP Rate=FP / (FP + TN)
Precision=TP/ (TP+FP)
Accuracy=(TP+FP) / (TP+FP+FN+TN)=0.944
Sensitivity=TP/P=0.991
Specificity=TN / (FP + TN)=0.865

Correctly Classified Instances	1541	94.1356 %
Incorrectly Classified Instances	96	5.8644 %

系統介面成果

圖3為我們的開始畫面，可以在此處選取想要的演算法去建立模型，再輸入病歷號碼即可查詢。圖4則是顯示出預測結果，可以在此處看到病人的個人資料及是否具有危險性，並且我們還會擷取出4個主要影響預測的特徵，可以觀察到與正常值差距。

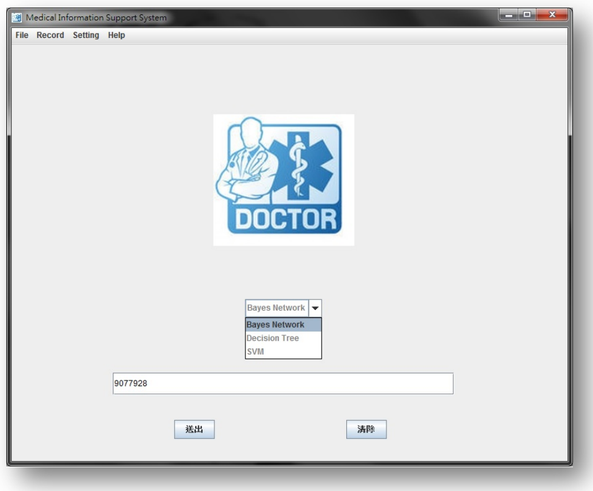

圖3、查詢畫面

圖4、結果顯示

參考文獻

1.A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data
Qinbao Song， Jingjie Ni， and Guangtao Wang