

運用AI進行自殺分析&預測

班級:四電三甲

製作學生:李福祥、陳家斌、黃千宸

指導老師:戴鴻傑 教授

組別:C10

摘要

National Kaohsiung University of Science and Technology

> 自殺防治是全世界重要的公共衛生與醫療議題,自殺不僅僅對個人造成傷害,對於家庭 也是一個不可抹滅的創傷。導致自殺的成因錯綜複雜,需考量多面向的資訊來達成有效的自 殺防治,例如:具有巨大經濟與社會負擔或非致命性自殺企圖的患者皆為自殺的高風險族群。

自殺問題的範圍及嚴重性已引起大量研究關注,隨著近年人工智慧技術的快速發展,本 專題以AI的方式分析臨床醫生評估報告、患者的自述結果以及電子健康記錄應用LSTM、CNN 深度學習技術、並結合開發C#去識別化系統,保護資料中患者的隱私, 將敏感資料去識別 化,供院方將相關資料釋出,便於專題研究使用。 取得已去識別化的資料集後,依據資料 集中病患是否再次自殺,將群體分為六個月內再自殺、一年內再自殺、一年後再自殺以及無 再自殺。隨後我們將根據此資料集分別進行資料前處理、特徵工程以及特徵還取並基於多種 不同的機器學習演算法進行模型的開發並評估其效能,輔助醫事人員評估病惠的自殺風險, 落實自殺防治並降低再自殺人數比例。

本專題實作過程分為三部份(如圖1) ,去識別化系統開發、資料前處理、機器學習。



▲圖1

資料前處理

(一)、資料內容

院方釋出去識別化資料後,資料項目多達數百項, 透過資料前處理刪除非必要欄位、抽取與自殺風險有 關的變項。

(二)、資料統計

合作醫院釋出資料中,病患可能多次住院及

看診,數量統計結果如下表、圖3。

7H 47	** *	. 190 0 1 1	- D - Je x	1 10	1240	
	第名器		- 李敦			
住院主持	5130					
在院明》	521403					
門蒙主		102092				
門會明命	82199					
		再自撤	<=30夫	<=180	<=365	>365
				夫	夫	夫
季業	2007	583	53	124	106	300

(三)、資料整理

1.年齡,我們將患者年齡以10歲為間隔分成十一群,避免 年齡特徵過於分散,降低模型準確度。

2. 住院及門診,考量到季節會影響人類的心理因素,我們以三個月做為分類,結合醫師診斷碼,用四位元編碼表示90天數內有無包含該類診斷碼。如ICD10 (F32 憂鬱症)患者90天內住院2次、90天後門診3次,F32欄編碼為1001。
3. 國際疾病分類(ICD 10編码),原始資料中,ICD10編码為患者疾病細項,將同為呼吸系統疾病詳細區分為 ICD10 (J06.9)急性上呼吸道感染、ICD10 (J96.91)呼吸衰竭併缺氧……等,將此細項列為機器學習特徵顯得複雜,因此,我們將細項統計後,將其編碼分為二十二大類,並與院方討論後,僅將精神相關疾病的代碼加以細分,如表,篩選

要特徵。	自殺道報机計			
	2000 2000 1000			
圖3▶				
7-7	0 011			

文献在多数	英文製在本籍	THE
9.60	sex	第:1
	700	k:1
		紀錄講報時間時的年齡、並以12歲存間限分為一群、如:
		1-10A:1
Y-60	1800	11-29A; : 2
THE	100	21-39 AL : 3
		and the same of th
		101点上:11
10円 元数(別え内)	ada 900	連報音数 - 人居音数
3.院北東(大松別大)	ada other	
可协议数(预失性)	opd_500	通報日期 - 門珍日期
5分之数(大价90天)	opd other	
		根據[CDO部配住院及門侍之數乘通行編碼·如謂:
		1銀的集為6001(理補定4格)
-2386		3101各代表D、E、F、G那筆資料內有包含此類的疾病代號
-23M	leategory-23category	0 E F G H
		adm_900 adm_mberspt_900 rept_mber Emirgory
		1 1 4 2 8
		回接1CEG新配在股车門作力數值進行構構、如果:
	F01-F90	F3256 Vo (6, 8-0100) (8 HE JF 6-16.)
特什么名分别		DISSAGAD - E - F - CHERTHARA - DAMMAGAR
RITION OF STREET	141-139	0 6 7 6 000
		ule 90 als obried 90 and obe 70
		1 1 4 2 10
		照據ATCrode搭配住院及門珍文數來進行編碼:
製品 (項目) 代統	DESC_NOCORDER_CODE)	III III 本代表D·E·F·G斯蘭資料內有包含此類的ITCode
		4-W:
		PSRT, NONS 集長100(領領北三馬)
	PART_NO	第一端代表重大编码(代码801)
将珠身会		第二烯代表热收入产(代明803)
		第三碼代表持有殘障手術(第一碼為英文字母且第三碼為3·es:A03·B03)
特种科技院文獻	ada PSY	一条内包的现在分别只在精神开的比赛
精神科門發皮數	cod. PSY	一条内内的经验的信用系统设计的企业
医沙黎伊科门沙伐斯	opd DI	一年內門你時看你部門為急你醫學科的文教
多属科門抄收數	opd_PN	一年內門亦時看功部門為疼痛料的次數
特种科技健康	regist_count	一年內就珍祝ங春"保珍"
免经双金融种料次数	ipd_count	一年內急珍照會維持科文數
注股照會條件料次數	er_count	一年內住稅照會精神科次數
		一次自務:0
	toka t	90人内在市报:1
电视	tabel	90 K (株 A (8 科) 2

精神相關疾病代碼	疾病名稱
F00	阿茲海默病 (<u>失智症</u>)
F01	血管性痴呆
F02	分類於他處的其他疾病引起的 <u>痴呆</u>
F03	未特指的 <u>痴呆</u>
F04	器質性遺忘綜合症,非由酒精 和其他精神活性物質所致
F05	讀妄,非由酒精和其他精神活性物質所致
F06	由於腦損害和機能障礙及軀體 疾病引起的其他精神障礙
F07	由於腦部疾病、損害和功能障 礙引起的人格和 <u>行為障礙</u>
F09	未特指的器質性或微狀性精神 障礙

去識別化系統開發

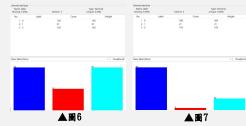
專題擬運用合作醫院院區內 2005 至 2020年間所有急診就診患者的相關數據與文字敘述的轉診單做為後續自殺風險評估因子模型的發展資料。然而,此資料包含病患的隱私資訊,在資料可以釋出前,須經過去識別化的處理方可做為研究使用。因此我們應用已發展的去識別化技術結合 C# 程式來開發非結構化文本資料的去識別化系統(如圖4,圖5為去識別化系統模型的蛇分),提供合作醫院的大數據中心使用。



持徵提取、機器學習

1. 利用weka內建的資料處理,使不平衡的數據依照Oversampling及Undersampling兩種方式,使得數據分布可以平衡,讓數據可以提升。由於需要利用虛擬資料來增加原有數據,一開始先以分成8:2的訓練及測試資料,再以訓練資料依照2:1:2的方式來平衡各類的數據,如關的所示。

2. 再以此訓練資料來 利 用 支 持 向 量 機 (Support Vector machine)、隨機森林 (RandomForest)、 決 策 樹 (Decision Tree)、貝式分類 (Naïve Bayes)四種 機器學習來建模。測 試資料如圖7所示。



3. 最後測試資料來對分別建好的模型做評分,評分數據及混淆矩陣如下圖所示:

1. 支持向量機(如圖8)

正確率: 49. 4141%

召回率: 0. 494

F-score: 0. 545

=== Confusion Matrix ===

a b c <-- classified as

231 55 100 | a = 0

5 9 7 | b = 1

50 20 35 | c = 2

▲圖8

=== Confusion Matrix ===

2.隨機森林(如圖9)

正確率: 59.9609% 召回率: 0.6 F-score: 0.623 a b c <-- classified as 250 5 131 | a = 0 8 1 12 | b = 1 49 0 56 | c = 2

▲圖9

=== Confusion Matrix ===

3.決策樹(如圖10)

正確率:53.7109% 召回率:0.537 F-score:0.583 a b c <-- classified as 231 55 100 | a = 0 5 9 7 | b = 1 50 20 35 | c = 2

▲圖10

=== Confusion Matrix ===

4.貝式分類(如圖11)

正確率: 66.7969 召回率: 0.668 F-score: 0.661 a b c <-- classified as 315 13 58 | a = 0 11 1 9 | b = 1 73 6 26 | c = 2

▲圖11

結論

精神相關疾病為重

由於資料取得不易,只能從醫院單方面取得資料,這種情況下,資料分布會顯得十分不平均,進而導致測試分數大大降低,以測試資料分布來看,class b(即為<=90天內在自殺者)為最少類,且在訓練過程中也時常遇到class b預測值為0的情況。以四種機器學習得出的結論,雖以貝式分類平均召回率最高,但若以各類拆分開來看,決策樹判斷class b 為0.429最高分,其次為支持向量機 0.381分。