Breast Cancer Classification Using Machine Learning

Ⅰ.INTRODUCTION

前陣子，因為家人有罹患乳癌，所以正好找了和這個議題相關的乳癌數據，來分析究竟乳癌的產生，究竟和何種因素有關。

有了機器學習，我們能更快的處理和分析數據的結構和資訊，利用機器學習可以解決不同的分類問題，在乳癌的資訊當中，可以快速且有效的分類出那些大小或形狀的乳癌腫瘤是良性或是惡性。

在這份報告當中我們首先要對資料進行前置處理，資料前處理是機器學習必要且耗時的一項步驟，再來是建模和測試，最後比較出何種的分類器對資料有比較好的預測能力。

Ⅱ.DATA PREPROCESSING

首先，要把整個乳癌數據組分成 Training set 和 Testing set，再來，如若有遺漏值，分別對遺漏值做處理，接著，Feature selection 來檢視所有特徵的重要性，或降維做 Feature Extraction 來增加運算效率，最後，把數據丟入分類器來比較準確值。

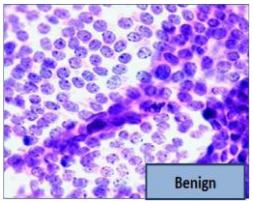
Ⅲ.DATA SETS

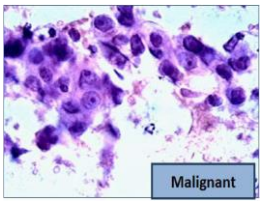
我所使用的是 UCI (Diagnostic) Breast Cancer Wisconsin 的數據集，它一共有四種類型:

Name of the dataset	Number of instances	Number of attributes
Breast Cancer	286	9
Breast Cancer Wisconsin(original)	699	10
Breast Cancer Wisconsin(Prognostic)	198	34
Breast Cancer Wisconsin(Diagnostic)	569	32

Diagnostic dataset 與良性或惡性有關

Prognostic dataset 與復發不復發有關





Ⅳ.TRAINING MODEL CLASSIFIER

最後用不同的模型來檢測何種為最理想的分類器，常用到的模型有:

- Logistic Regression
- Naive Bayes Classifier
- Support Vector Machine
- Decision Tree
- Random Forest
- Baggings
- Multi Class Classifier

Ⅴ.REFERENCE

[1]. S. Sathya, Sundeep Joshi, S. Padmavathi, "Classification of Breast Cancer Dataset by Different Classification Algorithms", IEEE, 2017

[1]. Meriem Amrane ,Saliha Oukid ,Ikram Gagaoua, Tolga Ensarİ, "Breast Cancer Classification Using Machine Learning", IEEE, 2018