# Random-projection ensemble classification

出處:

探討: 我們介紹了一種非常通用的高維分類方法，該方法基於將任意基礎分類器應用於特徵向量在低維空間的隨機投影的結果進行仔細組合。在我們詳細研究的一種特殊情況下，隨機投影被分成不相干的組，在每組中我們選擇產生最小的測試誤差估計的投影。然後，我們的隨機投影集合分類器將基礎分類器在所選投影上的應用結果匯總起來，用數據驅動的投票閾值來決定最終的分配。我們的理論結果闡明了增加投影數量對性能的影響。此外，在充分降維假設所隱含的邊界條件下，我們表明隨機投影集合分類器的測試超額風險可以由不依賴於原始數據維度的項和隨著投影數量增加而變得可以忽略的項來控制。通過廣泛的模擬研究，我們還將該分類器與其他幾個流行的高維分類器進行了實證比較，結果顯示其具有出色的有限樣本性能。

技術: 我們的理論結果分為三個部分。在第一部分中，我們考慮了一個通用的基礎分類器和一個生成隨機投影的通用方法，並量化了隨機投影集合分類器的測試誤差和其無限模擬對應物之間隨著投影數量增加而產生的差異。然後，我們考慮從不重疊的組中選擇隨機投影，最初根據Haar度量法抽取這些投影，然後在每組中保留測試誤差估計值最小的投影。在一個被廣泛使用的充分降維假設所暗示的條件下（Li，1991；Cook，1998；Lee等人，2013），我們可以控制隨機投影分類器的測試誤差和Bayes風險之間的差異，作為一個取決於基於投影數據的基礎分類器的性能和我們估計測試誤差的方法的函數，以及一個隨著投影數量增加而變得可以忽略的項。我們理論的最後部分給出了這些項中前兩項的風險界線，適用於基礎分類器的具體選擇，即Fisher的線性判別法和K-近鄰分類器。這裡的關鍵點是，這些界限只取決於d、樣本量n和投影數，而不取決於原始數據維度p。

結論: 我們引入了一個通用的高維分類框架，通過將基礎分類器的應用結果與精心挑選的數據的低維隨機投影相結合。其吸引人的特點之一是它的通用性：該方法可以與任何基礎分類器一起使用。此外，儘管我們詳細探討了一種結合隨機投影的方法（部分原因是它有利於嚴格的統計分析），但這裡還有許多其他選擇。例如，與其只保留每個區塊內產生最小的測試誤差估計值的投影，不如給不同的投影加權，加權隨著測試誤差估計值的增加而減少。在這項工作中，我們主要關注的是隨機投影集合分類器的分類性能，而不是類別分配的可解釋性。然而，所選擇的投影提供了權重，表明了模型中不同變量的相對重要性。因此，另一個有趣的方向是了解由隨機投影集合分類器引起的變量排名的特性。

與論文的相關性: