本研究所使用之垃圾郵件（Spam）資料集來自UCI機器學習資料庫中。該資料集是由 Hewlett-Packard Laboratories 的 Mark Hopkins、Erik Reeber、George Forman 及 Jaap Suermondt 等人於 1999 年整理提供。此資料集的建立，部分內容參考了 Cranor 與 LaMacchia（1998）於 Communications of the ACM 期刊中發表的「Spam!」一文。

表5.1為垃圾郵件資料集的變數說明，此資料集共包含 58 個變數，涵蓋了單字出現比例、特殊字元出現比例、與大寫字母使用情形相關的統計指標，以及是否為垃圾郵件的分類標籤。資料集中共有 4,601 筆觀察值，其中標記為非垃圾郵件（記作 0）者共有 2,788 筆，占整體資料的約 60.6%；標記為垃圾郵件（記作 1）者共有 1,813 筆，占整體資料的約 39.4%。

表5.1 垃圾郵件資料集變數說明

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 編號 | 變數名稱 | 意義 | 變數型態 | 備註 |
| 1 | V1 | 單字匹配的百分比 | 數值型 | 單字為「make」 |
| 2 | V2 | 單字匹配的百分比 | 數值型 | 單字為「address」 |
| 3 | V3 | 單字匹配的百分比 | 數值型 | 單字為「all」 |
| 4 | V4 | 單字匹配的百分比 | 數值型 | 單字為「3d」 |
| 5 | V5 | 單字匹配的百分比 | 數值型 | 單字為「our」 |
| 6 | V6 | 單字匹配的百分比 | 數值型 | 單字為「over」 |
| 7 | V7 | 單字匹配的百分比 | 數值型 | 單字為「remove」 |
| 8 | V8 | 單字匹配的百分比 | 數值型 | 單字為「internet」 |
| 9 | V9 | 單字匹配的百分比 | 數值型 | 單字為「order」 |
| 10 | V10 | 單字匹配的百分比 | 數值型 | 單字為「mail」 |
| 11 | V11 | 單字匹配的百分比 | 數值型 | 單字為「receive」 |
| 12 | V12 | 單字匹配的百分比 | 數值型 | 單字為「will」 |
| 13 | V13 | 單字匹配的百分比 | 數值型 | 單字為「people」 |
| 14 | V14 | 單字匹配的百分比 | 數值型 | 單字為「report」 |
| 15 | V15 | 單字匹配的百分比 | 數值型 | 單字為「addresses」 |
| 16 | V16 | 單字匹配的百分比 | 數值型 | 單字為「free」 |
| 17 | V17 | 單字匹配的百分比 | 數值型 | 單字為「business」 |
| 18 | V18 | 單字匹配的百分比 | 數值型 | 單字為「email」 |
| 19 | V19 | 單字匹配的百分比 | 數值型 | 單字為「you」 |
| 20 | V20 | 單字匹配的百分比 | 數值型 | 單字為「credit」 |
| 21 | V21 | 單字匹配的百分比 | 數值型 | 單字為「your」 |
| 22 | V22 | 單字匹配的百分比 | 數值型 | 單字為「font」 |
| 23 | V23 | 單字匹配的百分比 | 數值型 | 單字為「000」 |
| 24 | V24 | 單字匹配的百分比 | 數值型 | 單字為「money」 |
| 25 | V25 | 單字匹配的百分比 | 數值型 | 單字為「hp」 |
| 26 | V26 | 單字匹配的百分比 | 數值型 | 單字為「hpl」 |
| 27 | V27 | 單字匹配的百分比 | 數值型 | 單字為「george」 |
| 28 | V28 | 單字匹配的百分比 | 數值型 | 單字為「650」 |
| 29 | V29 | 單字匹配的百分比 | 數值型 | 單字為「lab」 |
| 30 | V30 | 單字匹配的百分比 | 數值型 | 單字為「labs」 |
| 31 | V31 | 單字匹配的百分比 | 數值型 | 單字為「telnet」 |
| 32 | V32 | 單字匹配的百分比 | 數值型 | 單字為「857」 |
| 33 | V33 | 單字匹配的百分比 | 數值型 | 單字為「data」 |
| 34 | V34 | 單字匹配的百分比 | 數值型 | 單字為「415」 |
| 35 | V35 | 單字匹配的百分比 | 數值型 | 單字為「85」 |
| 36 | V36 | 單字匹配的百分比 | 數值型 | 單字為「technology」 |
| 37 | V37 | 單字匹配的百分比 | 數值型 | 單字為「1999」 |
| 38 | V38 | 單字匹配的百分比 | 數值型 | 單字為「parts」 |
| 39 | V39 | 單字匹配的百分比 | 數值型 | 單字為「pm」 |
| 40 | V40 | 單字匹配的百分比 | 數值型 | 單字為「direct」 |
| 41 | V41 | 單字匹配的百分比 | 數值型 | 單字為「cs」 |
| 42 | V42 | 單字匹配的百分比 | 數值型 | 單字為「meeting」 |
| 43 | V43 | 單字匹配的百分比 | 數值型 | 單字為「original」 |
| 44 | V44 | 單字匹配的百分比 | 數值型 | 單字為「project」 |
| 45 | V45 | 單字匹配的百分比 | 數值型 | 單字為「re」 |
| 46 | V46 | 單字匹配的百分比 | 數值型 | 單字為「edu」 |
| 47 | V47 | 單字匹配的百分比 | 數值型 | 單字為「table」 |
| 48 | V48 | 單字匹配的百分比 | 數值型 | 單字為「conference」 |
| 49 | V49 | 符號匹配的百分比 | 數值型 | 符號為「;」 |
| 50 | V50 | 符號匹配的百分比 | 數值型 | 符號為「(」 |
| 51 | V51 | 符號匹配的百分比 | 數值型 | 符號為「[」 |
| 52 | V52 | 符號匹配的百分比 | 數值型 | 符號為「!」 |
| 53 | V53 | 符號匹配的百分比 | 數值型 | 符號為「$」 |
| 54 | V54 | 符號匹配的百分比 | 數值型 | 符號為「#」 |
| 55 | V55 | 連續大寫字母序列的平均長度 | 數值型 |  |
| 56 | V56 | 最長的連續大寫字母序列的長度 | 數值型 |  |
| 57 | V57 | 所有連續大寫字母序列長度的總和 | 數值型 |  |
| 58 | y | 垃圾郵件判斷 | 數值型 | 1=是；0=否 |

表5.2為垃圾郵件資料集各變數變異數膨脹因子（Variance inflation factor），為了避免變數共線性的問題，使用邏輯斯迴歸計算各變數的變異數膨脹因子，由於連續大寫字母序列的平均長度（V55）與最長的連續大寫字母序列的長度（V56）具有較高的變異數膨脹因子，分別為為 6.0892 和 5.3020 ，這表明這兩個變數之間可能存在共線性問題。因此，決定將連續大寫字母序列的平均長度變數（V55）刪除，並重新計算。結果顯示，刪除 連續大寫字母序列的平均長度（V55） 後，最長的連續大寫字母序列的長度（V56）的降低變異數膨脹因子至 2.5347 ，已回到正常範圍，顯示出共線性問題得以改善。

表5.2 垃圾郵件資料集各變數變異數膨脹因子

|  |  |  |  |
| --- | --- | --- | --- |
| 編號 | 變數名稱 | 變異數膨脹因子 | 變異數膨脹因子（移除V55） |
| 1 | V1 | 1.3687 | 1.3586 |
| 2 | V2 | 1.0236 | 1.0221 |
| 3 | V3 | 1.0531 | 1.0511 |
| 4 | V4 | 1.0116 | 1.0116 |
| 5 | V5 | 1.0546 | 1.0544 |
| 6 | V6 | 1.0578 | 1.0474 |
| 7 | V7 | 1.0586 | 1.0584 |
| 8 | V8 | 1.0652 | 1.0654 |
| 9 | V9 | 1.1345 | 1.1042 |
| 10 | V10 | 1.0388 | 1.0371 |
| 11 | V11 | 1.4760 | 1.4697 |
| 12 | V12 | 1.105 | 1.1024 |
| 13 | V13 | 1.0535 | 1.0531 |
| 14 | V14 | 1.0359 | 1.0347 |
| 15 | V15 | 1.0608 | 1.0604 |
| 16 | V16 | 1.1127 | 1.1123 |
| 17 | V17 | 1.0911 | 1.0886 |
| 18 | V18 | 1.0957 | 1.0902 |
| 19 | V19 | 1.2165 | 1.2161 |
| 20 | V20 | 1.1866 | 1.1781 |
| 21 | V21 | 1.4014 | 1.3980 |
| 22 | V22 | 2.8747 | 2.8932 |
| 23 | V23 | 1.0708 | 1.0693 |
| 24 | V24 | 1.056 | 1.0563 |
| 25 | V25 | 1.5359 | 1.5317 |
| 26 | V26 | 1.4926 | 1.4903 |
| 27 | V27 | 3.1748 | 2.1941 |
| 28 | V28 | 1.1207 | 1.1196 |
| 29 | V29 | 1.0127 | 1.0127 |
| 30 | V30 | 1.0411 | 1.0393 |
| 31 | V31 | 1.0134 | 1.0146 |
| 32 | V32 | 1.0248 | 1.0283 |
| 33 | V33 | 1.1155 | 1.1122 |
| 34 | V34 | 1.0162 | 1.0185 |
| 35 | V35 | 1.0930 | 1.0915 |
| 36 | V36 | 1.1512 | 1.1506 |
| 37 | V37 | 1.0673 | 1.0661 |
| 38 | V38 | 1.0295 | 1.0290 |
| 39 | V39 | 1.0660 | 1.0660 |
| 40 | V40 | 1.0615 | 1.0610 |
| 41 | V41 | 1.0433 | 1.0409 |
| 42 | V42 | 1.0469 | 1.0481 |
| 43 | V43 | 1.0178 | 1.0177 |
| 44 | V44 | 1.1539 | 1.1484 |
| 45 | V45 | 1.1481 | 1.1482 |
| 46 | V46 | 1.0609 | 1.0598 |
| 47 | V47 | 1.0313 | 1.0305 |
| 48 | V48 | 1.0684 | 1.0659 |
| 49 | V49 | 2.4571 | 2.4688 |
| 50 | V50 | 1.0808 | 1.0798 |
| 51 | V51 | 1.0227 | 1.0219 |
| 52 | V52 | 1.0627 | 1.0625 |
| 53 | V53 | 1.1985 | 1.2015 |
| 54 | V54 | 1.4730 | 1.4776 |
| 55 | V55 | 6.0892 |  |
| 56 | V56 | 5.3020 | 2.5347 |
| 57 | V57 | 1.6845 | 1.5274 |