

用於基因微陣列資料的兩階段式基因選擇

Two-stage gene selection for microarray

楊正宏

Cheng-Hong Yang

高雄應用科技大學

chyang@cc.kuas.edu.tw

柯兆軒

Chao-Hsuan Ke

高雄應用科技大學

a092205237@cc.kuas.edu.tw

莊麗月

Li-Yeh Chuang

義守大學

chuang@isu.edu.tw

楊正三

Cheng-San Yang

嘉義基督教醫院

摘要

基因微陣列是一個具有良好成效的醫學診斷工具，目前被廣泛的用於分析基因與疾病之間的表現特性，藉由其具有高特徵維度的資料特性，可以同時的分析數萬個基因對於疾病的表現關係，但也因為擁有高特徵維度的特性，因此在進行分類計算時，容易造成運算上的負擔。基因（特徵）的選取可以有效的選取出對於分類計算具有影響的基因，剔除不必要基因，藉此提高分類計算的效率，並提高分類正確率。在本研究中，我們利用資訊增益與基因演算法對於基因微陣列資料進行兩個階段的基因選擇，將選取出的基因代替原始的資料集，並搭配 K 個最近鄰居法計算分類正確率。研究結果顯示，利用本研究提出的方法可以選擇出數量較少且對於分類計算較有影響性的基因資料，而利用選取出的基因資料進行分類計算後，可以得到更好的分類正確率。

關鍵字：基因選擇、資訊增益、基因微陣列

Abstract

The microarray is a medical diagnostic tool with good effect, and it was used for analyzing the behavior characteristic between the gene and disease by the extensive one at present. Microarray data are characterized by a high dimension, which could be analyze more than thousand of gene and diseases simultaneously, but it will lead to need more calculate time when implement classification. The genes (features) selected can extraction genes which influence classification accuracy effectively, to eliminate the useless genes, and to improve the calculate performance and the classification accuracy. In this study, we aimed

for microarray used two-stage methods which are information gain and genetic algorithm to select genes, and used K-nearest neighbor to calculate the classification accuracy. Experimental results show that the proposed methods can extraction smaller number of genes which can influence classification accuracy effectively.

Keyword : gene select, information gain, microarray expression data

1、前言

基因微陣列(Microarray)目前主要的被用於分析基因與疾病之間的關聯，利用一次可以存放數千至數萬個基因的特性，針對具有疾病與沒有疾病的基因同時進行比對，如此可以有效的提升分析效率並節省實驗成本，因此利用基因微陣列資料對於當前的基因體研究越來越重要，特別是在疾病的探討和藥學發展上更為迫切，良好的分析結果可以讓研究人員在研究的過程逐一處理複雜的生物問題，另外在醫藥的發展上也可以更迅速的獲得所需的資訊。

特徵選取的目的是在於從原始資料的特徵集合中選擇出較為重要的特徵子集合，而此特徵子集合要能夠代表原始的特徵資料，目前在具有大量資料的研究中，如圖形辨識、生物資訊和文件分類[8][12]，針對這些資料進行處理前，先進行特徵選取的動作，選擇出具有重要性的特徵子集合來代表原始資料，利用少量的特徵資料來進行如分類計算，如此將可以有效的提升計算效率，並提高分類正確率[1][9]。

而特徵選取主要分為兩種架構，一種稱為 filter 架構，而另一種是 wrapper 架構。一般來說，當執行特徵選取時與分類演算法是獨立進行時，通稱為 filter 架構；而其他方式，如同時與分類演算法一起進行時則稱為

wrapper 架構。Filter 的架構目的在於計算出每個特徵資料對於區別類別結果的權重值，選擇出權重值較高的特徵資料來代表原始的資料，如此可以有效的降低原始資料的特徵維度，並且選擇出的特徵子集合與原始資料不會有太大的差異；目前屬於 filter 架構常見的方法有資訊增益 (information gain, IG)、互斥資訊 (mutual information, MI)、信號與雜訊 (signal to noise, S2N)、和卡方檢定法 (chi-squared method, χ^2 -method)。

Wrappers 特徵選取方法的研究方式是藉由持續的加入或刪除特徵，並且利用學習演算法中的目標函數評估特徵集合的優劣，再利用搜尋策略朝向最佳化的特徵集合進行搜尋。這種特徵選取的研究方式相當於最佳化的技術，雖然針對單一個特徵維度而言或許無法獲得較高的重要性，但是對於整個特徵集合而言，不僅特徵集合與類別之間具有相關性，同時，特徵與特徵之間也具有相關性。而 Wrappers 的架構中最常利用最佳化演算法來針對特徵進行加入或刪除的動作，並搭配分類器計算分類正確率來評估所選擇出的特徵組合對於分類結果是否具有影響；而常用的最佳化演算法有基因演算法 (Genetic Algorithm, GA) 和粒子族群最佳化 (Particle Swarm optimization, PSO) [3]等。

基因演算法是一種模仿自然界藉由交配和突變等演化機制所提出的一種最佳化演算法，在過去的研究中，利用基因演算法在分類問題的資料上先做特徵選取 [8][10]，再將選取出的特徵當作原始資料做分類計算，研究結果顯示，利用特徵選取可以選出較有意義的資料，並排除例如雜訊等會影響分類結果的資料，如此可以有效的提升之後分類時計算效率，並提升其分類正確率。

在本研究中，我們針對基因微陣列資料進行兩個步驟的基因選擇，第一個步驟是利用資訊增益計算出每個基因的資訊增益值，將對於區分類別結果具有明顯的基因挑選出來；而第二步驟則是利用基因演算法將第一步驟所選擇出的基因再進行一次的基因選擇，並搭配 K 最近鄰居法進行分類正確率的評估，用於判斷所挑選出的基因是否真的對於類別的計算是否具有影響。我們選用了 6 個基因微陣列來進行測試，研究結果顯示，經過兩個階段的基因選擇後進行分類計算，可以得到較少的基因數量，以及更好的分類正確率。

2、相關研究

2.1、資訊增益 (Information gain, IG)

資訊增益 (information gain, IG) 是由 Quinlan 於 1979 年提出 [7]，主要是做為 ID3 (Interactive Dichotomizer 3, ID3) 決策樹歸納演算法建構決策樹的依據，用於判斷分割點的屬性。

計算分類問題時，資訊增益也可用於判斷特徵屬性與其他類別屬性的差異，常用於決策樹演算法上選擇分裂點 (split point)；而資訊增益也可用於特徵的選取，主要是以計算單一個特徵獲得資訊增益值的大小做為特徵是否被挑選或被刪除的依據，挑選特徵的時候必須設定一個門檻值 (threshold)，當特徵計算的資訊增益值大於門檻值代表特徵具有足夠的分類辨識資訊必須被挑選；若特徵計算的資訊增益值小於門檻值代表特徵並不具有足夠的分類辨識資訊必須被刪除。

舉例而言，假設分類問題包含有資料 $N = \{1, 2, \dots, n\}$ 、特徵維度 $D = \{1, 2, \dots, d\}$ 以及類別 $K = \{1, 2, \dots, k\}$ ，針對單一個特徵的資訊增益測量方式必須計算兩個相關性的數值，這兩個相關性的數值皆稱為熵值 (entropy)，如方程式 (1) 和方程式 (2)，而這兩個熵值的差即稱作為資訊增益，如方程式 (3)：

$$Entropy(N) = \sum_{i=1}^K P_i \log_K \left(\frac{1}{P_i} \right) = - \sum_{i=1}^K p_i \log_K p_i \quad (1)$$

$$Entropy(D_j) = \sum_{i=1}^{|D_j|} \frac{D_{ji}}{N} \times Entropy(D_{ji}) \quad (2)$$

$$IG(D_j) = Entropy(N) - Entropy(D_j) \quad (3)$$

在方程式 (1) 中，Entropy(N) 主要計算的是整個分類問題具有的總資訊含量，以此總資訊含量做為計算單一個特徵獲得資訊增益的依據。其中，N 代表的是資料數量，K 代表的是類別數量， P_i 代表的是第 i 個類別在 N 筆資料中出現的機率。

在方程式 (2) 中，Entropy(D_{ji}) 主要計算的是第 j 個特徵維度中，第 i 種數值、類別與資料數量之間的資訊含量，將特徵中不同數值、類別與資料數量之間的資訊含量加總即為單一個特徵的總資訊含量。其中， D_{ji} 代表的是第 j 個特徵維度中包含有 i 種不同的數值，N 代表的是資料數量。

在方程式 (3) 中，IG(D_j) 代表的即為分類問題中第 j

個特徵所獲得的資訊增益，是以整個分類問題具有的總資訊含量與第 j 個特徵的總資訊含量之間的差為計算方式。經由方程式 (1)、方程式 (2) 和方程式 (3) 的計算後，就可以獲得分類問題中每一個特徵的資訊增益，再藉由所設定的門檻值進行評估後，挑選資訊增益值大於門檻值的特徵，組合成具有足夠分類辨識資訊的特徵集合。

2.2 基因演算法 (Genetic Algorithm, GA)

基因演算法是一種進化式演算法 (Evolutionary Algorithm, EA)，在 1975 年由 John Holland 所提出[6]，藉由觀察生物演化的步驟，提出染色體選擇、交配和突變的步驟來解決最佳化問題；透過選擇出的染色體互相交配，可以產生出較佳的下一代，也利用突變的特性，產生突破性的發展。在基因演算法的設計中，通常將許多的解編碼成一個染色體(Chromosome)，而染色體中每一個基因(Gene)代表著一個解，染色體長度越長，其複雜度越高。基因演算法俱有影響效能的四個主要變數依序為：族群數量、迭代次數、交配率和突變率。通常擁有大量的族群和較多的迭代次數較容易得到最佳解，但相對的其缺點是會增加計算的時間。而基因演算法的步驟依序為：1.初始化染色體 2.計算適應值 3.選擇染色體 4.交配 5.突變 6.取代 7.終止條件。

2.3 K 最近鄰居法 (K-nearest neighbor)

K 最近鄰居法(K-nearest neighbor)是由 Fix 和 Hodges 在 1951 年所提出，利用特定的距離測量方式如：歐幾里得距離(Euclidean distance)，來找出最接近預測點的 K 個已知點，利用此 K 個已知點的結果來當作預測點的結果。在分類問題中，可由從已知類別的資料利用 K 最近鄰居法對於未知類別的資料做分類預測，經由計算未知類別資料與已知類別資料的最小距離，依序可得知有 K 個已知類別資料點與未知類別資料點的最小距離，此 K 個已知類別資料點的類別結果，其出現頻率最高的類別即是此未知類別資料的分類結果。

2.4 正確率評估

目前在分類計算的研究中最常被使用的正確率評估方式有 K-fold 評估法與留一交叉評估法(LOOCV)[1]；然

而針對不同的資料集必須採取不同的評估方法，以避免產生 overfitting，導致獲得不客觀的正確率，通常在針對具有大數量的資料集，會使用 K-fold 評估法，而針對只有小數量的資料集，則是採用留一交叉評估法(LOOCV)，在本研究中，我們使用留一交叉評估法來評估分類正確率。

2.5 IG-GA/KNN

在本研究中，我們針對基因微陣列資料使用了資訊增益與基因演算法兩個階段的方法進行基因選擇的動作，在第一階段中先利用資訊增益選出特徵定數量的基因後，再利用基因演算法搭配 K 最近鄰居法當作分類器進行評估，進行第二階段的基因選擇。

在第一個階段中，我們利用 Weka 軟體[4]先計算出基因微陣列資料中每個基因（特徵）的資訊增益值，經過計算後若所得到的資訊增益值越高，表示此基因對於與其他類別的區別性越高，也就表示對於之後的分類計算具有較重要的影響程度。經過計算每個基因的資訊增益值後，須設定一個數值當作門檻值，用來區別基因的重要性，從研究中發現，利用資訊增益計算後大部分基因的資訊增益值皆為 0，這表示基因微陣列資料中具有許多對於類別結果的區別沒有重要的影響性，因此在本研究中，我們將資訊增益值的門檻值設定為 0，當基因的資訊增益值有大於 0 的則用來進行下一階段的計算，而等於 0 的基因將不被用於下一個階段的計算。舉例來說，在一個具有 10 個基因（10 個特徵數量）的微陣列資料集中 $f_1 f_2 f_3 f_4 f_5 f_6 f_7 f_8 f_9 f_{10}$ ，若利用資訊增益計算出符合門檻值的基因有 5 個，分別是 f_1 、 f_2 、 f_4 、 f_7 和 f_{10} ，而在第二階段時，則只使用這 5 個基因（ $f_1 f_2 f_4 f_7 f_{10}$ ）繼續執行選擇的動作。經過第一個階段資訊增益值的計算後，符合門檻值的基因將被繼續用於第二階段的計算；而在第二階段的基因選擇中，我們利用基因演算法來進行基因的選取，整個基因演算法的流程如下列步驟：首先要先初始化染色體，而為了使基因演算法能運用於基因的選取，所以在染色體的編碼部分中，我們將染色體編碼為二進位來表示；當位元值為{1}時，表示此基因被選取，當位元值為{0}時，表示此基因不被選取；在具有 10 個基因的微陣列資料集來說，經過第一步驟的計算資訊增益後，只剩下 5 個基因 $f_1 f_2 f_4 f_7 f_{10}$ ，則基因演

算法的染色體在進行初始化時，只能針對此5個基因進行初始化的動作，而初始化的結果是使用隨機方式產生，例如 $C_i = 10011$ ，表示在第 C_i 個染色體中只有第 f_1 、 f_7 和 f_{10} 的基因被選取；在適應函數的部份則是使用K最近鄰居法，而被選擇出的基因利用K最近鄰居法所計算出的分類正確率，即是適應函數值，利用適應函數值可用來判斷選擇出的基因是否確實對於分類結果具有影響性；而基因演算法經過不斷的選擇、交配與突變的動作後，當達到終止條件後，即可完成整個基因選擇的動作，此時即可得知哪些基因有被選擇出。

在本研究中基因演算法的參數設定依序為：族群數為30、交配率 0.8、突變率 0.05、迭代次數 100 次，而終止條件則是當分類正確率達到 100%或是達到最大的迭代次數時；K 最近鄰居法中的 K 值設定為 1。

3. 結果與討論

針對基因微陣列表現資料選擇出有意義的基因資訊，目前對於資訊科學家是一個極大的挑戰，正確的預測結果可以幫助生物學家解決複雜的基因與疾病等相關問題。由於基因微陣列表現資料具有高特徵維度與低資料量的特性，因此在計算分類問題時會耗費大量的運算時間，如何有效的選擇出對於分類運算具有影響的基因是當前重要的研究主題。

在本研究中，我們先計算基因微陣列表現資料中每個基因的資訊增益值，先選擇出對於分類結果有明顯影響性的基因，再利用基因演算法進行一次基因的選取。此次研究所使用的資料集是從 <http://www.gems-system.org> 下載[1]，並分別針對 6 筆的基因微陣列表現資料進行基因選擇並計算分類正確率，而此 6 筆資料詳細的資料特性如表 1 所示。

表 1 中也記錄著此 6 筆基因微陣列表現資料的原始基因數量，以及利用資訊增益計算後符合門檻值的基因數量，其中符合門檻值最高的是 Lung_Cancer 資料集，符合門檻值的基因數量占了原始特徵數量的 75.88%，而最少的則是 9_Tumors 資料集，符合門檻值的基因數量只占了原始特徵數量的 2.88%。

在表格 2 中比較了其他文獻的實驗結果與本研究提出的方法，在 Statnikov *et al.* 的研究中[1]，在分類器部分區分為 Non-SVM 與 MC-SVM 兩種，其中 Non-SVM

部分最好的分類結果是 KNN，其分類正確率平均是 78.41%，而 MC-SVM 部分則是 CS，其分類正確率平均是 90.87%。而本研究所提出的方法針對此 6 筆基因微陣列表現資料進行分類計算，得到的分類正確率平均是 94.41%，比較其他的研究平均多了 3.54%，從結果可知，本研究所提的方法確實可以有效的選擇出對於分類計算具有影響性的基因，因此之後在計算分類結果時，確實可以有效的提升分類正確率。我們另外也針對只利用資訊增益選擇出的基因進行分類計算，得到的分類正確率是 86.82%，雖然也得到良好的分類正確率，但仍然無法與 MC-SVM 部分的分類結果要來的更好；由此可知，本研究所提出兩階段的基因選擇方式，確實可以有效的選擇出對於分類計算具有影響性的基因，並使得在日後的分類計算得到更好的結果。

在此 6 筆的基因微陣列表現資料中，先經過基因選擇之後再進行分類計算後所得到的分類正確率比較於未經過基因選擇而直接進行分類計算所得到的分類正確率要來的更好，但從表 2 中可知，11_Tumors 資料集的微陣列表現資料先經過基因選取後再進行分類計算所得到的分類正確率並無法勝過 MC-SVM 架構的分類正確率，其主要原因在於在本次的研究中，我們所使用的分類計算是利用 K 個最近鄰居法，而不是使用支持向量機 (Support Vector Machine, SVM)；在分類計算相關的研究中已經指出，支持向量機對於計算高特徵維度與類別數量高的資料集中，會有較好的分類表現結果[11]，因此相較於 MC-SVM 架構的分類正確率，我們經由基因的選擇後再利用 K 個最近鄰居法進行分類計算時，還是不容易在分類正確率上取得較好的結果，因此我們建議，在未來進行計算基因微陣列表現資料的相關研究中，可以使用支持向量機用於分類計算。然而在 Lung_Cancer 此筆基因微陣列表現資料中，經由基因選擇後進行分類運算所得到的分類結果也是較 MC-SVM 架構的分類正確率稍差，主要的原因在於在本研究中第一階段計算每個基因的資訊增益值先選取出重要的基因，在此部分 Lung_Cancer 符合我們所設定的門檻值的基因數量占了原始的資料集達到了 75.88%，相較於其他 10 筆資料的 2.88%至 28.99%要高出許多，我們認為這是影響了最後分類正確率的主要原因；而原因在於 Lung_Cancer 的每個基因資訊對於分類結果皆有不少的影響性，因此資訊增益值無法有效的區別哪一

個基因對於區別分類結果是具有更大的影響性，因此在第二階段中利用基因演算法再進行一次基因選擇的時候，要找到更佳的分類結果其困難度將會提升，為了避免類似的問題產生，我們建議在第一階段中利用計算資訊增益值來做基因選擇的時候，可以將資訊增益值得門檻值提升，藉由提高的門檻值來區別出更重要的基因，用來選擇出對於計算分類結果有更重要影響性的基因。

在本研究中利用了兩個階段的基因選擇方式，在第一個階段先計算出基因微陣列資料中每個基因的資訊增益值，目的在於先從全部的基因中選擇出對於區別不同的分類結果有重要影響性的基因，在完成利用資訊增益值選擇出重要的基因後，利用選擇出的基因進行分類計算，相較於不進行選擇而直接進行分類計算，確實可以有效的提升分類正確率；然而基因的選取一直是一種複雜的組合問題，沒有特定數量的基因組合才能夠獲得最佳的分類正確率，在 Yeang *et al.*, [2]的研究中也說明了基因數量與分類正確率不具有明顯的關聯性，但在 Huang *et al.*, [5]的研究中說明，針對基因微陣列的資料集，選擇越小數量的基因將可有效的提升分類正確率；為了能有效的降低基因數量，本研究將第一階段利用計算資訊增益所選出的基因資料，第二階段中利用基因演算法選擇出更好的基因組合，如此兩階段的基因選擇方式，經研究結果顯示，確實可以有效的降低基因數量，並提高分類正確率。

4. 結論

在本研究中，我們利用了資訊增益與基因演算法兩個階段的基因選擇方式，針對基因微陣列資料進行基因的選擇，將選擇出的基因用來代表原始的資料集，搭配 K 個最近鄰居法用來計算分類正確率。實驗結果顯示，我們所提出的方法可以有效的降低選擇出的基因資料，如此可以有效的降低運算時間以及提高運算效率，並且可以有效的提升分類正確率。我們使用了 6 筆的基因微陣列資料進行研究，利用提出的方法運算而得到的分類成果中，有 4 筆的基因微陣列資料具有較好的分類結果，而整體平均也提升了 3.54%。在日後的研究中，可以將我們提出的方法用於其他需要進行特徵選取的相關研究，用於當作是一種資料的前處理步驟。

參考文獻

- [1] Alexander R. Statnikov, Constantin F. Aliferis, Ioannis tsamardinos, Douglas Hardin, Shawn Levy. "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis". *Bioinformatics*(21): 631-643 (2005).
- [2] C.H. Yeang, S. Ramaswamy, P. Tamayo, S. Mukherjee, R.M. Rifkin, M. Angelo, M. Reich, E. Lander, J. Mesirov, and T. Golub, "Molecular classification of multiple tumor types", *Bioinformatics*, vol. 1, 2001, pp. 1-7.
- [3] E. Elbeltagi, T. Hegazy, and D. Grierson, "Comparison among five evolutionary-based optimization algorithms," *Advanced Engineering Informatics*, vol. 19, no. 1, pp. 43-53, 2005.
- [4] E. Frank, M. Hall, L. Trigg, G. Holmes, I.H. Witten, "Data mining in bioinformatics using Weka," *Bioinformatics*, Vol. 20, pp. 2479-81, 2004.
- [5] H.L. Huang, C.C. Lee, and S.Y. Ho, "Selecting a minimal number of relevant genes from microarray data to design accurate tissue classifiers," accepted by *BioSystems*, 2006.
- [6] J.H. Holland, "Adaptation in Nature and Artificial Systems," MIT Press, 1992.
- [7] J. R. Quinlan., "Induction of decision trees," *Machine Learning*, No. 1, pp. 81-106, 1986.
- [8] J. Yang and V. Honavar., "Feature subset selection using a genetic algorithm". *IEEE Intelligent Systems*, 13:44-49, 1998.
- [9] Leping Li, Clarice R. Weinberg, Thomas A. Darden, Lee G. Pedersen, "Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. ", *Bioinformatics*, Vol.17, No.12, pp. 1131-1142, 2001
- [10] Oh, I.- S., Lee, J.-S., and Moon, B.-R. Hybrid Genetic algorithms for Feature Selection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.26,

No.11, nov. 2004.

[11] T. Li, C. Zhang, and M. Ogihara, "A Comparative Study of Feature Selection and Multiclass Classification Methods for Tissue Classification Based on Gene Expression," *Bioinformatics*, vol. 20, no. 15, pp. 2429-2437, 2004.

[12] Wenqian Shang, Houkuan Huang, Haibin Zhu, Yongmin Lin, Youli Qu, and Zhihai Wang, "A Novel Feature Selection Algorithm for Text Categorization", *Expert Systems with Applications*, Vol. 33, Issue 1, July 2007.

表 1 本研究所採用的 11 筆基因微陣列資料與利用兩階段式基因選擇所選取出的基因資料

資料集	資料數量	基因數量(A)	類別數	經過 IG 計算後符合門 檻值的基因數量(B)	(B) / (A) 比例	經由 IG-GA 選出的 基因數量
9_Tumors	60	5726	9	165	2.88 %	52
11_Tumors	174	12533	11	3181	25.38 %	479
Brain_Tumor1	90	5920	5	1612	27.23 %	244
Leukemia1	72	5327	3	848	15.92 %	82
Lung_Cancer	203	12600	5	9561	75.88 %	2101
SRBCT	83	2308	4	669	28.99 %	56

表 2 利用不同分類演算法用於計算基因微陣列資料的分類正確率

方法 資料集	Non-SVM			MC-SVM					IG	IG-GA
	KNN	NN	PNN	OVR	OVO	DAG SVM	WW	CS	KNN	KNN
9_Tumors	43.90	19.38	34.00	65.10	58.57	60.24	62.24	65.33	66.67	85.00
11_Tumors	78.51	54.14	77.24	94.68	90.36	90.36	94.68	95.30	83.33	92.53
Brain_Tumor1	87.94	84.72	79.61	91.67	90.56	90.56	90.56	90.56	88.89	93.33
Leukemia1	83.57	76.61	85.00	97.50	97.32	96.07	97.50	97.50	93.06	100.0
Lung_Cancer	89.64	87.80	85.66	96.05	95.59	95.59	95.55	96.55	90.15	95.57
SRBCT	86.90	91.03	79.50	100.00	100.00	100.00	100.00	100.00	98.80	100.0
Average	78.41	68.95	73.50	90.83	88.73	88.80	90.09	90.87	86.82	94.41

附註: (1) Non-SVM: Tradition classification method. (2) MC-SVM: Multi-class support vector machines. (3) KNN: K-Nearest Neighbors. (4) NN: Backpropagation Neural Networks. (5) PNN: Probabilistic Neural Networks. (6) OVR: One-Versus-Rest. (7) OVO: One-Versus-One. (8) DAG: DAGSVM. (9) WW: Method by Weston and Watkins. (10) CS: Method by Crammer and Singer. (11)IG: Information Gain. (12) IG-GA: Information Gain – Genetic Algorithm