Clustering Comparison between EM and Kmeans + NIPS text analysis the Ultimate 白片

統碩一 賴東昇

報告概要

EM 與K-means 實作與比較

1. 介紹EM演算法的概念，流程
2. 找兩組資料來分看看
3. R Package Demo

NIPS 文字分析

1. 燒爛電腦

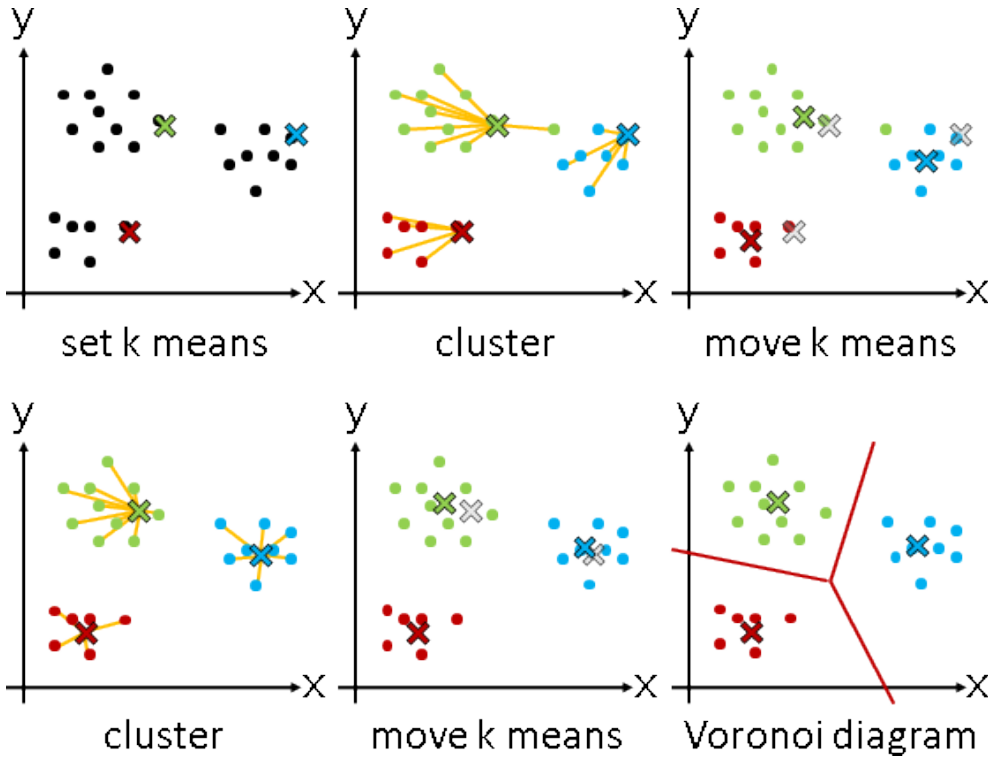
Code & Detail Available :
https://github.com/hyades910739/cluster_NIPS

K-means

演算法流程

1. 從資料中隨機選取K點作為起始點，所有觀測值分別計算與這K點的距離，把各觀測值分配到最近的點
2. 分群後計算群的中心點，作為下次迭代之起始點
3. 重複1.2. 直至收斂或電腦受不了

K-means



Reference: http://www.csie.ntnu.edu.tw/~u91029/Classification.html

K-means

特性

1. 時間複雜度: O(NCI), N為資料數量，C為集群數，I為迭代次數
2. 可以用不同的dissimilarity來計算

EM Algorithm

演算法流程

設有觀測值 X 與隱變數 Y ，透過以下步驟估計X與Y相關的參數

1. E：給定觀測值與第t次的參數估計值，計算x與y的**對數概似函數的期望值**
2. M：用數值方法找出參數估計值，使得E步驟的函數極大化
3. 重複1,2，直至收斂或電腦受不了

EM Algorithm

注意事項

1. 比起自己設定起始點外，使用**隨機起始點**是比較實際的方法(除非有人托夢給你)
2. 再估計共變異矩陣時，需要求權重，即P(Y|X) 不得為0，若皆為0必須做出調整
3. 再進行參數估計時，可以對參數做出限制，例如限制所有高斯模型的共變異矩陣皆相同，或為相異，但共變異為0的矩陣。

EM Algorithm

pseudo code?!

for n個隨機起始點:

while 沒收斂:

計算P(Y|X,$\theta^{(t)}$);

計算α, μ, Σ;

if P(Yi|X,$\theta^{(t)}$)很小:

調整值

if收斂或迭代次數達上限: 離開迴圈

傳回參數估計，log likelihood function

比較n個log likelihood function，回傳值最大者

EM Algorithm

特色

1. 有統計背景的演算法，somehow比較優雅不俗
2. 計算複雜，跑豪久rrrr
3. 概似函數越高，就代表分的越好???

與其他集群方法的比較

有此一說[1]

1. EM 與K-means 的準確率(accuracy)低於SOM與Hierarchical
2. EM 與K-means 在大型資料上表現較佳
3. EM 與K-means 對於雜訊較敏感

[1]: comparison between data clustering algorithms,2008

E-K大對抗

使用資料

利用iris與seeds資料集來比較分群能力

IRIS :

1. 資料來源：R內建資料
2. 資料筆數：150筆
3. 欲分群數：3群

SEEDS:

1. 資料來源：UCI ML[2]
2. 資料筆數：210筆
3. 欲分群數：3群

[2]: https://archive.ics.uci.edu/ml/datasets/seeds

E–K大對抗

比較準則

對於分群方法，我們的比較大致分為三個方向，以下為一些相關準則：[3]

1. 內部比較：silhouette width，connectivity [4]
2. 穩定性比較：WADP[5,6]，APN
3. 正確性比較：purity

[3]: clValid: An R Package for Cluster Validation
[4]:Computational cluster validation in post-genomic data analysis
[5]:Evaluation and Comparison of Clustering Algorithms in Analyzing ES Cell Gene Expression Data
[6]:Molecular classification of cutaneous malignant melanoma by gene expression profiling

silhouette width

Definition:

Let

a(i) : the average distance between i and all other data within same cluster.

b(i) : the lowest average distance of i to all points in any other cluster.

Then:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

所以silhouette 越接近1表示分群結果越好

WADP

定義

設原始分群結果有K 群，其中第j 群有m_j個觀測值。現將各變數(columns)加入雜訊後重新分群，計算這m_j個觀測值兩兩分在不同群的對數，記為D_j。若有m_j個觀測值，則會有$M_j = \frac{m_j(m_j-1)}{2}$組對數。定義WADP為

$$\text{WADP} = \frac{\Sigma_{j=1}^{k} m_j \frac{D_j}{M_j}}{\Sigma_{j=1}^{k} m_j}$$

則WADP接近0則分群較為穩定(robustness)，接近1則易受噪音影響。

Purity

定義

若有原始標籤，就能比較分群準確率，簡單來說，purity就是分對的比率。

設分群結果各群數量為$\Omega = (\omega_k, \dots)$，原始標籤各群數量為$C = (c_j, \dots)$，觀察值總數為N

則定義purity為：

$$Purity(\Omega, C) = \frac{1}{N}\Sigma_k max_j \mid \omega_k \cap c_j \mid$$

越接近1越好

Note：single or average?

由於k-means與EM演算法皆涉及隨機起始點，無法保證每次收斂的結果都一樣，特別是當資料很大，或在較寬鬆的收斂準則下。所以，再不同群數、不同變數、或不同方法下使用這些比較準則，應該進行多次實驗後取平均，比較保守。

其中，WADP更涉及加入隨機噪音，因此必須要多次實驗取平均[5]，才是一合理的估計。

Data : iris

Purity and Confused matrix:

EM :

Purity : 0.97

\Label Cluster	setosa	versicolor	virginica
1	50	0	0
2	0	5	50
3	0	45	0

K-Means :

Purity : 0.89

\Label Cluster	setosa	versicolor	virginica
1	50	0	0
2	0	48	14
3	0	2	36

Data : iris

	Purity	Silhouette	WADP(0.1)	WADP(1)
EM	0.97	0.50	0.19	0.44
K-Means	0.89	**0.55**	0.05	0.64

“各有千秋” 來自水鏡先生的評語

ref: 火鳳燎原

Data : Seeds

	Purity	Silhouette	WADP(0.1)	WADP(1)
EM	0.89	0.44	0.22	0.45
K-Means	0.90	**0.47**	0.06	0.63

結論：

1. 在資料較小時，EM 與K–means 的表現非常相似

   惟 Silhouette 都是K–means較優

2. 雖說分群結果大略相同，但實際運行時間EM遠大於K–means

   當然也是因為EM寫得不夠有效率qq

NIPS 文字分析：

定義問題：

1. 找出論文產量前十大的作者，試著對這些論文進行分群，檢驗分群結果是否能對應到原本的十位作者(群)
2. 刪除掉共寫的論文後，共有460篇

文字處理：

處理流程：

斷詞、清理停用詞、字尾處理

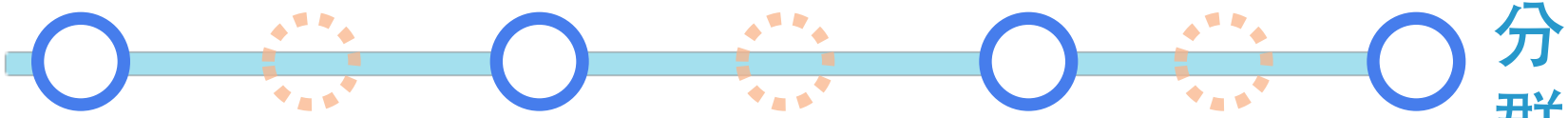
以TF-IDF呈現 Vector Space Model (7241X213090)

利用SVD降至100維的論文向量 (460X100)

分群

文字處理：

處理流程：



斷詞、清理停用詞、字尾處理

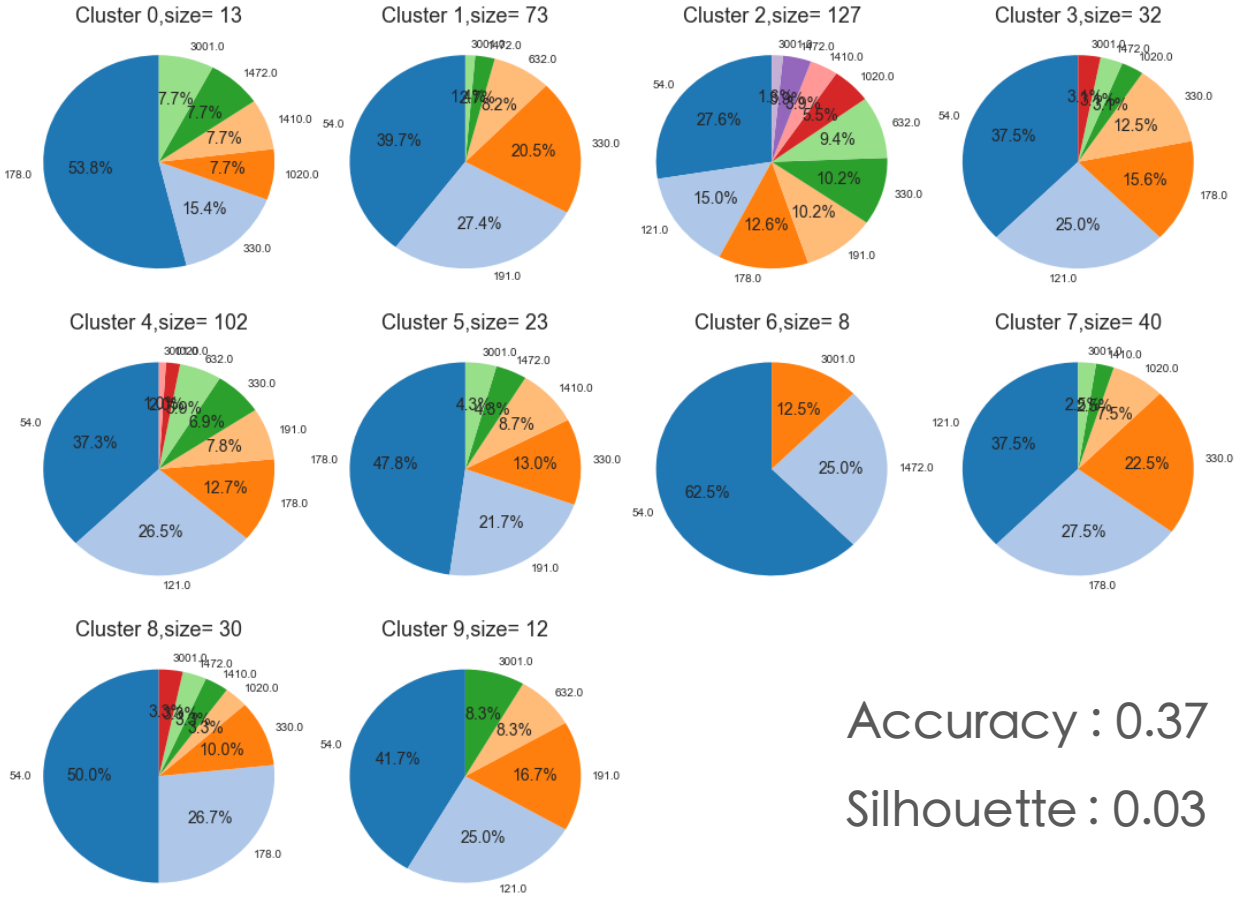
以TF-IDF呈現 Vector Space Model (7241X213090)

利用SVD降至100維的論文向量 (460X100)

分群

問題：要在哪裡取子集

K-means分群結果：



Accuracy : 0.37

Silhouette : 0.03

加入其他特徵：

POS 詞性標記：

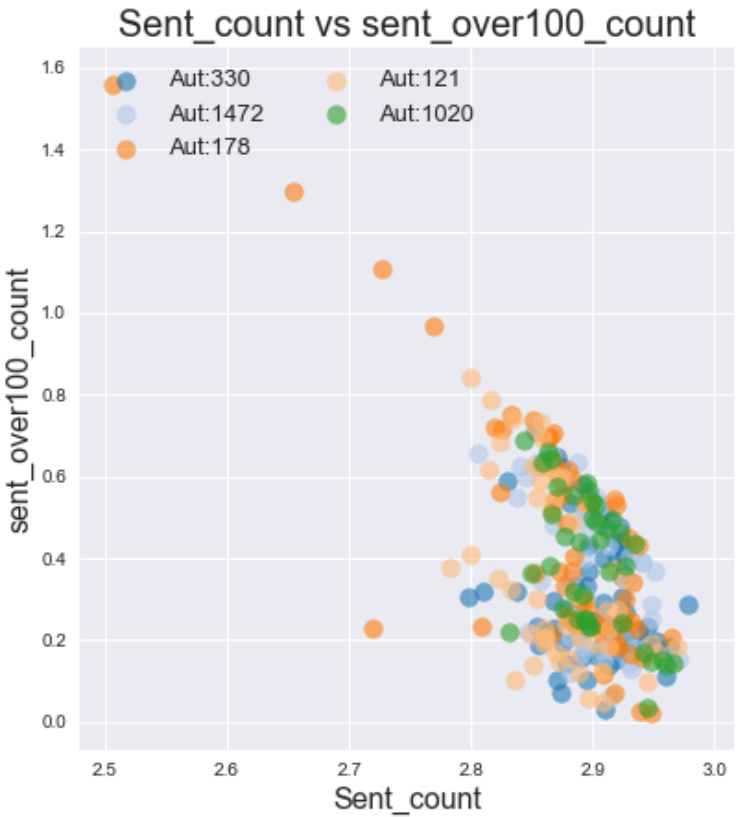
1. 定義超過10字元的句子為一句，計算各論文之句數
2. 找出特定詞性(‘CC’,‘DT’,‘EX’,‘IN’,‘MD’,‘WP’,‘WRB’,’WDT’)的詞語

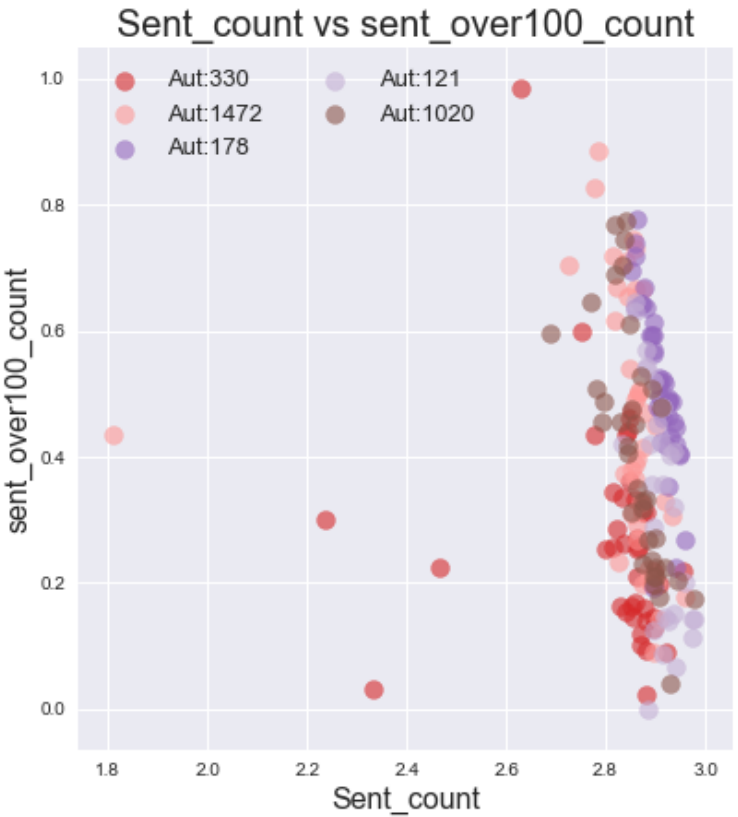
   計算這些詞平均每句出現次數，作為變數

其他特徵：

1. 句數、平均句長、字元數大於100的句數(長句)

加入其他特徵：


Sent_count vs sent_over100_count
Aut:330
Aut:121
Aut:1472
Aut:1020
Aut:178
sent_over100_count
Sent_count
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
2.5
2.6
2.7
2.8
2.9
3.0


Sent_count vs sent_over100_count
Aut:330
Aut:121
Aut:1472
Aut:1020
Aut:178
sent_over100_count
Sent_count
0.0
0.2
0.4
0.6
0.8
1.0
1.8
2.0
2.2
2.4
2.6
2.8
3.0

分群結果： EM(一千次的平均）

	emf_acc	emfp_acc	ems_acc
mean	0.345	0.345	**0.374**
std	0.024	0.023	0.024
50%	0.343	0.343	0.374

	emf_sil	emfp_sil	ems_sil
mean	0.035	0.034	**0.041**
std	0.005	0.005	0.005
50%	0.035	0.035	0.041

分群結果： K-means(一千次的平均)

	kmf_acc	kmfp_acc	kms_acc
mean	0.336	0.336	**0.371**
std	0.021	0.023	0.024
50%	0.335	0.345	0.370

	kmf_sil	kmfp_sil	kms_sil
mean	0.044	0.043	**0.048**
std	0.005	0.005	0.004
50%	0.044	0.043	0.048

“我怕眼淚撐不住!” 來自周杰倫的評語

ref: 火鳳燎原

結論

1. 不論再大資料與小資料中，EM 與 K–means 的表現都滿相似的
2. 通常EM的purity略高於K–means， silhouette 則相反(可能與演算法本身特性有關)
3. 文字分析還有很長一段路要走