藉由分散式系統減少由基因演算法進行資料探勘所花費的時間

指導教授：陳朝鈞
專題成員：邱哲揚
開發工具：JVM、Hadoop2.7
測試環境：Ubuntu14.04

一、簡介

我的專題是為陳教授的合作論文< Mining Group Stock Portfolio by Using Grouping Genetic Algorithms >進行實驗；這一論文的內容，是以基因演算法找出一群公司中適合的投資組合。而投資組合形成的目的為「將投資風險最小化」以及「將投資報酬率的最大化」。

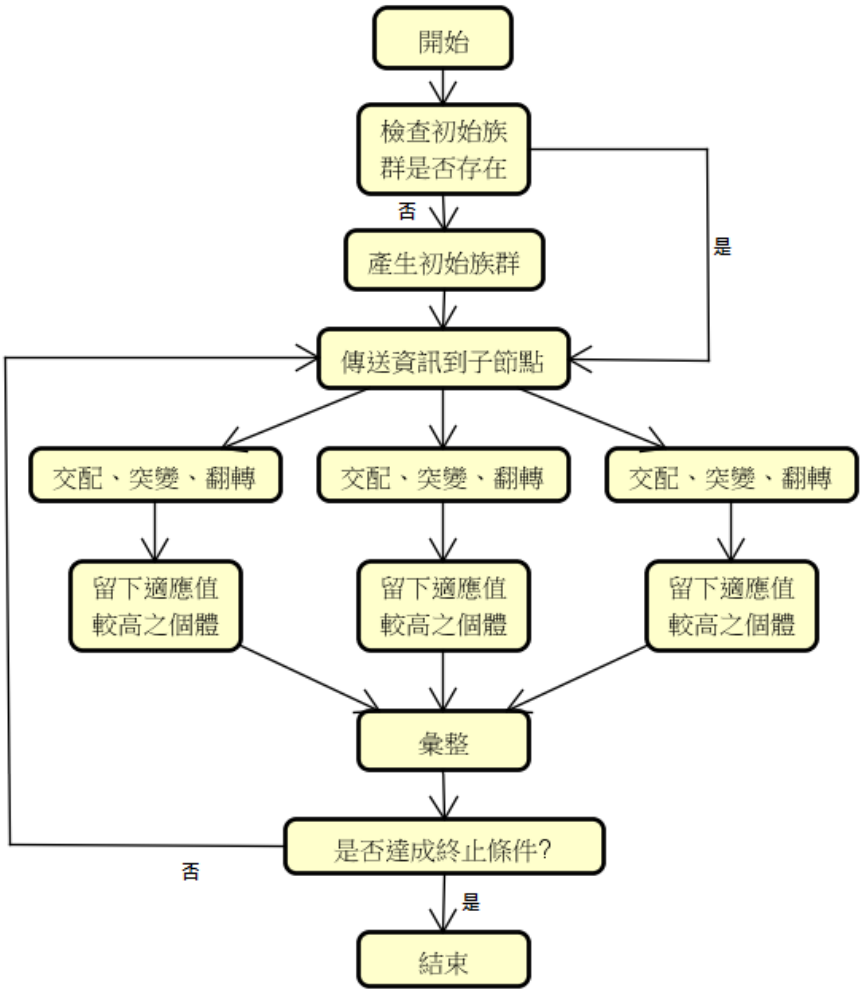
在這個演算法中，我們給予投資組合如下限制條件(1)投資者總預算(2)投資者希望自幾間公司購買股票(3)投資者能自同一間公司買多少張股票(4)公司分組數(5)「投資風險最小化」以及「投資報酬率的最大化」的評比。接著，由於這一演算法為追求分險分散，能形成之投資組合越多越好，故時間複雜度相當高，故使用 hadoop 分散式系統減少演算法找出最佳投資組合所耗費之時間。

基因演算法是演化式演算法 (Evolutionary Algorithms)的一種，模擬生物演化中「適者生存，不適者淘汰」以及「隨機性質」的概念。

在基因演算法中，將各種必要資訊形成一組合，並將其稱之為「染色體(Chromosome) 」。接著，隨機選擇二染色體，並互相交換資訊，形成與原本染色體不完全相同的染色體，此一步驟稱之為「交配(crossover)」。為了產生更多的可能性，基因演算法加入「突變(mutation)」步驟，這一步驟中染色體的任意資訊將隨機改變。接著，再藉由「天擇(selection)」，先計算各個染色體對要求的符合程度，並去除較不適合者，留下較優秀之染色體去產生下一代。重複此一過程，直到達成終止條件，逐步接近所需要之最佳解。

分散式系統允許檔案透過網路在多台主機上分享的檔案系統，可讓多機器上的多使用者分享檔案和儲存空間。在這樣的檔案系統中，客戶端並非直接存取底層的資料儲存區塊，而是透過網路，以特定的通訊協定和伺服器溝通。Hadoop 提供了分布式檔案系統，用以存儲所有計算節點的資料。並且實現了名為 MapReduce 的編程範式：應用程式被分割成許多小部分，而每個部分都能在集群中的任意節點上執行或重新執行。

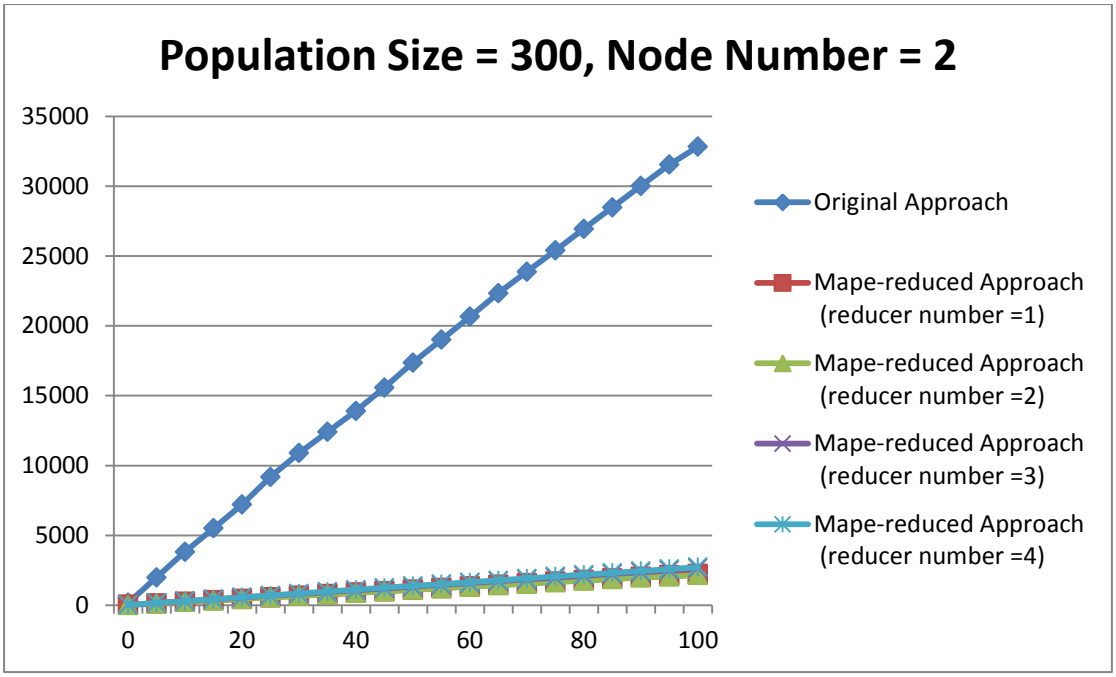
假設有今 n 間公司，將其分為 K 組，則今天染色體中就必須記錄各組別$(G_1, G_1, \dots G_k)$中有哪些公司，該組是否被購買(b)、購買幾張(u)；若今組別一(G_1)的 b 為不買而 u 為 15 張，則G_1中購買之張數為 0*15=0 張；若今組別一(G_1)的 b 為買而 u 為 15 張，則G_1中購買之張數為 1*15=15 張。自各組中各取一間公司，形成這一個投資組合中的一組可能性，當我們評估投資組合時，需考慮這一投資組合的各種組合方式，故時間複雜度將達到 O(n^k)，消耗時間相當大，因此我們使用 Hadoop 將時間分散。



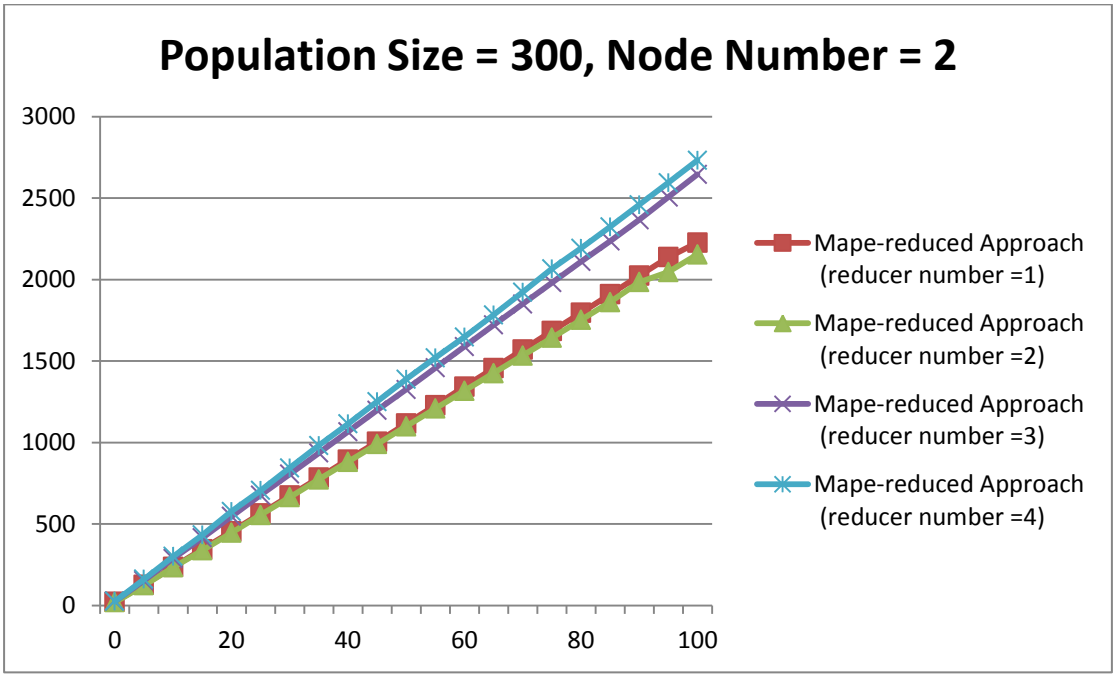
<程式流程圖>

二、測試結果

PopulationSize = 300	Execution Time(SEC)				
Generation	Original Approach	Map-reduced Approach (reducer number =1)	Map-reduced Approach (reducer number =2)	Map-reduce Approach (reducer number =3)	Map-reduce Approach (reducer number =4)
0	167	22	20	26	26
5	1991	127	124	153	161
10	3818	236	234	288	301
15	5508	345	338	415	432
20	7198	454	446	546	577
25	9176	564	555	677	705
30	10892	674	664	806	844
35	12397	784	773	935	981
40	13903	895	881	1066	1114
45	15573	1004	990	1198	1251
50	17360	1116	1099	1325	1388
55	19003	1228	1208	1458	1520
60	20647	1343	1317	1588	1647
65	22338	1456	1425	1721	1783
70	23868	1570	1533	1850	1922
75	25397	1684	1642	1980	2066
80	26940	1796	1752	2110	2191
85	28476	1910	1861	2236	2323
90	30014	2024	1984	2366	2459
95	31537	2137	2046	2506	2594
100	32815	2227	2154	2645	2732



<執行結果表一>



<執行結果表二>

由執行結果表一，相較於單機實驗結果，可以看出 Hadoop 使執行時間減少了。在表二中，比較以 Hadoop 執行，但 reducer 數量不同之結果，我們可依發現一個 Node 上有過多 reducer 需處理時，會因為資源不足而使執行時間增加。