



(19)中華民國智慧財產局

(12)發明說明書公告本

(11)證書號數：TW I796286 B

(45)公告日：中華民國 112 (2023) 年 03 月 21 日

(21)申請案號：106104294

(22)申請日：中華民國 106 (2017) 年 02 月 09 日

(51)Int. Cl. : G06N20/00 (2019.01)

G06N7/00 (2006.01)

(30)優先權：2016/02/29 中國大陸

201610113716.1

(71)申請人：香港商阿里巴巴集團服務有限公司 (香港地區) ALIBABA GROUP SERVICES
LIMITED (HK)

香港

(72)發明人：周俊 (CN)

(74)代理人：林志剛

(56)參考文獻：

CN 102156907A

CN 103136361A

US 2009/0092313A1

US 2012/0041574A1

US 2014/0079297A1

審查人員：王珮如

申請專利範圍項數：14 項 圖式數：6 共 36 頁

(54)名稱

一種機器學習系統的訓練方法和訓練系統

(57)摘要

本發明揭露一種機器學習系統的訓練方法和訓練系統，利用多個樣本資料對機器學習系統進行訓練，該方法包括：獲得多個樣本集合，每個樣本集合包括對應取樣時間段內的樣本資料；根據每一個樣本集合對應的取樣時間段，設置該樣本集合對應的取樣率；獲得多個根據取樣率取樣後的樣本集合；分別確定所述多個取樣後的樣本集合的重要程度值；利用該重要程度值修正所述多個取樣後的樣本集合中的每一個樣本資料，獲得修正後的樣本資料；將每一個所述修正後的樣本資料輸入機器學習系統，對該機器學習系統進行訓練。本發明實施例在將樣本資料輸入機器學習系統之前對樣本資料進行處理，在減輕機器的記憶體資源需求的同時儘量降低對機器學習系統的學習效果的影響。

指定代表圖：

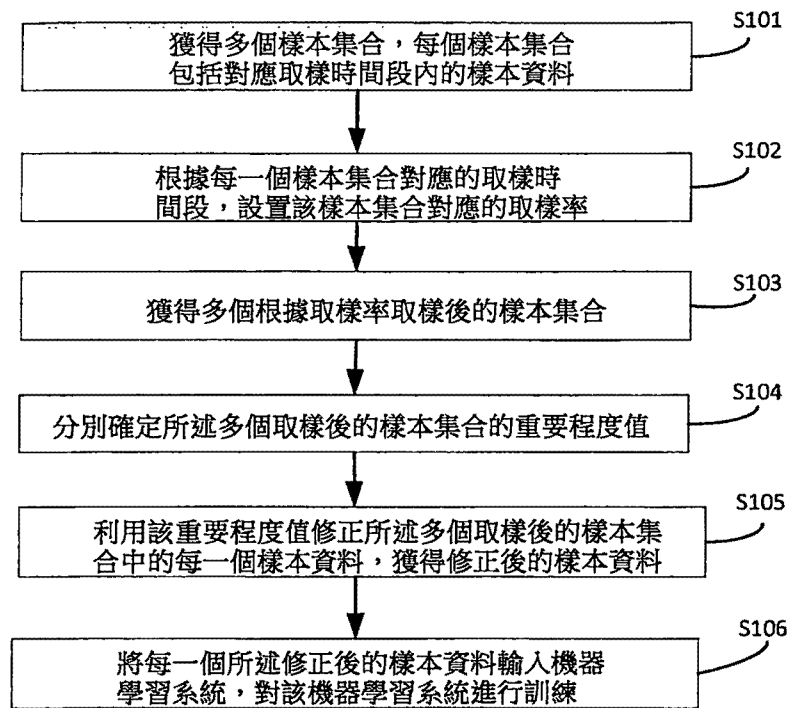


圖 1

I796286

發明摘要

※申請案號：106104294

※申請日：106 年 02 月 09 日

※IPC 分類：

【發明名稱】(中文/英文)

一種機器學習系統的訓練方法和訓練系統

【中文】

本發明揭露一種機器學習系統的訓練方法和訓練系統，利用多個樣本資料對機器學習系統進行訓練，該方法包括：獲得多個樣本集合，每個樣本集合包括對應取樣時間段內的樣本資料；根據每一個樣本集合對應的取樣時間段，設置該樣本集合對應的取樣率；獲得多個根據取樣率取樣後的樣本集合；分別確定所述多個取樣後的樣本集合的重要程度值；利用該重要程度值修正所述多個取樣後的樣本集合中的每一個樣本資料，獲得修正後的樣本資料；將每一個所述修正後的樣本資料輸入機器學習系統，對該機器學習系統進行訓練。本發明實施例在將樣本資料輸入機器學習系統之前對樣本資料進行處理，在減輕機器的記憶體資源需求的同時儘量降低對機器學習系統的學習效果的影響。

【英文】

【代表圖】

【本案指定代表圖】：第(1)圖。

【本代表圖之符號簡單說明】：無

【本案若有化學式時，請揭示最能顯示發明特徵的化學式】：無

發明專利說明書

(本說明書格式、順序，請勿任意更動)

【發明名稱】(中文/英文)

一種機器學習系統的訓練方法和訓練系統

【技術領域】

本發明關於大資料處理領域，尤其關於一種機器學習系統的訓練方法和訓練系統。

【先前技術】

在如今的大資料時代，網際網路公司獲取超大規模資料已非常容易。據不完全統計，谷歌 2012 年每天 30 億 query/300 億廣告，臉書使用者 2013 年每天分享 43 億內容，阿里巴巴 2015 雙十一當天就有超過 7 億筆交易。這些公司藉由機器學習系統，去挖掘資料裡面的金礦，包括使用者興趣/行為/習慣等等。

機器學習系統設計為模仿人腦的神經網路，用於預測使用者的行為。在機器學習系統上線之前，需要藉由大規模的資料進行訓練。然而在訓練過程中，大規模的資料必然要求大規模的機器資源才能有效處理，例如騰訊的廣告資料，都是 PB 層級，必然要用到千台機器以上，這對大部分公司來說，都是個巨大的成本。

為了降低成本，提高機器學習系統的效率，通常的處理方式是藉由隨機樣本取樣的手段減少機器學習系統處理

的資料量。隨機樣本取樣就是以一定概率丟棄樣本，例如對每一個樣本隨機生成 1 個 0-1 範圍內的浮點數，當浮點數大於閾值時則直接丟棄該樣本。然而，隨機丟棄樣本的方式會丟棄大量的有用資料，損害機器學習系統的訓練效果，降低預測的精度。

【發明內容】

鑒於上述問題，提出了本發明實施例以便提供一種克服上述問題或者至少部分地解決上述問題的機器學習系統的訓練方法和訓練系統。

為解決上述問題，本發明一實施例揭露一種機器學習系統的訓練方法，利用多個樣本資料對機器學習系統進行訓練，所述訓練方法包括：

獲得多個樣本集合，每個樣本集合包括對應取樣時間段內的樣本資料；

根據每一個樣本集合對應的取樣時間段，設置該樣本集合對應的取樣率；

獲得多個根據取樣率取樣後的樣本集合；

分別確定所述多個取樣後的樣本集合的重要程度值；

利用該重要程度值修正所述多個取樣後的樣本集合中的每一個樣本資料，獲得修正後的樣本資料；

將每一個所述修正後的樣本資料輸入機器學習系統，對該機器學習系統進行訓練。

本發明另一實施例揭露一種機器學習系統的訓練系

第 106104294 號

民國 111 年 10 月 5 日修正

統，利用多個樣本資料對機器學習系統進行訓練，其中，所述訓練系統包括：

第一獲取模組，用於獲得多個樣本集合，每個樣本集合包括對應取樣時間段內的樣本資料；

取樣率設置模組，用於根據每一個樣本集合對應的取樣時間段，設置該樣本集合對應的取樣率；

第二獲取模組，用於獲得多個根據取樣率取樣後的樣本集合；

重要程度值確定模組，用於分別設置所述多個取樣後的樣本集合的重要程度值；

樣本資料修正模組，用於利用該重要程度值修正所述多個取樣後的樣本集合中的每一個樣本資料，獲得修正後的樣本資料；

訓練模組，用於將每一個所述修正後的樣本資料輸入機器學習系統，對該機器學習系統進行訓練。

本發明實施例至少具有以下優點：本發明實施例揭露一種機器學習系統的訓練方法和訓練系統，在將樣本資料輸入機器學習系統之前對樣本資料進行處理，包括獲取根據取樣時間段劃分的樣本集合、根據取樣時間段設置每個樣本集合的取樣率、根據取樣率進行取樣、確定取樣後樣本集合的重要程度值以及利用該重要程度值將樣本資料進行修正，並將樣本資料輸入機器學習系統進行訓練，在降低機器學習系統處理的資料量的同時保證了重要資料的採用率和利用程度，在減輕機器的記憶體資源需求的同時儘

量降低對機器學習系統的學習效果的影響。

【圖式簡單說明】

圖 1 是本發明第一實施例的機器學習系統的訓練方法的流程圖。

圖 2 是本發明第二實施例的機器學習系統的訓練方法的流程圖。

圖 3 是本發明第三實施例的機器學習系統的訓練方法的流程圖。

圖 4 是本發明第四實施例的機器學習系統的訓練系統的方塊圖。

圖 5 是本發明第五實施例的機器學習系統的訓練系統的方塊圖。

圖 6 是本發明第六實施例的機器學習系統的訓練系統的方塊圖。

【實施方式】

下面將結合本發明實施例中的附圖，對本發明實施例中的技術方案進行清楚、完整地描述，顯然，所描述的實施例僅僅是本發明一部分實施例，而不是全部的實施例。基於本發明中的實施例，本領域普通技術人員所獲得的所有其他實施例，都屬於本發明保護的範圍。

本發明的核心思想之一在於，提出一種機器學習系統的訓練方法和訓練系統，利用多個樣本資料對機器學習系

統進行訓練，包括根據樣本資料的取樣時間段將樣本資料劃分為多個樣本集合；根據取樣時間段設置每一個樣本集合的取樣率；根據取樣率對每一個樣本集合取樣，並修改每一個取樣後的樣本集合對應的重要程度值；利用重要程度值修正每一個樣本資料，並將該修正後的樣本資料輸入機器學習系統，對該機器學習系統進行訓練。

第一實施例

本發明第一實施例提出一種機器學習系統的訓練方法，如圖 1 所示為本發明一實施例的機器學習系統的訓練方法的流程圖，本實施例提出的機器學習系統的訓練方法包括如下步驟：

S101，獲得多個樣本集合，每個樣本集合包括對應取樣時間段內的樣本資料；

在這一步驟中，每一個樣本資料例如為一個向量，該向量中的其中一個維度例如為該樣本資料的取樣時間。在本步驟中可以將所有樣本資料的取樣時間劃分為多個取樣時間段，並將多個樣本資料根據取樣時間段劃分為多個樣本集合，每一個樣本集合對應一個取樣時間段。

例如，所有樣本資料的取樣時間是從 1 月 24 日至 1 月 29 日，則可以將這一取樣時間劃分為多個取樣時間段，例如 1 月 29 日、1 月 27 日至 1 月 28 日、1 月 24 日至 1 月 26 日三個取樣時間段。按照上述三個取樣時間段，將樣本資料劃分為 1 月 29 日取樣的樣本集合、1 月

第 106104294 號

民國 111 年 10 月 5 日修正

27 日至 1 月 28 日取樣的樣本資料集合、1 月 24 日至 1 月 26 日取樣的樣本集合。因此，每一個樣本集合對應一個取樣時間段。

值得注意的是，上述取樣時間段可以是依據開發者或使用者設定的規則劃分，可以平均分佈或者不平均分佈，本發明並不以此為限。

S102，根據每一個樣本集合對應的取樣時間段，設置該樣本集合對應的取樣率；

在這一步驟中，可以根據取樣時間段設置對應的每一個取樣集合的樣本率。例如，可以按照取樣時間段越新的樣本集合對應的取樣率越高的原則設置取樣率。即，所述樣本集合的取樣率隨著該樣本集合對應的取樣時間段從舊到新而增加。例如在上述實例中，可以將 1 月 29 日取樣的樣本資料對應的樣本集合的取樣率設置為 1.0，將 1 月 27 日至 1 月 28 日取樣的樣本資料對應的樣本集合的取樣率設置為 0.5，將 1 月 24 日至 1 月 26 日取樣的樣本資料對應的樣本集合的取樣率設置為 0.1。

S103，獲得多個根據取樣率取樣後的樣本集合；

在這一步驟中，可以根據上一步驟中設置的取樣率，對每一個樣本集合內的樣本進行取樣。例如某一個樣本集合中包含的樣本資料為 1000 個，取樣率為 0.1，則取樣後該樣本集合中包含的樣本資料的個數為 $1000 \times 0.1 = 100$ 個。藉由取樣後，樣本集合中的樣本資料為 100 個，這 100 個樣本資料對應的集合可以稱為取樣後的樣本集合。

S104，分別確定所述多個取樣後的樣本集合的重要程度值；

在一實施例中，重要程度值可以是人為或者機器演算法設定的係數，每一個取樣後的樣本集合對應的重要程度值分別可以人為設定或藉由機器以一定規則設定。在上述步驟中，可以在該樣本集合原重要程度值的基礎上，設置新的重要程度值。

S105，利用該重要程度值修正所述多個取樣後的樣本集合中的每一個樣本資料，獲得修正後的樣本資料；

在這一步驟中，可以利用該重要程度值修正多個取樣後的樣本集合中的每一個樣本資料，獲得修正後的樣本資料；

利用該重要程度值修正每一個樣本資料，可以是將每一個向量的每一個特徵維度與重要程度值相乘，使該向量等比例放大，獲得修正後的樣本資料。

例如，該樣本集合原有的或者預設的重要程度值為 1，在這一步驟中可以修正為 2，因此某個原來為 $a(1,1,1,2,\dots\dots n)$ 的樣本資料在這一步驟中可以修正為 $a(2,2,2,4,\dots\dots 2n)$ ，即為修正後的樣本資料。

然而，正如本領域技術人員可以得知的，重要程度值並不限於人為或者機器演算法設定的係數，在其他實施例，還可以有多種方法，例如對樣本資料 $a(1,1,1,2,\dots\dots n)$ 進行數學運算， $a1=f(a)$ 等等，這裡的函數 f 可以為等比相乘函數，或者類似指數運算等等各種

第 106104294 號

民國 111 年 10 月 5 日修正

數學函數，也可以對樣本進行修正。

S106，將每一個所述修正後的樣本資料輸入機器學習系統，對該機器學習系統進行訓練。

在這一步驟中，可以將修正後的樣本資料輸入機器學習系統，對該機器學習系統進行訓練。在訓練中，首先對損失函數求導，計算出梯度，再結合初始的權重以及設置的步長，根據公式“新的權重=舊的權重+步長*梯度”藉由反覆運算的方式計算出接近最優解的權重值。

綜上所述，本發明第一實施例揭露一種機器學習系統的訓練方法，在將樣本資料輸入機器學習系統之前對樣本資料進行處理，降低資料量的同時保證了重要資料的採用率和利用程度，在減輕機器的記憶體資源需求的同時儘量降低對機器學習系統的學習效果的影響。

第二實施例

本發明第二實施例提出一種機器學習系統的訓練方法，如圖 2 所示為本發明第二實施例的機器學習系統的訓練方法的流程圖，本實施例提出的機器學習系統的訓練方法包括如下步驟：

S201，獲得多個樣本集合，每個樣本集合包括對應取樣時間段內的樣本資料；

S202，根據每一個樣本集合對應的取樣時間段，設置該樣本集合對應的取樣率；

S203，獲得多個根據取樣率取樣後的樣本集合；

上述三個步驟與第一實施例中的步驟 S101、S102、S103 相同或相似，在此不再贅述。

S204，分別確定所述多個取樣後的樣本集合的重要程度值；

步驟 S204 例如可以包括：

子步驟 S204a：基於對應的取樣率對所述取樣後的樣本集合的初始重要程度值進行修正，得到所述取樣後的樣本集合的重要程度值；

所述重要程度值和初始重要程度值為正比關係，和所述取樣後的樣本集合的取樣率為反比關係。

在子步驟 S204a 中，例如可以藉由該樣本集合原先對應的重要程度值與取樣率的比值，計算新的重要程度值。例如，可以按照下述公式初次設置每一個樣本集合的重要程度值：

$$Y1=Y/a;$$

其中 Y1 為對應於該樣本集合的設置後的重要程度值；

Y 為對應於該樣本集合的原始的重要程度值；

a 為所述樣本集合的取樣率。

舉例來說，在第一實施例所提供的實例中，如果針對 1 月 24 日至 1 月 26 日這一取樣時間段的取樣率為 0.1，並且該集合對應的重要程度值設為 0.2；針對 1 月 27 日至 1 月 28 日這一取樣時間段的取樣率為 0.5，並且該集合對應的重要程度值設為 1；針對 1 月 29 日這一取樣時間段

的取樣率為 1，並且該集合對應的重要程度值設為 5，則根據 $Y1=Y/a$ ，可以得出按照取樣時間段由舊到新排列的這三個集合的重要程度值分別為 2、2、5。

步驟 S204 例如還可以包括：

子步驟 S204b，按照預設規則，提高最新的取樣時間段對應的樣本集合的重要程度值。

在子步驟 S204b 中，這一預設規則例如可以包括：

提高後的最新的取樣時間段對應的樣本集合的重要程度值與提高前的最新的取樣時間段對應的樣本集合的重要程度值成正比，並與樣本集合的總個數成正比。

在這一子步驟中，例如可以按照下述公式再次設置最新的取樣時間段對應的樣本集合的重要程度值：

$$Z1=Z*b；$$

其中 Z1 為對應於該樣本集合的再次修改後的重要程度值；

Z 為對應於該樣本集合的初次修改後的重要程度值；

b 為樣本集合的總個數。

舉例來說，根據步驟 S204b 得出的按照取樣時間段由舊到新排列的三個樣本集合對應的重要程度值分別為 2、2、5，在這一步驟中，可以針對取樣時間段最新的取樣後的樣本集合，即第三個樣本集合，再次提升其重要程度值。

例如，可以按照下述公式再次設置最新的取樣時間段對應的樣本集合的重要程度值：

$$Z1=Z*b;$$

其中 $Z1$ 為對應於該樣本集合的再次設置後的重要程度值；

Z 為對應於該樣本集合的初次設置後的重要程度值；

b 為樣本集合的總個數。

舉例來說，在子步驟 S204a 中獲得的取樣時間段最新的樣本集合對應的初次設置後的重要程度值為 5，在這一子步驟中，可以藉由 $Z1=Z*b$ 的公式，獲取再次設置後的重要程度值為 $5*3=15$ 。

值得注意的是，子步驟 S204b 可以在子步驟 S204a 之前或之後執行，或者是單獨執行。即，子步驟 S204b 相對於子步驟 S204a 是獨立的，並不依賴於子步驟 S204a。

S205，利用該重要程度值修正所述多個取樣後的樣本集合中的每一個樣本資料，獲得修正後的樣本資料；

在這一步驟例如可以包括如下子步驟：

S205a，將每一個所述重要程度值與對應的取樣後的樣本集合中的每一個樣本資料相乘，獲得修正後的樣本資料。

S206，將每一個所述修正後的樣本資料輸入機器學習系統，對該機器學習系統進行訓練。

這一步驟可以與第一實施例中的步驟 S106 相同或相似，在此不再贅述。

綜上所述，本發明第二實施例揭露一種機器學習系統的訓練方法，在將樣本資料輸入機器學習系統之前對樣本

資料進行處理，藉由對不同樣本集合的重要程度值的設置，在降低資料量的同時保證了重要資料的採用率和利用程度，在減輕機器的記憶體資源需求的同時儘量降低對機器學習系統的學習效果的影響。

第三實施例

本發明第三實施例提出一種機器學習系統的訓練方法，如圖 2 所示為本發明第二實施例的機器學習系統的訓練方法的流程圖，本實施例提出的機器學習系統的訓練方法包括如下步驟：

S301，獲得多個樣本集合，每個樣本集合包括對應取樣時間段內的樣本資料；

S302，根據每一個樣本集合對應的取樣時間段，設置該樣本集合對應的取樣率；

S303，獲得多個根據取樣率取樣後的樣本集合；

S304，分別確定所述多個取樣後的樣本集合的重要程度值；

S305，利用該重要程度值修正所述多個取樣後的樣本集合中的每一個樣本資料，獲得修正後的樣本資料；

上述步驟 S301 至 S305 可以與第一實施例揭露的步驟 S101 至 S105 相同或相似，也可以與第二實施例揭露的步驟 S201 至 S205 相同或相似，在此不再贅述。

本實施例還可以包括如下步驟：

S306，將每一個所述修正後的樣本資料輸入機器學習

系統，對該機器學習系統進行訓練。

在這一步驟中，可以將修正後的樣本資料輸入機器學習系統，對該機器學習系統進行訓練。在訓練中，首先對損失函數求導，計算出梯度，再結合初始的權重以及設置的步長，根據公式“新的權重=舊的權重+步長*梯度”藉由反覆運算的方式計算出接近最優解的權重值。

這一步驟可以包括如下子步驟：

S306a，計算出每一個所述修正後的樣本資料的梯度；

S306b，降低每一個所述樣本資料的梯度的精度；

S306c，將降低精度後的梯度輸入所述機器學習系統，對該機器模型進行訓練。

在步驟 S306a 中，可以首先計算每一個修正後的樣本資料的梯度，該梯度為損失函數的導數，藉由對損失函數求導，可以獲得梯度。

在步驟 S306b 中，機器學習系統的訓練一般採用梯度下降法，每一台機器都需要計算出梯度。如果儲存 1 個梯度需要 8 byte（位元組），則 100 億個梯度需要 $10000000000 * 8 / 1024 / 1024 / 1024 = 74.5\text{G}$ 的儲存空間。如果將儲存一個梯度的位元組數壓縮至 4 byte，則 100 億的梯度僅僅需要 32.25G 記憶體。

在步驟 S306b 中，可以使用下述公式，減少每一個本資料的梯度的儲存位元組，以實現降低精度：

$$X1 = \text{floor} (c * X + (\text{rand} ()) / d) / c$$

其中 `floor` 為向下取整；`rand()` 為產生 0-d 之間的浮點數；`X1` 為低精度浮點數，例如為電腦儲存需要 4 個位元組的 `float`，在這裡表示減少後每一個所述樣本資料的梯度的儲存位元組；`X` 為高精度浮點數，例如為電腦儲存需要 8 個位元組的 `double`，為減少前每一個所述樣本資料的梯度的儲存位元組。

另外，藉由利用 `rand` 函數引入隨機因素，來儘量降低浮點數的累計誤差。例如，利用 $(c * X + (\text{rand}()) / d)$ 的演算法，讓 `X` 乘以一個固定的數，接著加上一個在 0-1 範圍內的浮點數，目的在於在引入隨機因素。`c` 的值是個經驗值，例如可以為 536870912。`d` 例如可以為 $2^{32}-1$ ，即 2147483647，是 `rand` 函數所能產生的上限。

藉由上述公式，可以實現將一個高精度的浮點數，轉成一個低精度的浮點數，並且盡可能減低累計誤差。

綜上所述，本發明第三實施例揭露一種機器學習系統的訓練方法，在將樣本資料輸入機器學習系統之前對樣本資料進行處理，藉由對不同樣本集合的重要程度值的設置，以及在降低梯度精度時的處理，在降低資料量的同時保證了重要資料的採用率和利用程度，在減輕機器的記憶體資源需求的同時儘量降低對機器學習系統的學習效果的影響。

第四實施例

本發明第四實施例提出一種機器學習系統的訓練系

第 106104294 號

民國 111 年 10 月 5 日修正

統，如圖 4 所示為本發明第四實施例的機器學習系統的訓練系統的方塊圖，本實施例提出的機器學習系統的訓練系統利用多個樣本資料對機器學習系統進行訓練，所述訓練系統 400 包括：

第一獲取模組 401，用於獲得多個樣本集合，每個樣本集合包括對應取樣時間段內的樣本資料；

取樣率設置模組 402，用於根據每一個樣本集合對應的取樣時間段，設置該樣本集合對應的取樣率；

第二獲取模組 403，用於獲得多個根據取樣率取樣後的樣本集合；

重要程度值確定模組 404，用於分別設置所述多個取樣後的樣本集合的重要程度值；

樣本資料修正模組 405，用於利用該重要程度值修正所述多個取樣後的樣本集合中的每一個樣本資料，獲得修正後的樣本資料；

訓練模組 406，用於將每一個所述修正後的樣本資料輸入機器學習系統，對該機器學習系統進行訓練。

較佳地，在本實施例中，所述樣本集合的取樣率隨著該樣本集合對應的取樣時間段從舊到新而增加。

綜上所述，本發明第四實施例揭露一種機器學習系統的訓練系統，在將樣本資料輸入機器學習系統之前對樣本資料進行處理，降低資料量的同時保證了重要資料的採用率和利用程度，在減輕機器的記憶體資源需求的同時儘量降低對機器學習系統的學習效果的影響。

第五實施例

本發明第五實施例提出一種機器學習系統的訓練系統，如圖 5 所示為本發明第五實施例的機器學習系統的訓練系統的方塊圖，本實施例提出的機器學習系統的訓練系統利用多個樣本資料對機器學習系統進行訓練，所述訓練系統 500 包括：

第一獲取模組 501，用於獲得多個樣本集合，每個樣本集合包括對應取樣時間段內的樣本資料；

取樣率設置模組 502，用於根據每一個樣本集合對應的取樣時間段，設置該樣本集合對應的取樣率；

第二獲取模組 503，用於獲得多個根據取樣率取樣後的樣本集合；

重要程度值確定模組 504，用於分別設置所述多個取樣後的樣本集合的重要程度值；

樣本資料修正模組 505，用於利用該重要程度值修正所述多個取樣後的樣本集合中的每一個樣本資料，獲得修正後的樣本資料；

訓練模組 506，用於將每一個所述修正後的樣本資料輸入機器學習系統，對該機器學習系統進行訓練。

在本實施例中，所述樣本資料修正模組 505 用於：

將每一個所述重要程度值與對應的取樣後的樣本集合中的每一個樣本資料相乘，獲得修正後的樣本資料。

在本實施例中，所述重要程度值確定模組 504 包括：

初次修正子模組 504a，用於基於對應的取樣率對所述取樣後的樣本集合的初始重要程度值進行修正，得到所述取樣後的樣本集合的重要程度值；

所述重要程度值和初始重要程度值為正比關係，和所述取樣後的樣本集合的取樣率為反比關係。

例如，所述初次修正子模組可以按照下述公式初次設置每一個所述樣本集合的重要程度值：

$$Y1=Y/a;$$

其中 Y1 為對應於該樣本集合的設置後的重要程度值；

Y 為對應於該樣本集合的原始的重要程度值；

a 為所述樣本集合的取樣率。

在本實施例中，所述重要程度值確定模組 504 還可以包括：

二次修正子模組 504b，用於按照預設規則，提高最新的取樣時間段對應的樣本集合的重要程度值。

較佳地，所述預設規則包括：

提高後的最新的取樣時間段對應的樣本集合的重要程度值與提高前的最新的取樣時間段對應的樣本集合的重要程度值成正比，並與樣本集合的總個數成正比。

例如，可以按照下述公式再次設置最新的取樣時間段對應的樣本集合的重要程度值：

$$Z1=Z*b;$$

其中 Z1 為對應於該樣本集合的再次設置後的重要程

第 106104294 號

民國 111 年 10 月 5 日修正

度值；

Z 為對應於該樣本集合的初次設置後的重要程度值；

b 為樣本集合的總個數。

在本實施例中，所述樣本集合的取樣率隨著該樣本集合對應的取樣時間段從舊到新而增加。

綜上所述，本發明第五實施例揭露一種機器學習系統的訓練系統，在將樣本資料輸入機器學習系統之前對樣本資料進行處理，藉由對不同樣本集合的重要程度值的設置，在降低資料量的同時保證了重要資料的採用率和利用程度，在減輕機器的記憶體資源需求的同時儘量降低對機器學習系統的學習效果的影響。

第六實施例

本發明第六實施例提出一種機器學習系統的訓練系統，如圖 6 所示為本發明第六實施例的機器學習系統的訓練系統的方塊圖，本實施例提出的機器學習系統的訓練系統利用多個樣本資料對機器學習系統進行訓練，所述訓練系統 600 包括：

第一獲取模組 601，用於獲得多個樣本集合，每個樣本集合包括對應取樣時間段內的樣本資料；

取樣率設置模組 602，用於根據每一個樣本集合對應的取樣時間段，設置該樣本集合對應的取樣率；

第二獲取模組 603，用於獲得多個根據取樣率取樣後的樣本集合；

重要程度值確定模組 604，用於分別設置所述多個取樣後的樣本集合的重要程度值；

樣本資料修正模組 605，用於利用該重要程度值修正所述多個取樣後的樣本集合中的每一個樣本資料，獲得修正後的樣本資料；

訓練模組 606，用於將每一個所述修正後的樣本資料輸入機器學習系統，對該機器學習系統進行訓練。

在本實施例中，所述訓練模組 606 包括：

計算子模組 606a，用於計算出每一個所述修正後的樣本資料的梯度；

精度降低子模組 606b，用於降低每一個所述梯度的精度；

訓練子模組 606c，用於將降低精度後的梯度輸入所述機器學習系統，對該機器模型進行訓練。

在本實施例中，所述精度降低子模組 606b 用於：

利用下述公式，減少每一個梯度的儲存位元組，以實現降低精度：

$$X1 = \text{floor} (c * X + (\text{rand} ()) / d) / c$$

其中 floor 為向下取整；rand () 為產生 0-d 之間的浮點數；X1 為減少後的儲存位元組數；X 為減少前的儲存位元組數。

綜上所述，本發明第六實施例揭露一種機器學習系統的訓練系統，在將樣本資料輸入機器學習系統之前對樣本資料進行處理，藉由對不同樣本集合的重要程度值的設

置，以及在降低梯度精度時的處理，在降低資料量的同時保證了重要資料的採用率和利用程度，在減輕機器的記憶體資源需求的同時儘量降低對機器學習系統的學習效果的影響。

對於裝置實施例而言，由於其與方法實施例基本相似，所以描述的比較簡單，相關之處參見方法實施例的部分說明即可。

本說明書中的各個實施例均採用遞進的方式描述，每個實施例重點說明的都是與其他實施例的不同之處，各個實施例之間相同相似的部分互相參見即可。

本領域內的技術人員應明白，本發明實施例的實施例可提供為方法、裝置、或電腦程式產品。因此，本發明實施例可採用完全硬體實施例、完全軟體實施例、或結合軟體和硬體方面的實施例的形式。而且，本發明實施例可採用在一個或多個其中包含有電腦可用程式碼的電腦可用儲存媒體（包括但不限於磁碟記憶體、**CD-ROM**、光學記憶體等）上實施的電腦程式產品的形式。

在一個典型的配置中，所述電腦設備包括一個或多個處理器（**CPU**）、輸入/輸出介面、網路介面和記憶體。記憶體可能包括電腦可讀媒體中的非永久性記憶體，隨機存取記憶體（**RAM**）和/或非揮發性記憶體等形式，如唯讀記憶體（**ROM**）或快閃記憶體（**flash RAM**）。記憶體是電腦可讀媒體的實例。電腦可讀媒體包括永久性和非永久性、可移動和非可移動媒體可以由任何方法或技術來實現

訊號儲存。訊號可以是電腦可讀指令、資料結構、程式的模組或其他資料。電腦的儲存媒體的例子包括，但不限於相變記憶體（PRAM）、靜態隨機存取記憶體（SRAM）、動態隨機存取記憶體（DRAM）、其他類型的隨機存取記憶體（RAM）、唯讀記憶體（ROM）、電可抹除可程式設計唯讀記憶體（EEPROM）、快閃記憶體或其他記憶體技術、唯讀光碟唯讀記憶體（CD-ROM）、數位多功能光碟（DVD）或其他光學儲存、磁盒式磁帶，磁帶磁磁片儲存或其他磁性存放裝置或任何其他非傳輸媒體，可用於儲存可以被計算設備存取的訊號。按照本文中的界定，電腦可讀媒體不包括非持續性的電腦可讀媒體（transitory media），如調變的資料訊號和載波。

本發明實施例是參照根據本發明實施例的方法、終端設備（系統）、和電腦程式產品的流程圖和/或方塊圖來描述的。應理解可由電腦程式指令實現流程圖和/或方塊圖中的每一流程和/或方塊、以及流程圖和/或方塊圖中的流程和/或方塊的結合。可提供這些電腦程式指令到通用電腦、專用電腦、嵌入式處理機或其他可程式設計資料處理終端設備的處理器以產生一個機器，使得藉由電腦或其他可程式設計資料處理終端設備的處理器執行的指令產生用於實現在流程圖一個流程或多個流程和/或方塊圖一個方塊或多個方塊中指定的功能的裝置。

這些電腦程式指令也可儲存在能引導電腦或其他可程式設計資料處理終端設備以特定方式工作的電腦可讀記憶

體中，使得儲存在該電腦可讀記憶體中的指令產生包括指令裝置的製造品，該指令裝置實現在流程圖一個流程或多個流程和/或方塊圖一個方塊或多個方塊中指定的功能。

這些電腦程式指令也可裝載到電腦或其他可程式設計資料處理終端設備上，使得在電腦或其他可程式設計終端設備上執行一系列操作步驟以產生電腦實現的處理，從而在電腦或其他可程式設計終端設備上執行的指令提供用於實現在流程圖一個流程或多個流程和/或方塊圖一個方塊或多個方塊中指定的功能的步驟。

儘管已描述了本發明實施例的較佳實施例，但本領域內的技術人員一旦得知了基本創造性概念，則可對這些實施例做出另外的變更和修改。所以，所附申請專利範圍意欲解釋為包括較佳實施例以及落入本發明實施例範圍的所有變更和修改。

最後，還需要說明的是，在本文中，諸如第一和第二等之類的關係用語僅僅用來將一個實體或者操作與另一個實體或操作區分開來，而不一定要求或者暗示這些實體或操作之間存在任何這種實際的關係或者順序。而且，用語“包括”、“包含”或者其任何其他變體意在涵蓋非排他性的包含，從而使得包括一系列要素的過程、方法、物品或者終端設備不僅包括那些要素，而且還包括沒有明確列出的其他要素，或者是還包括為這種過程、方法、物品或者終端設備所固有的要素。在沒有更多限制的情況下，由語句“包括一個……”限定的要素，並不排除在包括所述要素的

過程、方法、物品或者終端設備中還存在另外的相同要素。

以上對本發明所提供的一種針對混淆指令碼語言的定位方法和系統，進行了詳細介紹，本文中應用了具體個例對本發明的原理及實施方式進行了闡述，以上實施例的說明只是用於幫助理解本發明的方法及其核心思想；同時，對於本領域的一般技術人員，依據本發明的思想，在具體實施方式及應用範圍上均會有改變之處，綜上所述，本說明書內容不應理解為對本發明的限制。

【符號說明】

400：訓練系統

401：第一獲取模組

402：取樣率設置模組

403：第二獲取模組

404：重要程度值確定模組

405：樣本資料修正模組

406：訓練模組

500：訓練系統

501：第一獲取模組

502：取樣率設置模組

503：第二獲取模組

504：重要程度值確定模組

504a：初次修正子模組

第 106104294 號

民國 111 年 10 月 5 日修正

504b：二次修正子模組

505：樣本資料修正模組

506：訓練模組

600：訓練系統

601：第一獲取模組

602：取樣率設置模組

603：第二獲取模組

604：重要程度值確定模組

605：樣本資料修正模組

606：訓練模組

606a：計算子模組

606b：精度降低子模組

606c：訓練子模組

申請專利範圍

1. 一種應用於電腦設備的機器學習系統的訓練方法，所述電腦設備利用多個樣本資料對機器學習系統進行訓練，所述訓練方法包括：

獲得多個樣本集合，每個樣本集合包括對應取樣時間段內的樣本資料；

根據每一個樣本集合對應的取樣時間段，設置該樣本集合對應的取樣率；

獲得多個根據取樣率取樣後的樣本集合；

基於對應的取樣率對所述取樣後的樣本集合的初始重要程度值進行修正，得到所述取樣後的樣本集合的重要程度值；

所述重要程度值和初始重要程度值為正比關係，和所述取樣後的樣本集合的取樣率為反比關係；

利用該重要程度值修正所述多個取樣後的樣本集合中的每一個樣本資料，獲得修正後的樣本資料；

將每一個所述修正後的樣本資料輸入機器學習系統，對該機器學習系統進行訓練；

其中所述重要程度值包括機器學習演算法設定的係數或對所述樣本資料進行數學運算以對所述樣本資料進行修正的函數。

2. 如申請專利範圍第 1 項所述的機器學習系統的訓練方法，其中，所述利用該重要程度值修正所述多個取樣後的樣本集合中的每一個樣本資料，獲得修正後的樣本資

料的步驟包括：

將每一個所述重要程度值與對應的取樣後的樣本集合中的每一個樣本資料相乘，獲得修正後的樣本資料。

3. 如申請專利範圍第 1 項所述的機器學習系統的訓練方法，其中，所述將每一個所述修正後的樣本資料輸入機器學習系統，對該機器學習系統進行訓練的步驟包括：

計算出每一個所述修正後的樣本資料的梯度；

降低每一個所述梯度的精度；

將降低精度後的梯度輸入所述機器學習系統，對該機器模型進行訓練。

4. 如申請專利範圍第 3 項所述的機器學習系統的訓練方法，其中，所述降低每一個所述梯度的精度的步驟包括：

利用下述公式，減少每一個梯度的儲存位元組，以實現降低精度：

$$X1 = \text{floor} (c * X + (\text{rand} ()) / d) / c$$

其中 floor 為向下取整；rand () 為產生 0-d 之間的浮點數；X1 為減少後的儲存位元組數；X 為減少前的儲存位元組數。

5. 如申請專利範圍第 1 項所述的機器學習系統的訓練方法，其中，所述分別設置所述多個取樣後的樣本集合的重要程度值步驟還包括：

按照預設規則，提高最新的取樣時間段對應的樣本集合的重要程度值。

6. 如申請專利範圍第 5 項所述的機器學習系統的訓練方法，其中，所述預設規則包括：

提高後的最新的取樣時間段對應的樣本集合的重要程度值與提高前的最新的取樣時間段對應的樣本集合的重要程度值成正比，並與樣本集合的總個數成正比。

7. 如申請專利範圍第 1 項所述的機器學習系統的訓練方法，其中，在根據每一個樣本集合對應的取樣時間段，設置該樣本集合對應的取樣率的步驟中，所述樣本集合的取樣率隨著該樣本集合對應的取樣時間段從舊到新而增加。

8. 一種應用於電腦設備的機器學習系統的訓練系統，所述電腦設備利用多個樣本資料對機器學習系統進行訓練，所述訓練系統包括：

第一獲取模組，用於獲得多個樣本集合，每個樣本集合包括對應取樣時間段內的樣本資料；

取樣率設置模組，用於根據每一個樣本集合對應的取樣時間段，設置該樣本集合對應的取樣率；

第二獲取模組，用於獲得多個根據取樣率取樣後的樣本集合；

重要程度值確定模組，用於分別設置所述多個取樣後的樣本集合的重要程度值；

樣本資料修正模組，用於利用該重要程度值修正所述多個取樣後的樣本集合中的每一個樣本資料，獲得修正後的樣本資料；

訓練模組，用於將每一個所述修正後的樣本資料輸入機器學習系統，對該機器學習系統進行訓練；

其中，所述重要程度值確定模組包括：

初次修正子模組，用於基於對應的取樣率對所述取樣後的樣本集合的初始重要程度值進行修正，得到所述取樣後的樣本集合的重要程度值；

所述重要程度值和初始重要程度值為正比關係，和所述取樣後的樣本集合的取樣率為反比關係；

其中所述重要程度值包括機器學習演算法設定的係數或對所述樣本資料進行數學運算以對所述樣本資料進行修正的函數。

9. 如申請專利範圍第 8 項所述的機器學習系統的訓練系統，其中，所述樣本資料修正模組用於：

將每一個所述重要程度值與對應的取樣後的樣本集合中的每一個樣本資料相乘，獲得修正後的樣本資料。

10. 如申請專利範圍第 8 項所述的機器學習系統的訓練系統，其中，所述訓練模組包括：

計算子模組，用於計算出每一個所述修正後的樣本資料的梯度；

精度降低子模組，用於降低每一個所述梯度的精度；

訓練子模組，用於將降低精度後的梯度輸入所述機器學習系統，對該機器模型進行訓練。

11. 如申請專利範圍第 10 項所述的機器學習系統的訓練系統，其中，所述精度降低子模組用於：

利用下述公式，減少每一個梯度的儲存位元組，以實現降低精度：

$$X1 = \text{floor} (c * X + (\text{rand} ()) / d) / c$$

其中 floor 為向下取整；rand () 為產生 0-d 之間的浮點數；X1 為減少後的儲存位元組數；X 為減少前的儲存位元組數。

12. 如申請專利範圍第 8 項所述的機器學習系統的訓練系統，其中，所述重要程度值確定模組還包括：

二次修正子模組，用於按照預設規則，提高最新的取樣時間段對應的樣本集合的重要程度值。

13. 如申請專利範圍第 12 項所述的機器學習系統的訓練系統，其中，所述預設規則包括：

提高後的最新的取樣時間段對應的樣本集合的重要程度值與提高前的最新的取樣時間段對應的樣本集合的重要程度值成正比，並與樣本集合的總個數成正比。

14. 如申請專利範圍第 8 項所述的機器學習系統的訓練系統，其中，所述取樣率設置模組用於，將所述樣本集合的取樣率設置為隨著該樣本集合對應的取樣時間段從舊到新而增加。

圖式

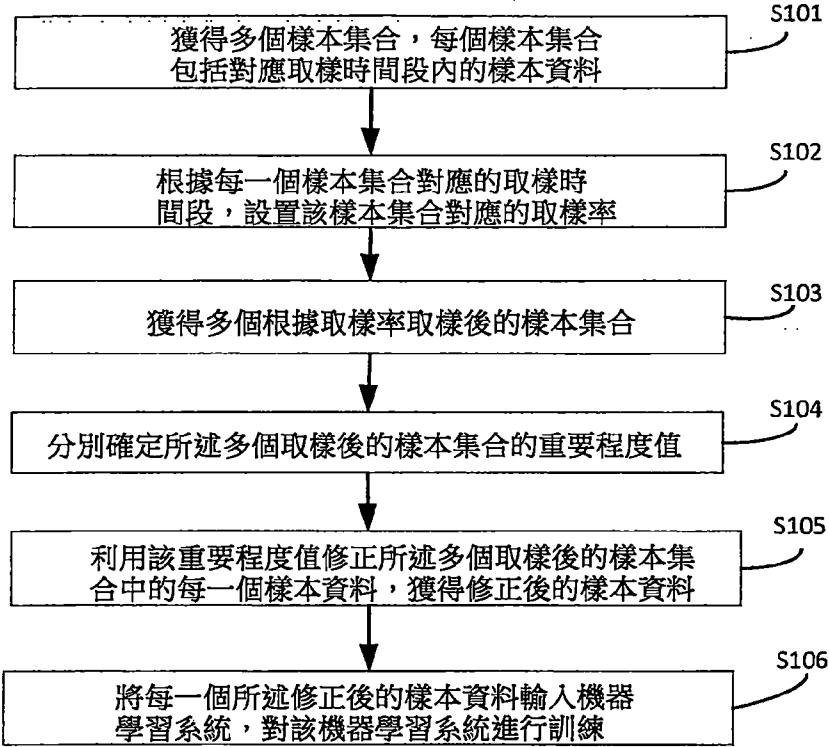


圖 1

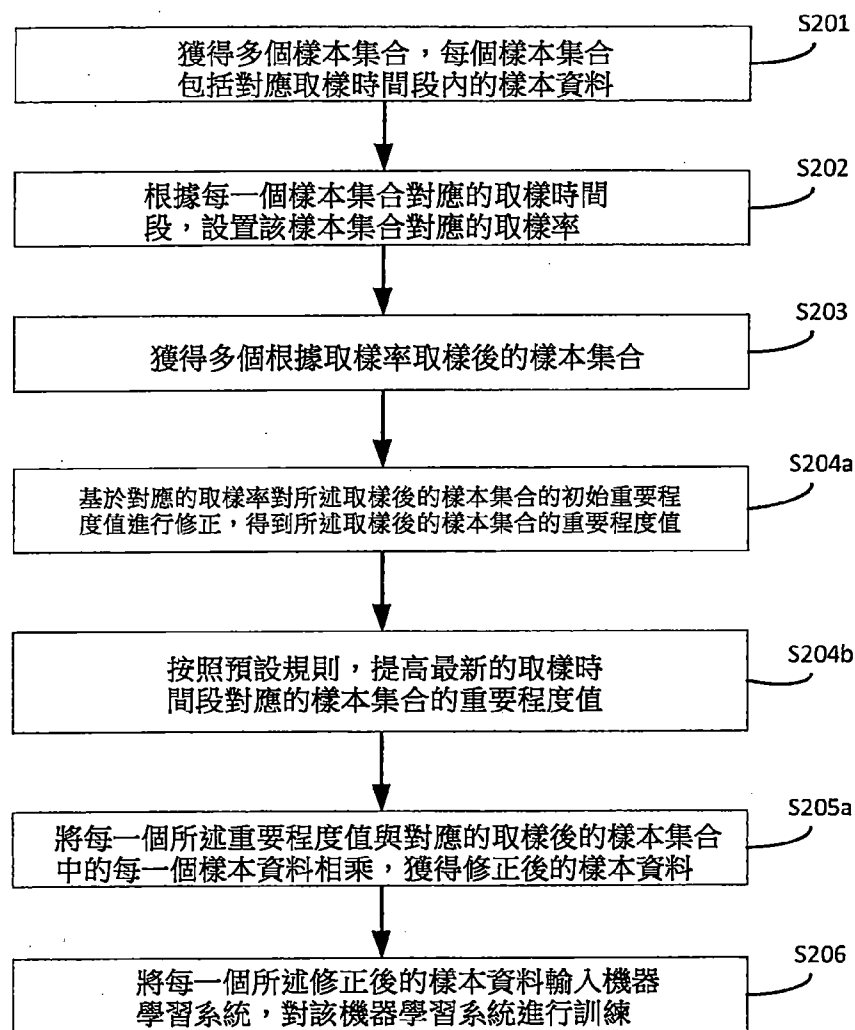


圖 2

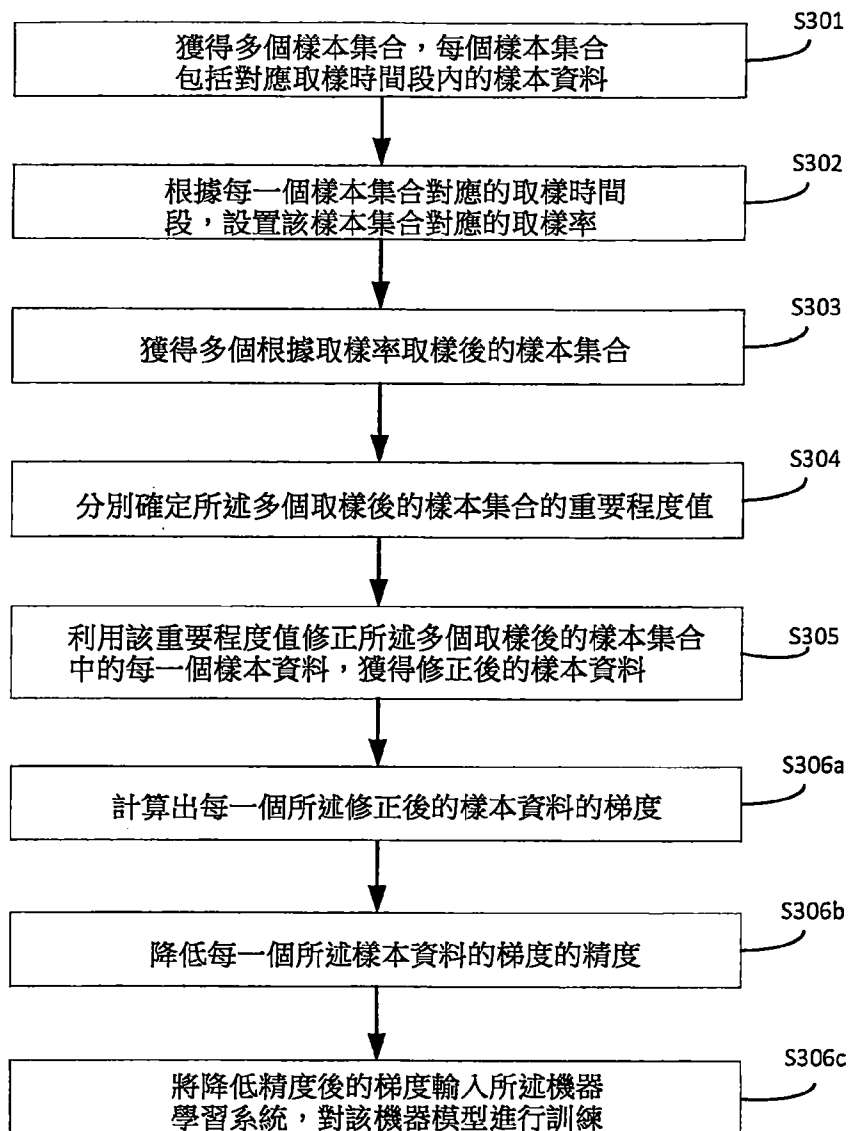


圖 3

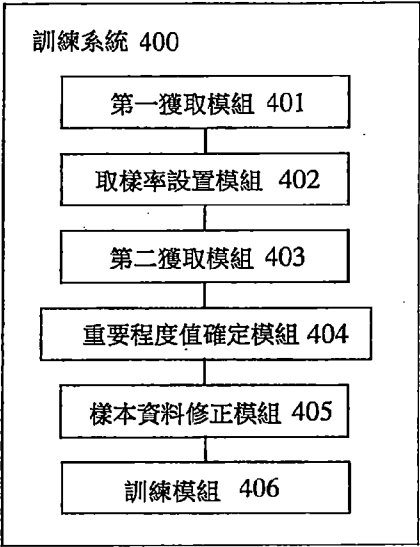


圖 4

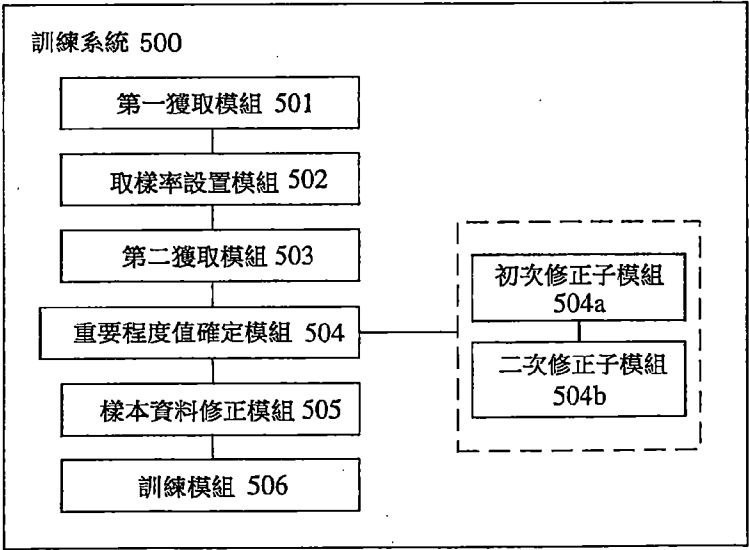


圖 5

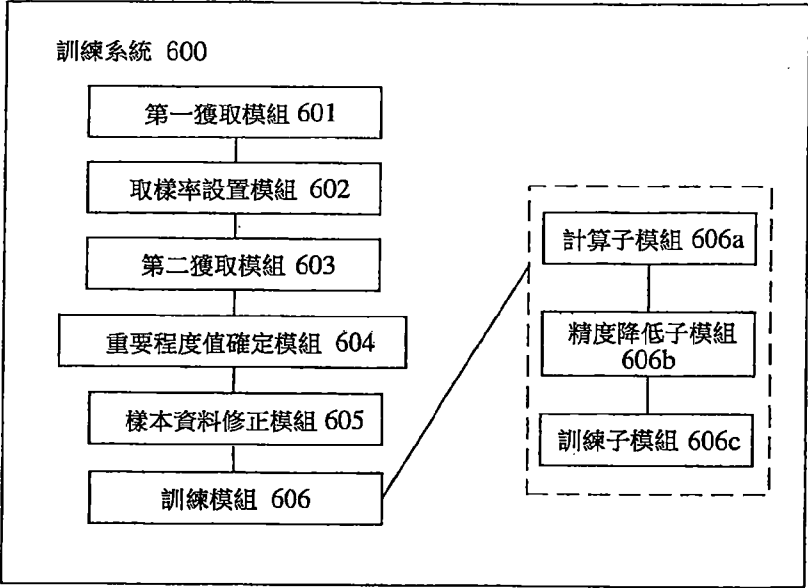


圖 6