



(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 I694341 B

(45)公告日：中華民國 109 (2020) 年 05 月 21 日

(21)申請案號：107132422

(22)申請日：中華民國 107 (2018) 年 09 月 14 日

(51)Int. Cl. : G06F17/27 (2006.01)

G06N20/00 (2019.01)

G06F17/50 (2006.01)

(30)優先權：2017/11/20 中國大陸

201711160012.0

(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：陳凡 (CN) ； 齊翔 (CN) ； 王德勝 (CN) ； 王韓彬 (CN) ； 郭棋林 (CN)

(74)代理人：林志剛

(56)參考文獻：

TW I560635

CN 104166706A

CN 107016406A

US 2016/0034788A1

審查人員：張耕誌

申請專利範圍項數：17 項 圖式數：6 共 39 頁

(54)名稱

一種資料樣本標籤處理方法、裝置及電腦設備

(57)摘要

揭露一種資料樣本標籤處理方法及裝置。一種資料樣本標籤處理方法包括：獲得不確定標籤樣本集和已確定標籤樣本集；利用以下步驟進行迭代處理，直到學習標籤的準確率達到預設的要求：對當前的不確定標籤樣本集和已確定標籤樣本集進行合併，訓練得到預測模型；將不確定標籤樣本輸入預測模型，將模型輸出的預測值確定為不確定標籤樣本的學習標籤；獲得帶有學習標籤的抽樣子集，對抽樣子集的學習標籤進行校驗以獲得學習標籤的準確率，如果學習標籤的準確率未達到預設的要求，則將標籤校驗結果確定為抽樣子集的標註結果，並將抽樣子集從不確定標籤樣本集移至已確定標籤樣本集；迭代結束後，將不確定標籤樣本當前的學習標籤確定為標註結果。

指定代表圖：

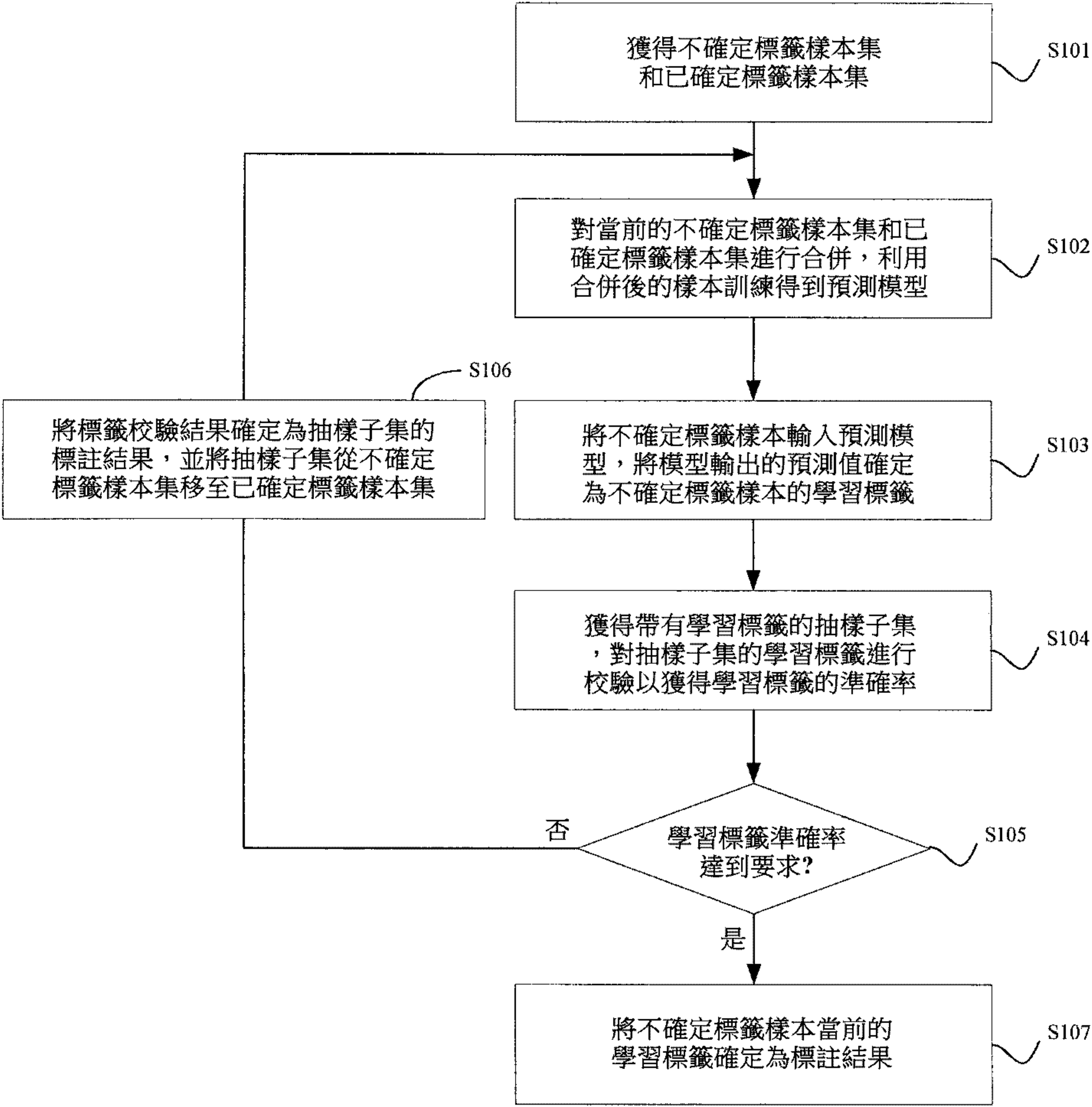


圖 1

I694341

【發明摘要】**【中文發明名稱】**

一種資料樣本標籤處理方法、裝置及電腦設備

【中文】

揭露一種資料樣本標籤處理方法及裝置。一種資料樣本標籤處理方法包括：獲得不確定標籤樣本集和已確定標籤樣本集；利用以下步驟進行迭代處理，直到學習標籤的準確率達到預設的要求：對當前的不確定標籤樣本集和已確定標籤樣本集進行合併，訓練得到預測模型；將不確定標籤樣本輸入預測模型，將模型輸出的預測值確定為不確定標籤樣本的學習標籤；獲得帶有學習標籤的抽樣子集，對抽樣子集的學習標籤進行校驗以獲得學習標籤的準確率，如果學習標籤的準確率未達到預設的要求，則將標籤校驗結果確定為抽樣子集的標註結果，並將抽樣子集從不確定標籤樣本集移至已確定標籤樣本集；迭代結束後，將不確定標籤樣本當前的學習標籤確定為標註結果。

【指定代表圖】第(1)圖。

【代表圖之符號簡單說明】無

【特徵化學式】無

【發明說明書】

【中文發明名稱】

一種資料樣本標籤處理方法、裝置及電腦設備

【技術領域】

本說明書實施例係關於資料分析之技術領域，更明確地係關於一種資料樣本標籤處理方法及裝置。

【先前技術】

在機器學習領域，利用大量的資料樣本進行訓練，可以獲得各種形式的資料模型以解決實際問題。機器學習可分為有監督學習和無監督學習，其中監督學習是利用一組已知類別的資料樣本來調整預測模型的參數、使其達到性能要求的過程。監督學習使用的訓練樣本均為已標記樣本，即每條樣本同時包含“特徵值”和“標籤值”。

有監督學習和無監督學習分別可以適用於一定的需求場景，然而在實際應用中，經常會遇到需要採用有監督學習解決問題、但是資料樣本標籤不準確甚至無標籤的情況。理論上雖然可以採用人工的方式對分別對每條資料樣本的標籤進行糾正或重新標註，然而在大資料的應用場景下，這種純人工的處理方式是不現實的。因此，如何對不確定標籤樣本實現高效、準確的標註，已經成為行業內備受關注的問題。

【發明內容】

針對上述技術問題，本說明書實施例提供一種資料樣本標籤處理方法及裝置，技術方案如下：

根據本說明書實施例的第一態樣，提供一種資料樣本標籤處理方法，該方法包括：

獲得不確定標籤樣本集和已確定標籤樣本集；利用以下步驟進行迭代處理，直到學習標籤的準確率達到預設的要求：

對當前的不確定標籤樣本集和已確定標籤樣本集進行合併，利用合併後的樣本訓練得到預測模型；

將不確定標籤樣本集中的不確定標籤樣本輸入預測模型，將模型輸出的預測值確定為不確定標籤樣本的學習標籤；

根據當前的不確定標籤樣本集，獲得帶有學習標籤的抽樣子集，對抽樣子集的學習標籤進行校驗以獲得學習標籤的準確率，如果學習標籤的準確率未達到預設的要求，則將標籤校驗結果確定為抽樣子集的標註結果，並將抽樣子集從不確定標籤樣本集移至已確定標籤樣本集；

迭代結束後，將不確定標籤樣本當前的學習標籤確定為標註結果。

根據本說明書實施例的第二態樣，提供一種資料模型訓練方法，該方法包括：

獲得不確定標籤樣本集和已確定標籤樣本集；利用以下步驟進行迭代處理，直到學習標籤的準確率達到預設的

要求：

對當前的不確定標籤樣本集和已確定標籤樣本集進行合併，利用合併後的樣本訓練得到預測模型；

將所述不確定標籤樣本集中的不確定標籤樣本輸入所述預測模型，將模型輸出的預測值確定為不確定標籤樣本的學習標籤；

對帶有學習標籤的不確定標籤樣本集進行抽樣得到抽樣子集，對抽樣子集的學習標籤進行校驗以獲得學習標籤的準確率，如果學習標籤的準確率未達到預設的要求，則將標籤校驗結果確定為抽樣子集的標註結果，並將抽樣子集從不確定標籤樣本集移動至已確定標籤樣本集；

迭代結束後，將當前的預測模型確定為適用於所述不確定標籤樣本集的預測模型。

根據本說明書實施例的第三態樣，提供一種資料樣本標籤處理裝置，該裝置包括：

輸入模組，用於獲得不確定標籤樣本集和已確定標籤樣本集；

學習模組，用於對當前的不確定標籤樣本集和已確定標籤樣本集進行合併，利用合併後的樣本訓練得到預測模型；將不確定標籤樣本集中的不確定標籤樣本輸入預測模型，將模型輸出的預測值確定為不確定標籤樣本的學習標籤；

校驗模組，用於根據當前的不確定標籤樣本集，獲得帶有學習標籤的抽樣子集，對抽樣子集的學習標籤進行校

驗以獲得學習標籤的準確率，如果學習標籤的準確率未達到預設的要求，則將標籤校驗結果確定為抽樣子集的標註結果，並將抽樣子集從不確定標籤樣本集移至已確定標籤樣本集；

所述學習模組和所述校驗模組相互配合實現迭代處理，直到學習標籤的準確率達到預設的要求；

輸出模組，用於在迭代結束後，將不確定標籤樣本當前的學習標籤確定為標註結果。

根據本說明書實施例的第四態樣，提供一種資料模型訓練裝置，該裝置包括：

輸入模組，用於獲得不確定標籤樣本集和已確定標籤樣本集；

學習模組，用於對當前的不確定標籤樣本集和已確定標籤樣本集進行合併，利用合併後的樣本訓練得到預測模型；將不確定標籤樣本集中的不確定標籤樣本輸入預測模型，將模型輸出的預測值確定為不確定標籤樣本的學習標籤；

校驗模組，用於根據當前的不確定標籤樣本集，獲得帶有學習標籤的抽樣子集，對抽樣子集的學習標籤進行校驗以獲得學習標籤的準確率，如果學習標籤的準確率未達到預設的要求，則將標籤校驗結果確定為抽樣子集的標註結果，並將抽樣子集從不確定標籤樣本集移至已確定標籤樣本集；

所述學習模組和所述校驗模組相互配合實現迭代處

理，直到學習標籤的準確率達到預設的要求；

輸出模組，用於在迭代結束後，將當前的預測模型確定為適用於所述不確定標籤樣本集的預測模型。

本說明書實施例所提供的技術方案，在已擁有大量已確定標籤樣本集的情況下，首先通過對已知標籤樣本資訊的學習，對不確定標籤樣本的標籤進行初步標註，以得到不確定標籤樣本的學習標籤，然後對學習標籤進行抽樣校驗，並將校驗後的結果回饋至學習階段，使其利用校驗後的結果重新進行學習。通過上述方式來不斷改善學習結果，直到滿足需求。應用上述方案，不僅可以在僅付出少量校驗成本的情況下，實現對不確定標籤的標註或糾正，還能夠針對不確定標籤樣本的自有特徵，得到可適用於不確定標籤樣本的預測模型。

應當理解的是，以上的一般描述和後文的細節描述僅是示例性和解釋性的，並不能限制本說明書實施例。

此外，本說明書實施例中的任一實施例並不需要達到上述的全部效果。

【圖式簡單說明】

為了更清楚地說明本說明書實施例或現有技術中的技術方案，下面將對實施例或現有技術描述中所需要使用的附圖作簡單地介紹，顯而易見地，下面描述中的附圖僅僅是本說明書實施例中記載的一些實施例，對於本領域普通技術人員來講，還可以根據這些附圖獲得其他的附圖。

圖 1 是本說明書實施例的資料樣本標籤處理方法的流程示意圖；

圖 2 是本說明書實施例的整體設計架構示意圖；

圖 3 是本說明書實施例的不確定標籤樣本生成示意圖；

圖 4 是本說明書實施例的資料模型訓練方法的流程示意圖；

圖 5 是本說明書實施例的標籤處理裝置及訓練裝置的結構示意圖；以及

圖 6 是用於配置本說明書實施例裝置的一種設備的結構示意圖。

【實施方式】

為了使本領域技術人員更好地理解本說明書實施例中的技術方案，下面將結合本說明書實施例中的附圖，對本說明書實施例中的技術方案進行詳細地描述，顯然，所描述的實施例僅僅是本說明書的一部分實施例，而不是全部的實施例。基於本說明書中的實施例，本領域普通技術人員所獲得的所有其他實施例，都應當屬於保護的範圍。

在對資料進行分類預測等應用場景（例如垃圾資訊識別、欺詐行為識別等），理想的情況是所有的資料樣本都帶有標籤，從而通過有監督學習得到預測模型。雖然對資料樣本進行標記的代價較高，但是為了得到性能較好的預測模型，在實現需求的初期也不可避免地要投入成本，以

得到數量足夠的有標籤樣本。

當訓練得到的模型性能達到需求後，就可以投入使用以對未知資料進行分類預測了。但是實際應用中，可能會出現的情況是：由於業務領域、應用場景等方面的差別，導致未知資料與訓練模型時所使用資料樣本的特徵並不完全相同，因此使用已有的模型對這些未知資料進行預測時，經常達不到期望的效果。例如，對於“識別垃圾資訊”這一需求，根據電商平臺的交易評論內容訓練出的識別模型，如果用來識別視訊網站評論內容中的垃圾資訊，很有可能會出現大量的識別錯誤。

另一種類似的問題是：給定一批“新類型”的資料，希望能夠據此訓練出適用於這批資料的識別模型。然而這批資料可能是完全不帶有任何標籤，也可能是全部或部分帶有標籤、但是無法保證這些標籤的準確性（例如可能是粗糙標註等原因導致）。例如，給定一批視訊網站中的評論內容，希望訓練出適用於網站中垃圾評論內容識別的模型。那麼，根據現有技術，如果要滿足模型訓練的要求，又需要花費大量成本對這批新類型資料進行標註。

針對上述需求，本說明書實施例提供一種資料樣本標籤處理方案，該方案的設計架構如圖1所示，具體架構說明如下：

1) 輸入部分：

以“不確定標籤樣本集”和“已確定標籤樣本集”作為輸入資訊。

其中“不確定標籤樣本集”中的樣本可能是完全不帶有任何標籤，也可能是全部或部分帶有標籤、但是無法保證這些標籤的準確性。

“已確定標籤樣本集”指當前已擁有的、攜帶可靠標籤的樣本集合。具體的標籤來源這裡不做限定，總之可以將這部分視為已知資訊。

實際情況中，“不確定標籤樣本集”和“已確定標籤樣本集”的整體使用需求相同，但又具有不同的資料特徵。例如前面例子中的“視訊網站評論內容”和“交易平臺評論內容”，都是用於識別垃圾評論內容，但是不同領域的評論內容又各有特色。則前者相當於“未確定標籤樣本集”，後者相當於“已確定標籤樣本集”。

為便於描述，後續將分別以符號“U”和“L”表示“不確定標籤樣本集”和“已確定標籤樣本集”。

2) 輸出部分：

以“U的標註結果”和“適用於U的預測模型”作為輸出資訊，當然在實際應用中，可能只需要上述兩種輸出資訊中的一種。

3) 學習模組：

學習模組又可以進一步分為訓練和標註兩個子模組（圖中未示出）：

a訓練子模組：

以U和L的合併結果作為依據，通過訓練得到預測模型。其中根據U的具體情況不同，將採用不同訓練方式：

如果U中樣本均帶有標籤，則對U和L進行合併後，使用有監督學習的方式訓練得到預測模型。

如果U中樣本完全不帶有任何標籤、或者僅部分樣本帶有標籤，則對U和L進行合併後，使用半監督學習的方式訓練得到預測模型。

可見，無論採用哪種訓練方式，由於訓練資料覆蓋了兩類資料的不同特徵，因此訓練出的模型都能夠適用於兩種不同資料類型的預測。

b標註子模組：

利用訓練子模組訓練出的模型，對U中的資料進行預測，將預測結果定義為“學習標籤”。

4) 校驗模組：

由於學習模組的訓練過程中，使用了大量了“不確定標籤樣本”，因此初期訓練得到的模型效果很一般是不理想的（除非U中有大量標註結果正確的樣本、而且這個結論是預先已知的，但是這種情況也就沒有必要使用合併的樣本進行訓練了），因此需要對學習標籤進行校驗。

為保證校驗結果的可靠性，這裡可以採用人工參與的方式進行校驗，校驗模組可提供樣本資料、標註結果等資訊的顯示功能，並且提供標註或糾錯等操作介面，以方便相關人員進行校驗。當然，在能夠保證校驗結果可靠性的前提下，還也可以採用其他方式實現校驗，本申請對此並不進行限定。

值得說明的是，由於校驗模組的功能僅是從整體上評

估標註結果是否理想，因此這裡並不需要對所有的學習標籤都進行校驗，只需對少量抽樣資料進行校驗即可，從而實現校驗代價的節省。

如果評估結果不理想，則需要觸發新一輪的標籤學習。另一方面，從“不確定標籤樣本”中抽樣出的資料，經過標籤校驗後，就可以當作“確定標籤樣本”使用，因此將校驗結果回饋給標籤學習系統後，能夠令每次標籤學習的準確率不斷趨於優化。兩個模組通過上述方式配合，可以進行多次再學習，直到校驗結果滿足需求。

基於上述設計方案，本說明書進一步提供相應的資料樣本標籤處理方法，參見圖2所示，該方法可以包括以下步驟：

S101，獲得不確定標籤樣本集和已確定標籤樣本集；

為描述方便，在本實施例中，仍以符號“U”和“L”表示“不確定標籤樣本集”和“已確定標籤樣本集”。

如前所述，U中的樣本可能是完全不帶有任何標籤，也可能是全部或部分帶有標籤、但是無法保證這些標籤的準確性。而L則指代當前已擁有的、攜帶可靠標籤的樣本集合。U和L的整體使用需求相同，但又具有不同的資料特徵。

根據前面的描述可知，訓練模型時，採用U和L的合併結果作為訓練樣本，而模型訓練的一個重要需求是：使得模型能夠適用於U和L兩種不同的資料類型的預測，這

就要求U和L都要達到一定的樣本數量，而且U和L的比例相差不能過於懸殊。由於單獨利用L已經能夠單獨訓練出性能滿足需求的模型，因此這裡可以認為L中樣本的絕對數量是足夠的；但是U的樣本數量則具有很大的不確定性，如果U中的樣本數量過少，則在無法在訓練過程中提供足夠的U的資料特徵，進而導致訓練出的模型對無法更好地適應對U類資料的預測。

如果U中的樣本數量不足，則可以使用GAN（Generative Adversarial Networks，生成式對抗網路），模擬U的情況再生成一部分樣本。

GAN是一種可以根據已有的真實樣本構建出新樣本的技術，GAN由生成模型（generative model）和判別模型（discriminative model）組成。生成模型的功能是捕捉已有樣本資料的分佈，用服從某一分佈（例如均勻分佈，高斯分佈等）的雜訊生成類似真實樣本據的新樣本，追求效果是越像真實樣本越好；判別模型是一個二分類器，用於判斷一個樣本是真實樣本還是生成樣本。

在GAN的訓練過程中固定一方，更新另一方的網路權重，交替迭代，在這個過程中，生成模型和判別模型雙方都極力優化自己的網路，從而形成競爭對抗，直到雙方達到一個動態的平衡，此時生成模型恢復了訓練資料的分佈（造出了和真實樣本一模一樣的樣本），判別模型也無法再判斷出是真實樣本還是生成樣本。

因此，假設初始給定的待處理樣本（本說明書中以 U_0

表示) 數量無法滿足訓練需求，則可以將 U_0 輸入 GAN，由 GAN 輸出與 U_0 同分佈的生成樣本集 (本說明書中以 U_G 表示)；然後將 U_0 與 U_G 進行合併，如圖 3 所示，即有以下關係：

$$U = U_0 + U_G$$

可以理解的是，本說明書中的“同分佈”，並不是嚴格數學意義上的同分佈，而是 GAN 所模擬出的同分佈。

由於 GAN 既可以生成有標籤樣本，也可以生成無標籤樣本，那麼可以根據 U_0 的不同情況，採用不同的樣本生成方式：

如果 U_0 中的樣本均為有標籤樣本，則可以有兩種處理方式：

1) 將 U_0 的特徵部分和標籤部分均輸入 GAN，得到帶有標籤的 U_G ，這種情況下， U 中的樣本也均為有標籤樣本。

2) 僅將 U_0 的特徵部分輸入 GAN，得到不帶標籤的 U_G ，這種情況下， U 中僅部分樣本帶有標籤。

如果 U_0 中的樣本中的樣本均為無標籤樣本、或僅部分樣本帶有標籤，則可以將 U_0 的特徵部分輸入 GAN，得到不帶標籤的 U_G ，這種情況下， U 中樣本的標籤攜帶情況與 U_0 一致。

需要生成 U_G 的樣本數量可以根據訓練需求確定，這裡的訓練既包括對樣本絕對數量的需求、也包括對樣本相對數量的需求。一般而言，希望 U 與 L 的比例不低於 1：4，

當然 U 也可以比 L 更大，該比例需求可以根據實際情況設計，本申請對此不需要進行限定。

$S102$ ，對當前的不確定標籤樣本集和已確定標籤樣本集進行合併，利用合併後的樣本訓練得到預測模型；

使用合併後的樣本集 S （其中 $S=U+L$ ）進行模型訓練，由於 L 中的樣本都是帶有標籤的，因此根據 U 中樣本的標籤攜帶情況將決定 S 中樣本的標籤攜帶情況。

如果 U 中樣本均帶有標籤，則 S 中樣本也均帶有標籤，此時訓練需求轉化為有監督學習問題。

如果 U 中樣本均不帶有標籤、或部分帶有標籤，則 S 中樣本也是部分帶有標籤的，此時訓練需求轉化為半監督學習問題。半監督學習是也一種利用部分有標籤樣本以及部分無標籤樣本實現模型訓練的技術，值得注意的是，現有技術中，半監督學習所針對的有標籤樣本和無標籤樣本是同一類型的資料樣本，而本申請中 U 和 L 是兩套並不完全一致的樣本，因此嚴格意義上講與半監督學習的傳統應用場景有所區別。由於其整體需求相同，因此在演算法層面仍然可以使用半監督學習演算法，但是其訓練結果需要配合後續的校驗步驟多次調整才能滿足應用需求。

根據具體的應用場景不同，可以選用不同形式的模型以及相應的學習演算法，對於本說明書對此並不限定。例如，對於文本識別應用，可以通過構建基於 **RNN**（**Recurrent neural Network**，遞迴類神經網路）深度學習模型訓練文本資料。

S103，將不確定標籤樣本集中的不確定標籤樣本輸入預測模型，將模型輸出的預測值確定為不確定標籤樣本的學習標籤；

對於 S102 所產出的預測模型，可以將 U 中的樣本輸入該模型，在本說明書中，將模型輸出的預測值稱為樣本的“學習標籤”，值得注意的是，該學習標籤與 U 中樣本是否帶有標籤或標籤是否準確並無必然關聯。

需要說明的是，這裡的“不確定標籤樣本”既可以是 U 中的全部樣本，也可用是 U 中樣本的一部分。

例如，如果在 S101 採用了 GAN 生成新樣本，則本步驟中可以僅將當前 U 中屬 U_0 的那部分樣本輸入預測模型。這樣處理的原因是，相對於 U_G 而言， U_0 才是真實的資料，後續對這部分資料進行校驗的意義更大，而且從“標註”需求而言，也只有 U_0 才是真正需要進行標註處理的對象。這裡需要明確的是“屬於 U_0 的那部分樣本”並不同於 U_0 ，這是因為隨著整個方案的迭代，U 的規模是逐步縮減的，相應地“屬於 U_0 的那部分樣本”也會逐步變小。

另外，在本步驟中也可以對 U（或 U_0 ）進行抽樣，僅將抽樣結果輸入預測模型，從而得到抽樣結果的學習標籤。抽樣的目的是降低校驗的代價，可以在本步驟實現，也可以在後續步驟中實現。

S104，根據當前的不確定標籤樣本集，獲得帶有學習標籤的抽樣子集，對抽樣子集的學習標籤進行校驗以獲得學習標籤的準確率；

如果在 S103 中未作抽樣處理，則在本步驟中，對所有帶有學習標籤的樣本進行抽樣，得到抽樣子集；如果在 S103 中已作抽樣處理，則在本步驟中可以直接使用 S103 輸出的結果作為抽樣子集。

抽樣的數量可以根據實際需求確定，一般綜合考慮以下幾個因素：

1) 是否有足夠的代表性：抽樣數量越大，代表性越好。

2) 對校驗代價的影響：抽樣數量越小，校驗代價越低。

3) 對迭代速度的影響：抽樣數量越大，則每次校驗後回饋給下一次學習的有用資訊越多，相應也會提高整體方案的迭代速度。

在實際應用中，也可以在迭代過程中使用動態的抽樣率，例如隨著迭代的進行，模型性能逐漸趨於穩定，可以逐步降低抽樣率。當然，本領域技術人員可以根據實際需求設計抽樣方案，本說明書對此不做限定。

S105，判斷學習標籤的準確率是否達到預設的要求，如果是則繼續執行 S107，否則執行 S106 後返回 S102；

S106，將標籤校驗結果確定為抽樣子集的標註結果，並將抽樣子集從不確定標籤樣本集移至已確定標籤樣本集；

S107，將不確定標籤樣本當前的學習標籤確定為標註結果。

假設以 U_s 表示抽樣子集，如果 U_s 的學習標籤準確率達到某個指標（例如 95%），則認為整個 U （或 U_0 ）的學習標籤是可信的，可以直接作為標註結果輸出，否則需要觸發下一輪學習。

經過校驗的 U_s 相當於已經具備了可靠的標籤，因此在下一輪學習中， U_s 可以作為“確定標籤樣本”使用，即在每次迭代之前，對 U 和 L 做以下更新：

$$U = U - U_s$$

$$L = L + U_s$$

通過 S102~S106 的迭代處理，由於具備可靠標籤的樣本逐步增加，而且這些樣本是具有“新類型資料”特徵的樣本，因此能夠令每次標籤學習的準確率不斷趨於優化，並且逐步適應“新類型資料”的預測，最終達到應用需求。

可以理解的是，迭代停止後的 U 已經和初始的 U 不同，因此最終的輸出的標註結果應包括“當前 U 中樣本最新的學習標籤”以及“歷次迭代過程中已經過校驗確認可靠的標籤”。

此外，迭代停止後，最終的預測模型也可以作為另一項輸出資訊，如圖 4 的 S108 所示（其他步驟與圖 2 所示一致，這裡不再重複說明），該模型都能夠適用於 U 和 L 兩種不同資料類型。

應用本說明書所提供的方案，對於已擁有巨量資料以及較成熟預測模型的企業而言，能夠有效將自身的預測能

力向外部輸出，為客戶或合作夥伴提供技術服務，同時也能夠不斷豐富自身模型的預測能力。

例如在電商平臺及支付平臺的風控系統中，已經具有強大的文本識別能力，可以從使用者生成內容中識別出灌水、廣告、暴恐政和黃賭毒等資訊。一些其他行業的外部商戶也具有類似的需求，例如微博、視訊、直播等 UGC（User Generated Content，使用者生成內容）相關領域，如果這些外部商戶沒有能力對使用者生成內容樣本進行準確標註，則可以基於本說明書所提供的技術方案，結合電商平臺及支付平臺自身已有的垃圾文本資料以及識別模型，對其他行業提供的樣本資料進行學習。相對於完全人工對外部樣本進行標註或糾正的方式而言效率更高，更容易實現規模化。

相應於上述方法實施例，本說明書實施例還提供一種資料樣本標籤處理裝置或資料模型訓練裝置，參見圖5所示，該裝置可以包括：

輸入模組 110，用於獲得不確定標籤樣本集和已確定標籤樣本集；

學習模組 120，用於對當前的不確定標籤樣本集和已確定標籤樣本集進行合併，利用合併後的樣本訓練得到預測模型；將不確定標籤樣本集中的不確定標籤樣本輸入預測模型，將模型輸出的預測值確定為不確定標籤樣本的學習標籤；

校驗模組 130，用於根據當前的不確定標籤樣本集，

獲得帶有學習標籤的抽樣子集，對抽樣子集的學習標籤進行校驗以獲得學習標籤的準確率，如果學習標籤的準確率未達到預設的要求，則將標籤校驗結果確定為抽樣子集的標註結果，並將抽樣子集從不確定標籤樣本集移至已確定標籤樣本集；

學習模組 120 和校驗模組 130 相互配合實現迭代處理，直到學習標籤的準確率達到預設的要求；

輸出模組 140，可以用於在迭代結束後，將不確定標籤樣本當前的學習標籤確定為標註結果。也可以用於在迭代結束後，將當前的預測模型確定為適用於不確定標籤樣本集的預測模型。

在本說明書提供的一種具體實施方式中，若不確定標籤樣本集中的樣本均為有標籤樣本；則學習模組 120 可以具體用於：利用有監督學習演算法對合併後的樣本進行訓練，得到預測模型。

若不確定標籤樣本集中的樣本均為無標籤樣本、或僅部分樣本帶有標籤；則學習模組 120 可以具體用於：利用半監督學習演算法對合併後的樣本進行訓練，得到預測模型。

在本說明書提供的一種具體實施方式中，輸入模組 110 可以包括：

生成子模組，用於在初始獲得的待處理樣本集中樣本數量未達到預設需求的情況下，將該待處理樣本集輸入生成式對抗網路，得到與待處理樣本集同分佈的生成樣本

集；

合併子模組，用於將待處理樣本集與生成樣本集合併，得到不確定標籤樣本集。

在本說明書提供的一種具體實施方式中，學習模組120可以具體用於：

將不確定標籤樣本集中，屬於待處理樣本集的部分輸入預測模型。

在本說明書提供的一種具體實施方式中，若待處理樣本集中的樣本均為有標籤樣本，則生成子模組130可以具體用於：

將該待處理樣本集的特徵部分和標籤部分輸入生成式對抗網路，得到與待處理樣本集同分佈的生成樣本集，生成樣本集中的樣本均為有標籤樣本；

或者

將該待處理樣本集的特徵部分輸入生成式對抗網路，得到與待處理樣本集同分佈的生成樣本集，生成樣本集中的樣本均為無標籤樣本。

若待處理樣本集中的樣本均中的樣本均為無標籤樣本、或僅部分樣本帶有標籤，則生成子模組130可以具體用於：

將該待處理樣本集的特徵部分輸入生成式對抗網路，得到與待處理樣本集同分佈的生成樣本集，生成樣本集中的樣本均為無標籤樣本。

本說明書實施例還提供一種電腦設備，其至少包括儲

存器、處理器及儲存在儲存器上並可在處理器上運行的電腦程式，其中，處理器執行所述程式時實現前述的資料樣本標籤處理方法資料模型訓練方法。該方法至少包括：

獲得不確定標籤樣本集和已確定標籤樣本集；利用以下步驟進行迭代處理，直到學習標籤的準確率達到預設的要求：

對當前的不確定標籤樣本集和已確定標籤樣本集進行合併，利用合併後的樣本訓練得到預測模型；

將不確定標籤樣本集中的不確定標籤樣本輸入預測模型，將模型輸出的預測值確定為不確定標籤樣本的學習標籤；

根據當前的不確定標籤樣本集，獲得帶有學習標籤的抽樣子集，對抽樣子集的學習標籤進行校驗以獲得學習標籤的準確率，如果學習標籤的準確率未達到預設的要求，則將標籤校驗結果確定為抽樣子集的標註結果，並將抽樣子集從不確定標籤樣本集移至已確定標籤樣本集；

迭代結束後，將不確定標籤樣本當前的學習標籤確定為標註結果、或者將當前的預測模型確定為適用於所述不確定標籤樣本集的預測模型。

圖6示出了本說明書實施例所提供的一種更為具體的計算設備硬體結構示意圖，該設備可以包括：處理器1010、儲存器1020、輸入/輸出介面1030、通信介面1040和匯流排1050。其中處理器1010、儲存器1020、輸入/輸出介面1030和通信介面1040通過匯流排1050實現彼此之間

在設備內部的通信連接。

處理器 1010 可以採用通用的 CPU（Central Processing Unit，中央處理器）、微處理器、應用特定積體電路（Application Specific Integrated Circuit，ASIC）、或者一個或多個積體電路等方式實現，用於執行相關程式，以實現本說明書實施例所提供的技術方案。

儲存器 1020 可以採用 ROM（Read Only Memory，唯讀記憶體）、RAM（Random Access Memory，隨機存取記憶體）、靜態儲存設備，動態儲存設備等形式實現。儲存器 1020 可以儲存操作系統和其他應用程式，在通過軟體或者韌體來實現本說明書實施例所提供的技術方案時，相關的程式碼保存在儲存器 1020 中，並由處理器 1010 來呼叫執行。

輸入/輸出介面 1030 用於連接輸入/輸出模組，以實現資訊輸入及輸出。輸入輸出/模組可以作為組件配置在設備中（圖中未示出），也可以外接於設備以提供相應功能。其中輸入設備可以包括鍵盤、滑鼠、觸控螢幕、麥克風、各類感測器等，輸出設備可以包括顯示器、揚聲器、振動器、指示燈等。

通信介面 1040 用於連接通信模組（圖中未示出），以實現本設備與其他設備的通信互動。其中通信模組可以通過有線方式（例如 USB、纜線等）實現通信，也可以通過無線方式（例如移動網路、WIFI、藍牙等）實現通信。

匯流排 1050 包括一通路，在設備的各個組件（例如處

理器 1010、儲存器 1020、輸入/輸出介面 1030和通信介面 1040) 之間傳輸資訊。

需要說明的是，儘管上述設備僅示出了處理器 1010、儲存器 1020、輸入/輸出介面 1030、通信介面 1040以及匯流排 1050，但是在具體實施過程中，該設備還可以包括實現正常運行所必需的其他組件。此外，本領域的技術人員可以理解的是，上述設備中也可以僅包含實現本說明書實施例方案所必需的組件，而不必包含圖中所示的全部組件。

本說明書實施例還提供一種電腦可讀儲存媒體，其上儲存有電腦程式，該程式被處理器執行時實現前述的資料樣本標籤處理方法資料模型訓練方法。該方法至少包括：

獲得不確定標籤樣本集和已確定標籤樣本集；利用以下步驟進行迭代處理，直到學習標籤的準確率達到預設的要求：

對當前的不確定標籤樣本集和已確定標籤樣本集進行合併，利用合併後的樣本訓練得到預測模型；

將不確定標籤樣本集中的不確定標籤樣本輸入預測模型，將模型輸出的預測值確定為不確定標籤樣本的學習標籤；

根據當前的不確定標籤樣本集，獲得帶有學習標籤的抽樣子集，對抽樣子集的學習標籤進行校驗以獲得學習標籤的準確率，如果學習標籤的準確率未達到預設的要求，則將標籤校驗結果確定為抽樣子集的標註結果，並將抽樣

子集從不確定標籤樣本集移至已確定標籤樣本集；

迭代結束後，將不確定標籤樣本當前的學習標籤確定為標註結果、或者將當前的預測模型確定為適用於所述不確定標籤樣本集的預測模型。

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資訊儲存。資訊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）、電可抹除可編程唯讀記憶體（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，磁帶磁碟儲存或其他磁性儲存設備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資訊。按照本文中的界定，電腦可讀媒體不包括暫態媒體（transitory media），如調變的資料信號和載波。

通過以上的實施方式的描述可知，本領域的技術人員可以清楚地瞭解到本說明書實施例可借助軟體加必需的通用硬體平臺的方式來實現。基於這樣的理解，本說明書實施例的技術方案本質上或者說對現有技術做出貢獻的部分可以以軟體產品的形式體現出來，該電腦軟體產品可以儲存在儲存媒體中，如ROM/RAM、磁碟、光碟等，包括若

干指令用以使得一台電腦設備（可以是個人電腦，伺服器，或者網路設備等）執行本說明書實施例各個實施例或者實施例的某些部分所述的方法。

上述實施例闡明的系統、裝置、模組或單元，具體可以由電腦晶片或實體實現，或者由具有某種功能的產品來實現。一種典型的實現設備為電腦，電腦的具體形式可以是個人電腦、膝上型電腦、蜂巢式電話、相機電話、智慧型電話、個人數位助理、媒體播放器、導航設備、電子郵件收發設備、遊戲主機、平板電腦、可穿戴設備或者這些設備中的任意幾種設備的組合。

本說明書中的各個實施例均採用漸進的方式描述，各個實施例之間相同相似的部分互相參見即可，每個實施例重點說明的都是與其他實施例的不同之處。尤其，對於裝置實施例而言，由於其基本相似於方法實施例，所以描述得比較簡單，相關之處參見方法實施例的部分說明即可。以上所描述的裝置實施例僅僅是示意性的，其中所述作為分離部件說明的模組可以是或者也可以不是物理上分開的，在實施本說明書實施例方案時可以把各模組的功能在同一個或多個軟體和/或硬體中實現。也可以根據實際的需要選擇其中的部分或者全部模組來實現本實施例方案的目的。本領域普通技術人員在不付出創造性勞動的情況下，即可以理解並實施。

以上所述僅是本說明書實施例的具體實施方式，應當指出，對於本技術領域的普通技術人員來說，在不脫離本

說明書實施例原理的前提下，還可以做出若干改進和潤飾，這些改進和潤飾也應視為本說明書實施例的保護範圍。

【符號說明】

S101：步驟

S102：步驟

S103：步驟

S104：步驟

S105：步驟

S106：步驟

S107：步驟

110：輸入模組

120：學習模組

130：校驗模組

140：輸出模組

1010：處理器

1020：儲存器

1030：輸入/輸出介面

1040：通信介面

【發明申請專利範圍】

【第 1 項】

一種資料樣本標籤處理方法，該方法包括：

獲得不確定標籤樣本集和已確定標籤樣本集；利用以下步驟進行迭代處理，直到學習標籤的準確率達到預設的要求：

對當前的不確定標籤樣本集和已確定標籤樣本集進行合併，利用合併後的樣本訓練得到預測模型；

將不確定標籤樣本集中的不確定標籤樣本輸入預測模型，將模型輸出的預測值確定為不確定標籤樣本的學習標籤；

根據當前的不確定標籤樣本集，獲得帶有學習標籤的抽樣子集，對抽樣子集的學習標籤進行校驗以獲得學習標籤的準確率，如果學習標籤的準確率未達到預設的要求，則將標籤校驗結果確定為抽樣子集的標註結果，並將抽樣子集從不確定標籤樣本集移至已確定標籤樣本集；以及

迭代結束後，將不確定標籤樣本當前的學習標籤確定為標註結果，

其中，取決於所述不確定標籤樣本集中的樣本是否均為有標籤樣本、是否均為無標籤樣本或僅部分樣本帶有標籤，利用有監督學習演算法或半監督學習演算法對合併後的樣本進行訓練，得到預測模型。

【第 2 項】

根據請求項 1 所述的方法，所述不確定標籤樣本集中

的樣本均為有標籤樣本；所述利用合併後的樣本訓練得到預測模型包括：

利用有監督學習演算法對合併後的樣本進行訓練，得到預測模型。

【第3項】

根據請求項1所述的方法，所述不確定標籤樣本集中的樣本均為無標籤樣本、或僅部分樣本帶有標籤；所述利用合併後的樣本訓練得到預測模型包括：

利用半監督學習演算法對合併後的樣本進行訓練，得到預測模型。

【第4項】

根據請求項1所述的方法，所述獲得不確定標籤樣本集，包括：

在初始獲得的待處理樣本集中樣本數量未達到預設需求的情況下，將該待處理樣本集輸入生成式對抗網路，得到與待處理樣本集同分佈的生成樣本集；以及

將待處理樣本集與生成樣本集合併，得到不確定標籤樣本集。

【第5項】

根據請求項4所述的方法，所述將不確定標籤樣本集中的不確定標籤樣本輸入所述預測模型，包括：

將不確定標籤樣本集中，屬於待處理樣本集的部分輸入所述預測模型。

【第6項】

根據請求項4所述的方法，若所述待處理樣本集中的樣本均為有標籤樣本，則所述將該待處理樣本集輸入生成式對抗網路，得到與待處理樣本集同分佈的生成樣本集，包括：

將該待處理樣本集的特徵部分和標籤部分輸入生成式對抗網路，得到與待處理樣本集同分佈的生成樣本集，所述生成樣本集中的樣本均為有標籤樣本；

或者

將該待處理樣本集的特徵部分輸入生成式對抗網路，得到與待處理樣本集同分佈的生成樣本集，所述生成樣本集中的樣本均為無標籤樣本。

【第7項】

根據請求項4所述的方法，若所述待處理樣本集中的樣本均中的樣本均為無標籤樣本、或僅部分樣本帶有標籤，則所述將該待處理樣本集輸入生成式對抗網路，得到與待處理樣本集同分佈的生成樣本集，包括：

將該待處理樣本集的特徵部分輸入生成式對抗網路，得到與待處理樣本集同分佈的生成樣本集，所述生成樣本集中的樣本均為無標籤樣本。

【第8項】

一種資料模型訓練方法，該方法包括：

獲得不確定標籤樣本集和已確定標籤樣本集；利用以下步驟進行迭代處理，直到學習標籤的準確率達到預設的要求：

對當前的不確定標籤樣本集和已確定標籤樣本集進行合併，利用合併後的樣本訓練得到預測模型；

將所述不確定標籤樣本集中的不確定標籤樣本輸入所述預測模型，將模型輸出的預測值確定為不確定標籤樣本的學習標籤；

對帶有學習標籤的不確定標籤樣本集進行抽樣得到抽樣子集，對抽樣子集的學習標籤進行校驗以獲得學習標籤的準確率，如果學習標籤的準確率未達到預設的要求，則將標籤校驗結果確定為抽樣子集的標註結果，並將抽樣子集從不確定標籤樣本集移動至已確定標籤樣本集；以及

迭代結束後，將當前的預測模型確定為適用於所述不確定標籤樣本集的預測模型，

其中，取決於所述不確定標籤樣本集中的樣本是否均為有標籤樣本、是否均為無標籤樣本或僅部分樣本帶有標籤，利用有監督學習演算法或半監督學習演算法對合併後的樣本進行訓練，得到預測模型。

【第9項】

一種資料樣本標籤處理裝置，該裝置包括：

輸入模組，用於獲得不確定標籤樣本集和已確定標籤樣本集；

學習模組，用於對當前的不確定標籤樣本集和已確定標籤樣本集進行合併，利用合併後的樣本訓練得到預測模型；將不確定標籤樣本集中的不確定標籤樣本輸入預測模型，將模型輸出的預測值確定為不確定標籤樣本的學習標

籤；

校驗模組，用於根據當前的不確定標籤樣本集，獲得帶有學習標籤的抽樣子集，對抽樣子集的學習標籤進行校驗以獲得學習標籤的準確率，如果學習標籤的準確率未達到預設的要求，則將標籤校驗結果確定為抽樣子集的標註結果，並將抽樣子集從不確定標籤樣本集移至已確定標籤樣本集；

所述學習模組和所述校驗模組相互配合實現迭代處理，直到學習標籤的準確率達到預設的要求；以及

輸出模組，用於在迭代結束後，將不確定標籤樣本當前的學習標籤確定為標註結果，

其中，取決於所述不確定標籤樣本集中的樣本是否均為有標籤樣本、是否均為無標籤樣本或僅部分樣本帶有標籤，利用有監督學習演算法或半監督學習演算法對合併後的樣本進行訓練，得到預測模型。

【第10項】

根據請求項9所述的裝置，所述不確定標籤樣本集中的樣本均為有標籤樣本；所述學習模組具體用於：

利用有監督學習演算法對合併後的樣本進行訓練，得到預測模型。

【第11項】

根據請求項9所述的裝置，所述不確定標籤樣本集中的樣本均為無標籤樣本、或僅部分樣本帶有標籤；所述學習模組具體用於：

利用半監督學習演算法對合併後的樣本進行訓練，得到預測模型。

【第12項】

根據請求項9所述的裝置，所述輸入模組包括：

生成子模組，用於在初始獲得的待處理樣本集中樣本數量未達到預設需求的情況下，將該待處理樣本集輸入生成式對抗網路，得到與待處理樣本集同分佈的生成樣本集；以及

合併子模組，用於將待處理樣本集與生成樣本集合併，得到不確定標籤樣本集。

【第13項】

根據請求項12所述的裝置，所述學習模組具體用於：

將不確定標籤樣本集中，屬於待處理樣本集的部分輸入所述預測模型。

【第14項】

根據請求項12所述的裝置，若所述待處理樣本集中的樣本均為有標籤樣本，則所述生成子模組具體用於：

將該待處理樣本集的特徵部分和標籤部分輸入生成式對抗網路，得到與待處理樣本集同分佈的生成樣本集，所述生成樣本集中的樣本均為有標籤樣本；

或者

將該待處理樣本集的特徵部分輸入生成式對抗網路，得到與待處理樣本集同分佈的生成樣本集，所述生成樣本集中的樣本均為無標籤樣本。

【第 15 項】

根據請求項 12 所述的裝置，若所述待處理樣本集中的樣本均中的樣本均為無標籤樣本、或僅部分樣本帶有標籤，則生成子模組具體用於：

將該待處理樣本集的特徵部分輸入生成式對抗網路，得到與待處理樣本集同分佈的生成樣本集，所述生成樣本集中的樣本均為無標籤樣本。

【第 16 項】

一種資料模型訓練裝置，該裝置包括：

輸入模組，用於獲得不確定標籤樣本集和已確定標籤樣本集；

學習模組，用於對當前的不確定標籤樣本集和已確定標籤樣本集進行合併，利用合併後的樣本訓練得到預測模型；將不確定標籤樣本集中的不確定標籤樣本輸入預測模型，將模型輸出的預測值確定為不確定標籤樣本的學習標籤；

校驗模組，用於根據當前的不確定標籤樣本集，獲得帶有學習標籤的抽樣子集，對抽樣子集的學習標籤進行校驗以獲得學習標籤的準確率，如果學習標籤的準確率未達到預設的要求，則將標籤校驗結果確定為抽樣子集的標註結果，並將抽樣子集從不確定標籤樣本集移至已確定標籤樣本集；

所述學習模組和所述校驗模組相互配合實現迭代處理，直到學習標籤的準確率達到預設的要求；以及

輸出模組，用於在迭代結束後，將當前的預測模型確定為適用於所述不確定標籤樣本集的預測模型，

其中，取決於所述不確定標籤樣本集中的樣本是否均為有標籤樣本、是否均為無標籤樣本或僅部分樣本帶有標籤，利用有監督學習演算法或半監督學習演算法對合併後的樣本進行訓練，得到預測模型。

【第17項】

一種電腦設備，包括：處理器；儲存器，及儲存在所述儲存器上並可在所述處理器上運行的電腦程式，其中，所述處理器執行所述程式時實現如請求項1至8中任一項所述的方法。

【發明圖式】

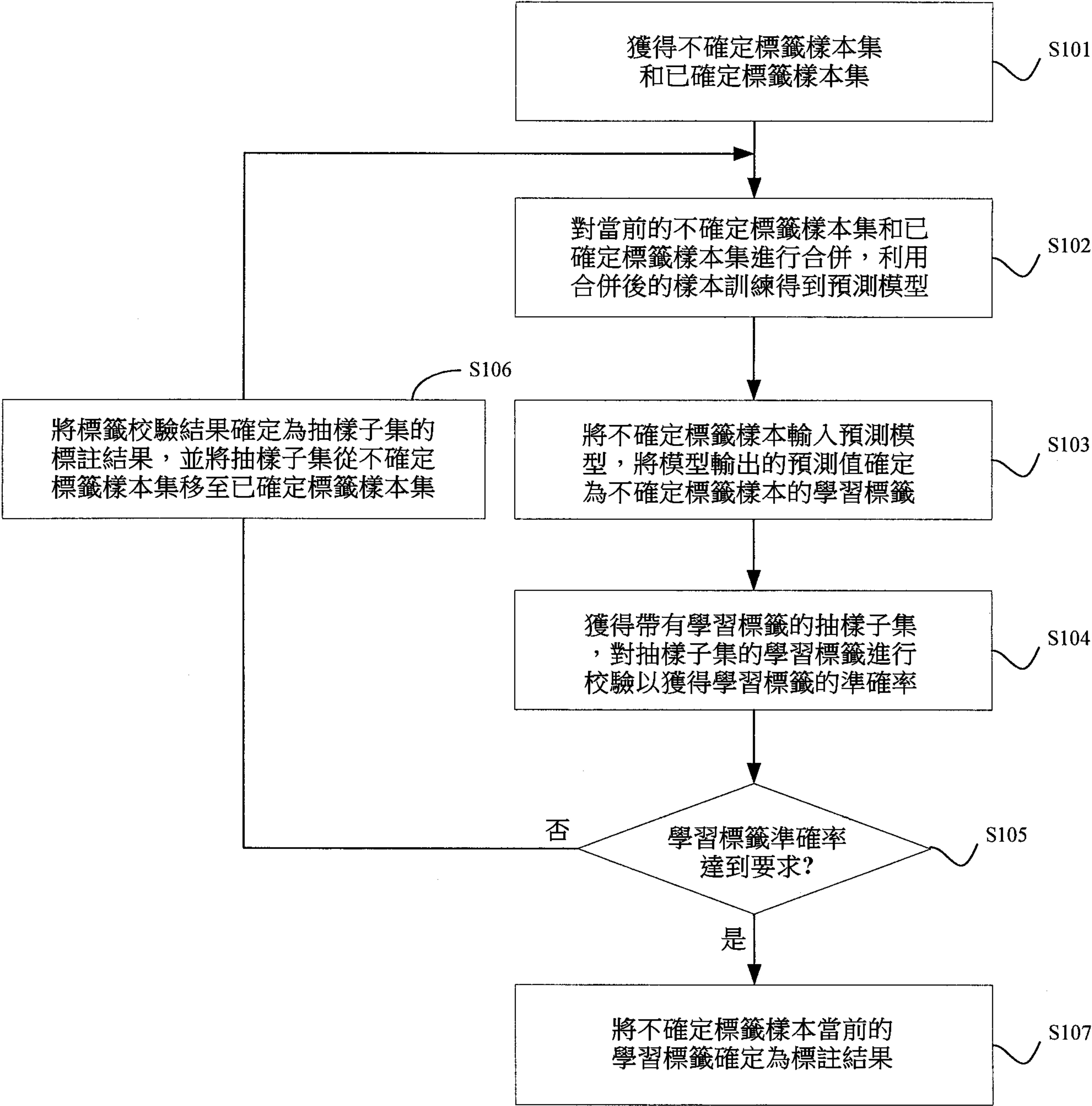


圖 1

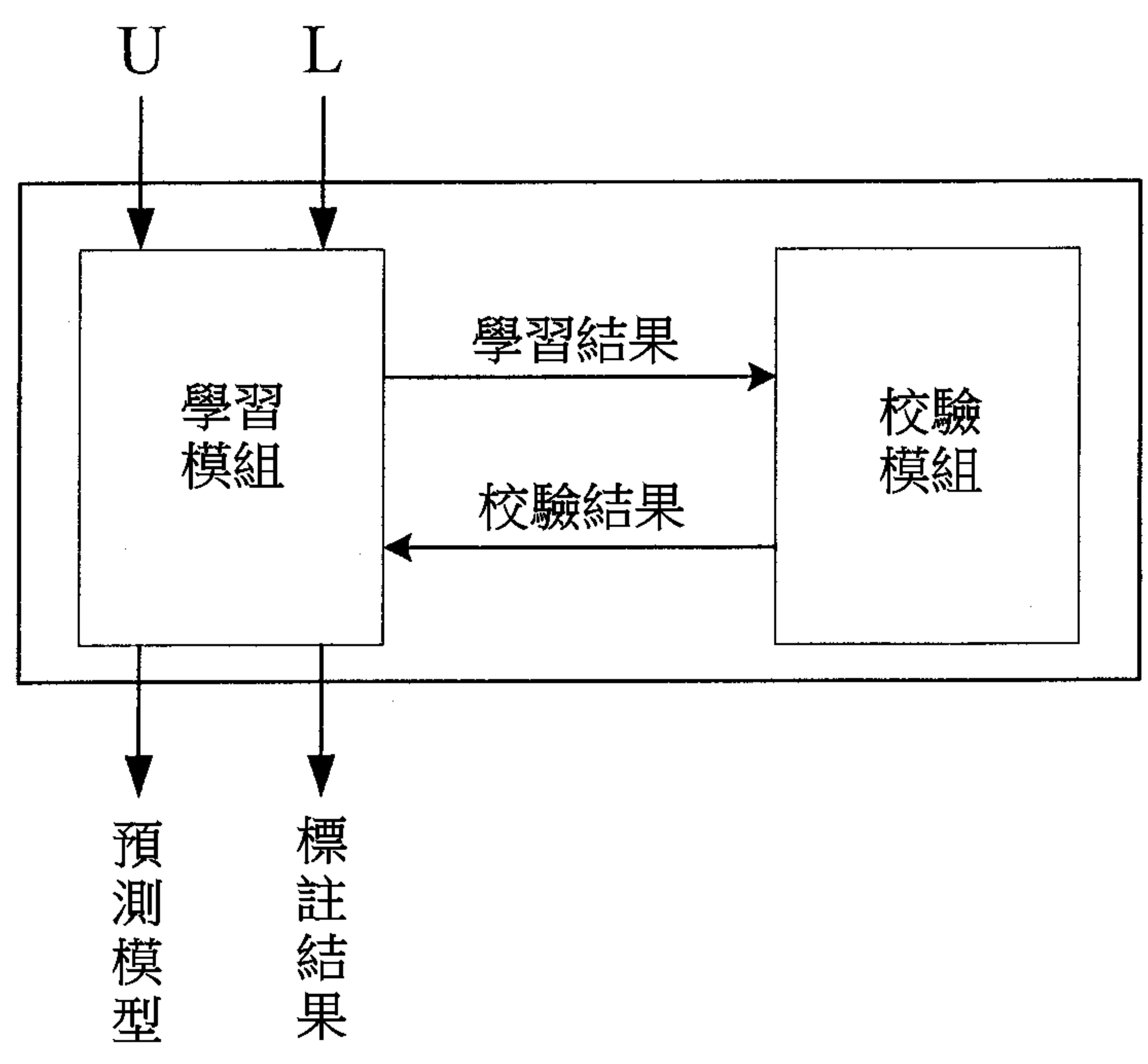


圖 2

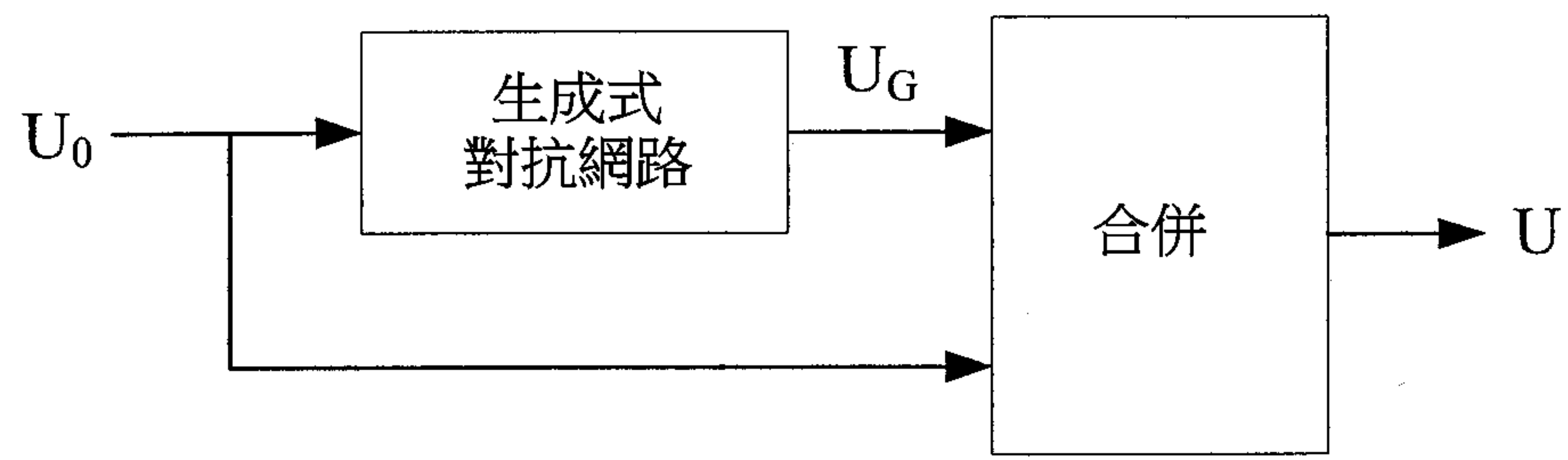


圖 3

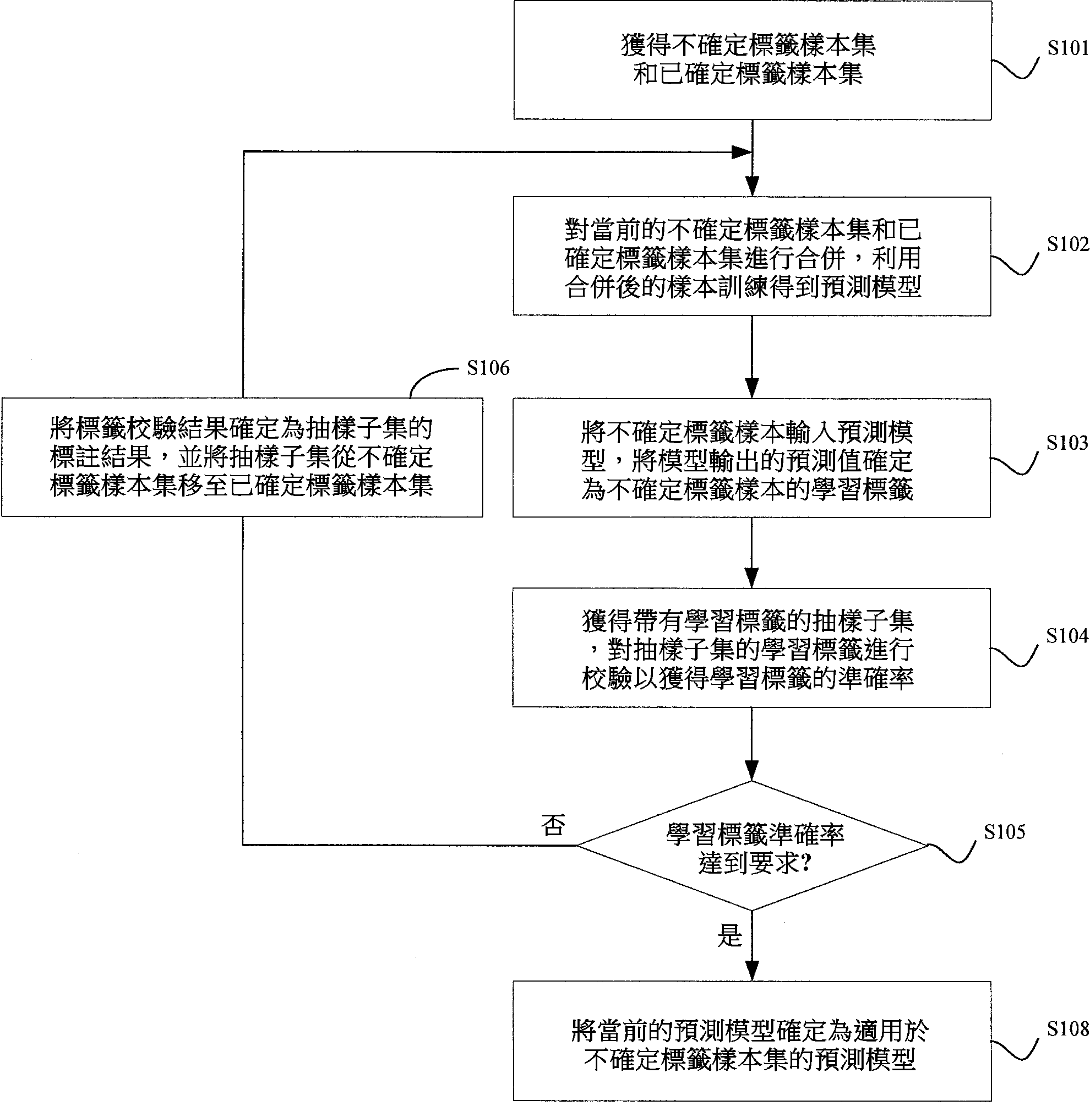


圖 4

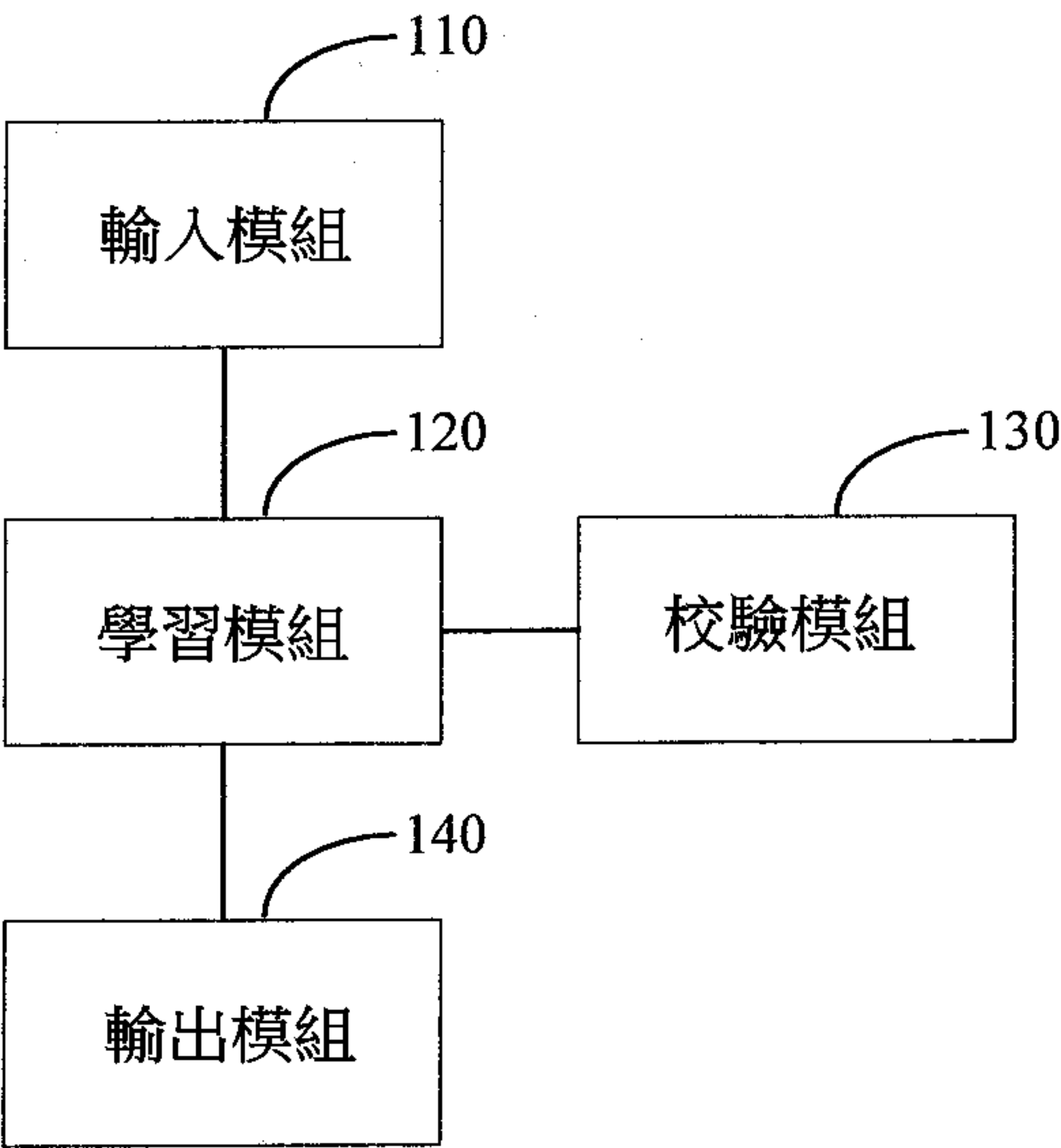


圖 5

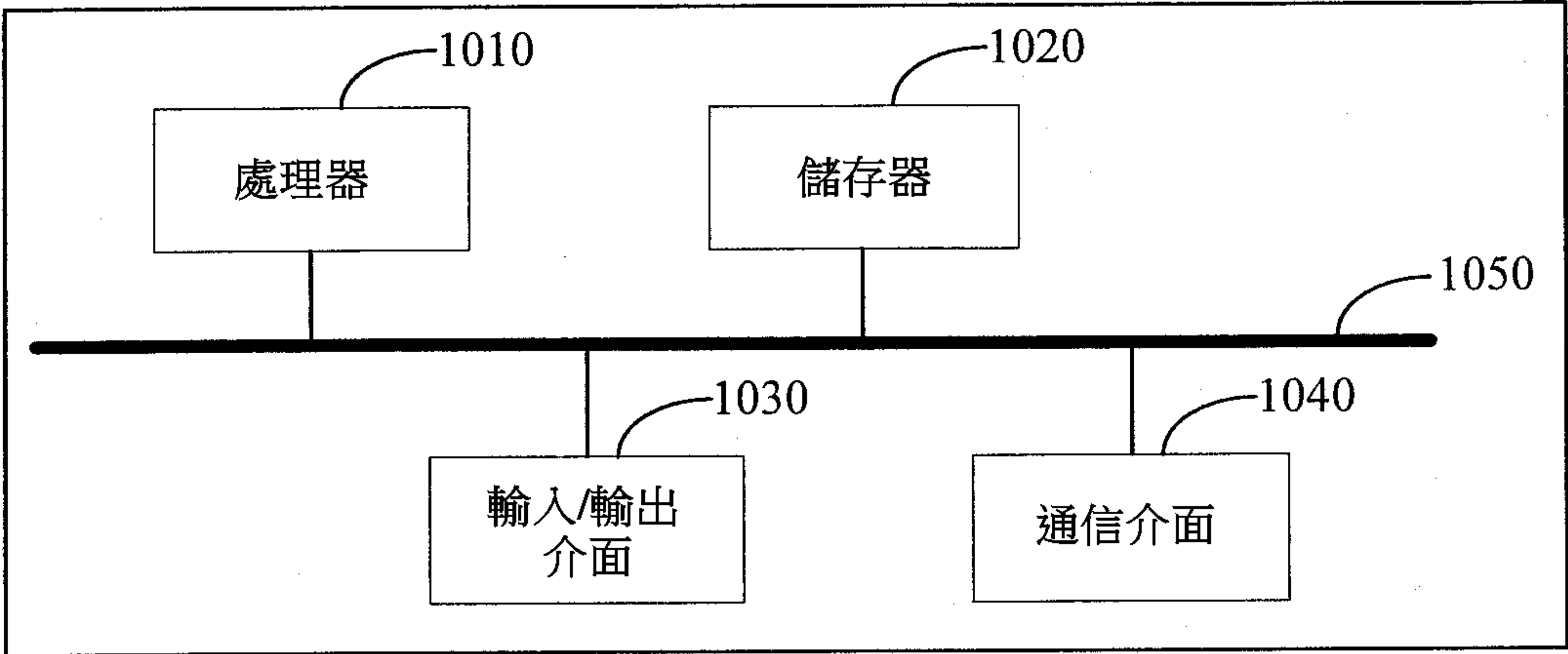


圖 6