【發明說明書】

【中文發明名稱】

模型的訓練方法、資料相似度的確定方法、裝置及設備

【技術領域】

本申請係關於電腦技術領域，尤其關於一種模型的訓練方法、資料相似度的確定方法、裝置及設備。

【先前技術】

人臉識別作為一種新型的身份核實方式，在為用戶提供便利的同時也產生了新的風險點。對於長相極為相似的多個用戶（如雙胞胎），透過人臉識別將很難有效區分不同用戶，從而極易造成因為無法正確識別導致的帳戶誤登錄，以及帳戶資金被盜用等風險。雙胞胎特別是同卵雙胞胎作為已知的長相極為相似的最典型情況，因為兩者彼此關係親密，非常容易產生上述風險行為。如何從大量資料中確定雙胞胎的用戶資料成為需要解決的重要問題。

通常，基於監督式的機器學習方法利用預先選取的樣本資料構造識別模型，具體地，調查人員透過問卷調查、有獎問答或人工觀察等方式進行社會調查，收集用戶資料，並透過人工觀察或向調查者詢問等方式得到的用戶之間的關聯關係或雙胞胎關係進行標注。透過人工標注的關聯關係或雙胞胎關係，使用相應的用戶資料作為樣本資料

構造識別模型。

然而，上述透過監督式機器學習方法構造的識別模型，其樣本資料需要進行人工標注，而人工標注的過程會消耗大量的人力資源，而且還會消耗大量的時間進行標注，從而使得模型訓練效率低下，且資源消耗較大。

【發明內容】

本申請實施例的目的是提供一種模型的訓練方法、資料相似度的確定方法、裝置及設備，以實現模型訓練的快速完成，提高模型訓練效率並減少資源消耗。

為解決上述技術問題，本申請實施例是這樣實現的：

本申請實施例提供的一種模型的訓練方法，所述方法包括：

獲取多個用戶資料對，其中，所述每個用戶資料對中的兩組用戶資料的資料欄位有相同的部分；

獲取每個用戶資料對所對應的用戶相似度，所述用戶相似度為每個用戶資料對中的兩組用戶資料對應的用戶之間的相似度；

根據所述每個用戶資料對所對應的用戶相似度和所述多個用戶資料對，確定用於訓練預設的分類模型的樣本資料；

基於所述樣本資料對所述分類模型進行訓練，得到相似度分類模型。

可選地，所述獲取每個用戶資料對所對應的用戶相似

度，包括：

獲取第一用戶資料對所對應的用戶的生物特徵，其中，所述第一用戶資料對為所述多個用戶資料對中的任意用戶資料對；

根據所述第一用戶資料對所對應的用戶的生物特徵，確定所述第一用戶資料對所對應的用戶相似度。

可選地，所述生物特徵包括面部圖像特徵，

所述獲取第一用戶資料對所對應的用戶的生物特徵，包括：

獲取第一用戶資料對所對應的用戶的面部圖像；

對所述面部圖像進行特徵提取，得到面部圖像特徵；

相應的，所述根據所述第一用戶資料對所對應的用戶的生物特徵，確定所述第一用戶資料對所對應的用戶相似度，包括：

根據所述第一用戶資料對所對應的用戶的面部圖像特徵，確定所述第一用戶資料對所對應的用戶相似度。

可選地，所述生物特徵包括語音特徵，

所述獲取第一用戶資料對所對應的用戶的生物特徵，包括：

獲取第一用戶資料對所對應的用戶的語音資料；

對所述語音資料進行特徵提取，得到語音特徵；

相應的，所述根據所述第一用戶資料對所對應的用戶的生物特徵，確定所述第一用戶資料對所對應的用戶相似度，包括：

根據所述第一用戶資料對所對應的用戶的語音特徵，確定所述第一用戶資料對所對應的用戶相似度。

可選地，所述根據所述每個用戶資料對所對應的用戶相似度和所述多個用戶資料對，確定用於訓練分類模型的樣本資料，包括：

對所述多個用戶資料對中的每個用戶資料對進行特徵提取，得到每個用戶資料對中的兩組用戶資料之間相關聯的用戶特徵；

根據所述每個用戶資料對中用戶資料之間相關聯的用戶特徵和所述每個用戶資料對所對應的用戶相似度，確定用於訓練分類模型的樣本資料。

可選地，所述根據所述每個用戶資料對中的兩組用戶資料之間相關聯的用戶特徵和所述每個用戶資料對所對應的用戶相似度，確定用於訓練分類模型的樣本資料，包括：

根據每個用戶資料對所對應的用戶相似度和預定的相似度閾值，從所述多個用戶資料對所對應的用戶特徵中選取正樣本特徵和負樣本特徵；

將所述正樣本特徵和負樣本特徵作為用於訓練分類模型的樣本資料。

可選地，所述用戶特徵包括戶籍維度特徵、姓名維度特徵、社交特徵和興趣愛好特徵；所述戶籍維度特徵包括用戶身份資訊的特徵，所述姓名維度特徵包括用戶姓名資訊的特徵和用戶姓氏的稀缺程度的特徵，所述社交特徵包

括用戶的社會關係資訊的特徵。

可選地，所述正樣本特徵和負樣本特徵中包含的特徵數目相同。

可選地，所述相似度分類模型為二分類器模型。

本申請實施例還提供的一種資料相似度的確定方法，所述方法包括：

獲取待測用戶資料對；

對所述待測用戶資料對中每組待測用戶資料進行特徵提取，得到待測用戶特徵；

根據所述待測用戶特徵和預先訓練的相似度分類模型，確定所述待測用戶資料對中的兩組待測用戶資料對應的用戶之間的相似度。

可選地，所述方法還包括：

如果所述待測用戶資料對中的兩組待測用戶資料對應的用戶之間的相似度大於預定相似度分類閾值，則確定所述待測用戶資料對所對應的待測用戶為雙胞胎。

本申請實施例提供的一種模型的訓練裝置，所述裝置包括：

資料獲取模組，用於獲取多個用戶資料對，其中，所述每個用戶資料對中的兩組用戶資料的資料欄位有相同的部分；

相似度獲取模組，用於獲取每個用戶資料對所對應的用戶相似度，所述用戶相似度為每個用戶資料對中的兩組用戶資料對應的用戶之間的相似度；

樣本資料確定模組，用於根據所述每個用戶資料對所對應的用戶相似度和所述多個用戶資料對，確定用於訓練預設的分類模型的樣本資料；

模型訓練模組，用於基於所述樣本資料對所述分類模型進行訓練，得到相似度分類模型。

可選地，所述相似度獲取模組，包括：

生物特徵獲取單元，用於獲取第一用戶資料對所對應的用戶的生物特徵，其中，所述第一用戶資料對為所述多個用戶資料對中的任意用戶資料對；

相似度獲取單元，用於根據所述第一用戶資料對所對應的用戶的生物特徵，確定所述第一用戶資料對所對應的用戶相似度。

可選地，所述生物特徵包括面部圖像特徵，

所述生物特徵獲取單元，用於獲取第一用戶資料對所對應的用戶的面部圖像；對所述面部圖像進行特徵提取，得到面部圖像特徵；

相應的，所述相似度獲取單元，用於根據所述第一用戶資料對所對應的用戶的面部圖像特徵，確定所述第一用戶資料對所對應的用戶相似度。

可選地，所述生物特徵包括語音特徵，

所述生物特徵獲取單元，用於獲取第一用戶資料對所對應的用戶的語音資料；對所述語音資料進行特徵提取，得到語音特徵；

相應的，所述相似度獲取單元，用於根據所述第一用

戶資料對所對應的用戶的語音特徵，確定所述第一用戶資料對所對應的用戶相似度。

可選地，所述樣本資料確定模組，包括：

特徵提取單元，用於對所述多個用戶資料對中的每個用戶資料對進行特徵提取，得到每個用戶資料對中的兩組用戶資料之間相關聯的用戶特徵；

樣本資料確定單元，用於根據所述每個用戶資料對中的兩組用戶資料之間相關聯的用戶特徵和所述每個用戶資料對所對應的用戶相似度，確定用於訓練分類模型的樣本資料。

可選地，所述樣本資料確定單元，用於根據每個用戶資料對所對應的用戶相似度和預定的相似度閾值，從所述多個用戶資料對所對應的用戶特徵中選取正樣本特徵和負樣本特徵；將所述正樣本特徵和負樣本特徵作為用於訓練分類模型的樣本資料。

可選地，所述用戶特徵包括戶籍維度特徵、姓名維度特徵、社交特徵和興趣愛好特徵；所述戶籍維度特徵包括用戶身份資訊的特徵，所述姓名維度特徵包括用戶姓名資訊的特徵和用戶姓氏的稀缺程度的特徵，所述社交特徵包括用戶的社會關係資訊的特徵。

可選地，所述正樣本特徵和負樣本特徵中包含的特徵數目相同。

可選地，所述相似度分類模型為二分類器模型。

本申請實施例還提供的一種資料相似度的確定裝置，

所述裝置包括：

待測資料獲取模組，用於獲取待測用戶資料對；

特徵提取模組，用於對所述待測用戶資料對中每組待測用戶資料進行特徵提取，得到待測用戶特徵；

相似度確定模組，用於根據所述待測用戶特徵和預先訓練的相似度分類模型，確定所述待測用戶資料對中的兩組待測用戶資料對應的用戶之間的相似度。

可選地，所述裝置還包括：

相似度分類別模組，用於如果所述待測用戶資料對中的兩組待測用戶資料對應的用戶之間的相似度大於預定相似度分類閾值，則確定所述待測用戶資料對所對應的待測用戶為雙胞胎。

本申請實施例提供的一種模型的訓練設備，所述設備包括：

處理器；以及

被安排成儲存電腦可執行指令的記憶體，所述可執行指令在被執行時使所述處理器執行以下操作：

獲取多個用戶資料對，其中，所述每個用戶資料對中的兩組用戶資料的資料欄位有相同的部分；

獲取每個用戶資料對所對應的用戶相似度，所述用戶相似度為每個用戶資料對中的兩組用戶資料對應的用戶之間的相似度；

根據所述每個用戶資料對所對應的用戶相似度和所述多個用戶資料對，確定用於訓練預設的分類模型的樣本資

料；

基於所述樣本資料對所述分類模型進行訓練，得到相似度分類模型。

本申請實施例提供的一種資料相似度的確定設備，所述設備包括：

處理器；以及

被安排成儲存電腦可執行指令的記憶體，所述可執行指令在被執行時使所述處理器執行以下操作：

獲取待測用戶資料對；

對所述待測用戶資料對中每組待測用戶資料進行特徵提取，得到待測用戶特徵；

根據所述待測用戶特徵和預先訓練的相似度分類模型，確定所述待測用戶資料對中的兩組待測用戶資料對應的用戶之間的相似度。

由以上本申請實施例提供的技術方案可見，本申請實施例透過獲取的多個用戶資料對，且每個用戶資料對中的兩組用戶資料的資料欄位有相同的部分，以及獲取的每個用戶資料對所對應的用戶相似度，確定用於訓練預設的分類模型的樣本資料，然後，基於樣本資料對分類模型進行訓練，得到相似度分類模型，以便後續可以透過相似度分類模型確定待測用戶資料對中的兩組待測用戶資料對應的用戶之間的相似度，這樣，僅透過相同的資料欄位得到多個用戶資料對，並透過用戶相似度確定每個用戶資料對中的兩組用戶資料對應的用戶之間的關聯關係，得到用於訓

練預設的分類模型的樣本資料，而不需要人工標注即可得到樣本資料，可以實現模型訓練的快速完成，提高了模型訓練效率並減少資源消耗。

【圖式簡單說明】

為了更清楚地說明本申請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申請中記載的一些實施例，對於本領域普通技術人員來講，在不付出創造性勞動性的前提下，還可以根據這些附圖獲得其他的附圖。

圖1為本申請一種模型的訓練方法實施例；

圖2為本申請一種資料相似度的確定方法實施例；

圖3為本申請一種檢測應用程式的介面示意圖；

圖4為本申請一種資料相似度的確定方法實施例；

圖5為本申請一種資料相似度的確定過程的處理邏輯示意圖；

圖6為本申請一種模型的訓練裝置實施例；

圖7為本申請一種資料相似度的確定裝置實施例；

圖8為本申請一種模型的訓練設備實施例；

圖9為本申請一種資料相似度的確定設備實施例。

【實施方式】

本申請實施例提供一種模型的訓練方法、資料相似度

的確定方法、裝置及設備。

為了使本技術領域的人員更好地理解本申請中的技術方案，下面將結合本申請實施例中的附圖，對本申請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申請一部分實施例，而不是全部的實施例。基於本申請中的實施例，本領域普通技術人員在沒有作出創造性勞動前提下所獲得的所有其他實施例，都應當屬於本申請保護的範圍。

實施例一

如圖1所示，本申請實施例提供一種模型的訓練方法，該方法的執行主體可以為終端設備或伺服器，其中的終端設備可以是個人電腦等，伺服器可以是獨立的一個伺服器，也可以是由多個伺服器組成的伺服器集群。本申請實施例中為了提高模型訓練的效率，該方法的執行主體以伺服器為例進行詳細說明。該方法具體可以包括以下步驟：

在步驟S102中，獲取多個用戶資料對，其中，每個用戶資料對中的兩組用戶資料的資料欄位有相同的部分。

其中，每個用戶資料對中可以包含多個不同用戶的用戶資料，例如，多個用戶資料對中包括用戶資料對A和用戶資料對B，其中，用戶資料對A中包括用戶資料1和用戶資料2，用戶資料對B中包括用戶資料3和用戶資料4等。用戶資料可以是與某用戶相關的資料，例如，用戶的姓名、

年齡、身高、住址、身份證號碼、社會保障卡（即社保卡）號碼等身份資訊，還可以包括用戶的興趣愛好、購買商品、旅遊等資訊。資料欄位可以是能夠表徵用戶資料對中的兩組不同用戶資料對應的用戶的身份，以及用戶之間的關聯關係的欄位或字元，例如，姓氏、身份證號碼中的預定位數的數值（如身份證號碼的前14位數字）、社會保障卡號碼或其它能夠確定用戶身份或資訊的證件號碼等。

在實施中，可以透過多種方式獲取用戶資料，例如，可以透過購買的方式從不同的用戶處購買其用戶資料，或者，用戶註冊某網站或應用程式時填寫的資訊，如註冊支付寶時填寫的資訊等，或者，用戶主動上傳的用戶資料等，其中，具體透過何種方式獲取用戶資料，本申請實施例對此不做限定。獲取到用戶資料後，可以將獲取的用戶資料中包含的資料欄位進行對比，從中查找出其資料欄位有相同的部分的用戶資料，並可以將資料欄位中有相同的部分的用戶資料組成一組，得到一個用戶資料對，透過上述方式，可以得到多組用戶資料對，且每個用戶資料對中都包含有資料欄位的相同部分。

例如，在實際應用中，為了盡可能的減少運算量、提高處理效率，可以設定資料欄位為身份證號碼和姓氏，則可以在用戶資料中查找用戶的身份證號碼和姓名等資訊，考慮到身份證號碼的某一位數字或多位數字可以表徵兩個用戶之間的關係，例如，身份證號碼的前14位數字等。本申請實施例中以身份證號碼的前14位數字作為判定資料欄

位中是否具有相同部分的依據為例，具體地，可以獲取每個用戶的身份證號碼的前14位數字和姓氏，並比較不同用戶的身份證號碼的前14位數字和姓氏。可以將具有相同姓氏且身份證號碼的前14位數字相同的兩組用戶資料劃分到同一個用戶資料對中。具體可以透過用戶對的形式儲存用戶資料對，例如，{用戶1身份證號碼，用戶2身份證號碼，用戶1姓名，用戶2姓名，用戶1其它資料，用戶2其它資料}等。

需要說明的是，上述兩組用戶資料的資料欄位有相同的部分可以理解為資料欄位中的一部分內容相同，如上述內容中18位身份證號碼的前14位數字等，也可以理解為資料欄位的全部內容相同等。

在步驟S104中，獲取每個用戶資料對所對應的用戶相似度，用戶相似度為每個用戶資料對中的兩組用戶資料對應的用戶之間的相似度。

其中，用戶相似度可以用於表徵多個用戶之間的相似程度，例如99%或50%等，在實際應用中，用戶相似度還可以透過其它方式表示，例如用戶相似度還可以以雙胞胎和非雙胞胎，或者同卵雙胞胎和異卵雙胞胎來表示等。

在實施中，本實施例的主要目的是訓練分類模型，這樣就需要訓練分類模型的樣本資料，以及該樣本資料對應的用戶相似度，用戶相似度可以預先儲存於伺服器或終端設備中。用戶相似度的確定可以包括多種方式，以下提供一種可選的處理方式，具體可以參見以下內容：可以預先

獲取用戶的圖像，該圖像可以是用戶註冊應用程式或網站的過程中用戶上傳的，其中的用戶可以是每個用戶資料對中包含的兩組用戶資料對應的用戶。可以將每個用戶資料對中的圖像進行對比，透過圖像的對比，可以計算該用戶資料對中的兩組用戶資料對應的用戶之間的相似度。在進行圖像對比的過程中，可以使用如圖像預處理、圖像特徵提取、圖像特徵對比等處理方式，本申請實施例對此不做限定。

在步驟S106中，根據每個用戶資料對所對應的用戶相似度和上述多個用戶資料對，確定用於訓練預設的分類模型的樣本資料。

其中，分類模型可以是任意分類模型，如樸素貝葉斯分類模型、Logistic回歸分類模型、決策樹分類模型或支援向量機分類模型等，本申請實施例中考慮到分類模型僅用於判斷兩個不同用戶之間是否相似，因此，該分類模型可以選用二分類模型。樣本資料可以是用於訓練分類模型的資料，該樣本資料可以是用戶資料對中的兩組用戶資料，也可以是上述用戶資料經過某種處理後得到的資料等，如對上述用戶資料進行特徵提取，得到相應的用戶特徵，該用戶特徵的資料可以作為樣本資料。

在實施中，可以預先設置相似度閾值，如80%或70%等，然後，可以將每個用戶資料對所對應的用戶相似度分別與相似度閾值相比較，可以將用戶相似度大於相似度閾值的用戶資料對劃分為一組，可以將用戶相似度小於相似

度閾值的用戶資料對劃分為一組，可以從上述兩組中各選取預定數目（如4萬或5萬等）的用戶資料對，並將選取的用戶資料對作為用於訓練預設的分類模型的樣本資料。

需要說明的是，選取用於訓練預設的分類模型的樣本資料的方式除了上述方式外，還可以包括多種，例如，提取每個用戶資料對中包含的兩組用戶資料的特徵，得到相應的用戶特徵，然後，可以透過每個用戶資料對所對應的用戶相似度和相似度閾值，將用戶特徵劃分為如上述的兩組，可以將兩組用戶特徵的資料作為用於訓練預設的分類模型的樣本資料。

在步驟S108中，基於上述樣本資料對分類模型進行訓練，得到相似度分類模型。

其中，相似度分類模型可以是用於確定不同用戶之間的相似程度的模型。

在實施中，基於上述選取的用戶資料對作為用於訓練預設的分類模型的樣本資料的情況，可以對選取的用戶資料對中的兩組用戶資料進行特徵提取，得到相應的用戶特徵，然後，可以將樣本資料中每個用戶資料對的用戶特徵輸入到分類模型中進行計算，計算完成後，可以輸出計算結果。可以將該計算結果與相應的用戶資料對所對應的用戶相似度進行比較，確定兩者是否相同，如果兩者不同，則可以修改分類模型的相關參數，然後，再將該用戶資料對的用戶特徵輸入到修改後的分類模型中進行計算，並判斷計算結果與用戶相似度是否相同，直到兩者相同為止。

如果兩者相同，則可以選取下一個用戶資料對執行上述處理過程，最終每個用戶資料對的用戶特徵輸入到分類模型後得到的計算結果與相應的用戶資料對所對應的用戶相似度均相同，則得到的分類模型即為相似度分類模型。

透過上述方式可以得到相似度分類模型，該相似度分類模型的使用可以參見下述相關內容：

如圖2所示，本申請實施例提供一種相似度的確定方法，該方法的執行主體可以為終端設備或伺服器，其中的終端設備可以是個人電腦等，伺服器可以是獨立的一個伺服器，也可以是由多個伺服器組成的伺服器集群。該方法具體可以包括以下步驟：

在步驟S202中，獲取待測用戶資料對。

其中，待測用戶資料對可以是待檢測的兩個用戶的用戶資料所組成的用戶資料對。

在實施中，為了檢測出兩個不同用戶之間的相似度，可以設置相應的檢測應用程式。如圖3所示，該檢測應用程式中可以包括用於上傳資料的按鍵，當需要對兩個不同用戶進行相似度檢測時，可以點擊上述用於上傳資料的按鍵，該檢測應用程式可以彈出資料上傳的提示框，資料上傳者可以在提示框中輸入待測用戶資料對的資料，輸入完成後，可以點擊該提示框中的確定按鍵，該檢測應用程式可以獲取資料上傳者輸入的待測用戶資料對。上述檢測應用程式可以安裝在終端設備上，也可以安裝在伺服器上，本申請實施例提供的相似度的確定方法的執行主體若為伺

服器，且如果檢測應用程式安裝在終端設備上，則檢測應用程式獲取到待測用戶資料對後，可以將該待測用戶資料對發送給伺服器，從而伺服器可以獲取到待測用戶資料對。如果檢測應用程式安裝在伺服器上，則伺服器透過檢測應用程式可以直接獲取到待測用戶資料對。

在步驟S204中，對上述待測用戶資料對中每組待測用戶資料進行特徵提取，得到待測用戶特徵。

其中，待測用戶特徵可以是待檢測的用戶的用戶資料的特徵。

在實施中，可以獲取上述待測用戶資料對中每組待測用戶資料，針對其中的任意一組待測用戶資料，可以使用預先設置的特徵提取演算法，從該待測用戶資料中提取相應的特徵，可以將提取的特徵作為該待測用戶資料對應的待測用戶特徵。透過上述方式可以得到待測用戶資料對中每組待測用戶資料對應的待測用戶特徵。

需要說明的是，特徵提取演算法可以是能夠從用戶資料中提取預定特徵的任意演算法，具體可以根據實際情況進行設定。

在步驟S206中，根據上述待測用戶特徵和預先訓練的相似度分類模型，確定上述待測用戶資料對中的兩組待測用戶資料對應的用戶之間的相似度。

在實施中，可以將透過上述步驟S204得到的待測用戶特徵輸入到透過上述步驟S102~步驟S108得到的相似度分類模型中進行計算，相似度分類模型輸出的結果即可以為

上述待測用戶資料對中的兩組待測用戶資料對應的用戶之間的相似度。

需要說明的是，在實際應用中，相似度分類模型的直接輸出結果可以以百分比的方式展現，例如90%或40%等，為了使得輸出結果對用戶來說更加直觀，可以根據實際情況對相似度分類模型的直接輸出結果進一步設定，例如，需要區分同卵雙胞胎和非同卵雙胞胎，或者，需要區分同卵雙胞胎和異卵雙胞胎等，對於上述情況，可以設置分類閾值，如果直接輸出結果大於該分類閾值，則確定上述待測用戶資料對中的兩組待測用戶資料對應的用戶之間為同卵雙胞胎，否則為非同卵雙胞胎或異卵雙胞胎等。這樣，透過預先訓練的相似度分類模型，可以快速判斷出待測用戶資料對中的兩組待測用戶資料對應的用戶之間的相似度，提高了用戶之間相似度的判定效率。

需要說明的是，上述用戶資料對和待測用戶資料對均是以包含兩組用戶資料來說明，在實際應用中，本申請提供的模型的訓練方法和相似度的確定方法還可以應用於包含兩組以上的用戶資料的用戶資料組合和待測用戶資料組合，具體處理可以參見本申請實施例中的相關內容，在此不再贅述。

本申請實施例提供一種模型的訓練方法和相似度的確定方法，透過獲取的多個用戶資料對，且每個用戶資料對中的兩組用戶資料的資料欄位有相同的部分，以及獲取的每個用戶資料對所對應的用戶相似度，確定用於訓練預設

的分類模型的樣本資料，然後，基於樣本資料對分類模型進行訓練，得到相似度分類模型，以便後續可以透過相似度分類模型確定待測用戶資料對中的兩組待測用戶資料對應的用戶之間的相似度，這樣，僅透過相同的資料欄位得到多個用戶資料對，並透過用戶相似度確定每個用戶資料對中的兩組用戶資料對應的用戶之間的關聯關係，得到用於訓練預設的分類模型的樣本資料，而不需要人工標注即可得到樣本資料，可以實現模型訓練的快速完成，提高了模型訓練效率並減少資源消耗。

實施例二

如圖4所示，本申請實施例提供了一種資料相似度的確定方法，該方法的執行主體可以為伺服器或者該方法可以由終端設備和伺服器共同實現，其中的終端設備可以是個人電腦等，伺服器可以是獨立的一個伺服器，也可以是由多個伺服器組成的伺服器集群。本申請實施例中為了提高模型訓練的效率，該方法的執行主體以伺服器為例進行詳細說明，對於由終端設備和伺服器共同實現的按情況，可以參見下述相關內容，在此不再贅述。該方法具體包括如下內容：

目前人臉識別作為一種用戶核實身份的新型方式，在為用戶提供便利的同時也產生了新的風險點，目前的人臉識別技術都是利用現場採集的用戶圖像與該用戶在人臉識別系統的資料庫中留存的用戶圖像進行比較，只要比對數

值達到預定閾值，則認為該用戶為留存的用戶圖像所對應的用戶，以達到核實用戶身份的目的。然而，針對長相極為相似的臉，上述方式將很難對用戶的身份進行有效核實，從而極易造成因為無法進行身份核實導致的帳戶誤登錄以及後續的資金盜用等。

雙胞胎特別是同卵雙胞胎作為已知的相似臉的最典型情況，因為彼此關係親密，這樣就更容易產生有關負面輿情。如果可以掌握盡可能多的雙胞胎用戶名單，就可以針對這部分用戶群體有單獨的人臉識別應對策略以預防上述風險。為此可以構造有效識別雙胞胎的模型，在保證高準確率的前提下輸出雙胞胎名單用於監控這些用戶的人臉識別行為以起到風險控制的作用。其中，構造有效識別雙胞胎的模型的處理可以參見下述步驟S402~步驟S412提供的模型的訓練方法，具體內容如下：

在步驟S402中，獲取多個用戶資料對，其中，每個用戶資料對中的兩組用戶資料的資料欄位有相同的部分。

在實施中，考慮到雙胞胎通常是姓氏相同且身份證號碼的前14位數字相同，因此，可以將姓氏和身份證號碼的前14位數字作為資料欄位來選取用戶資料對，上述步驟S402的具體處理過程可以參見上述實施例一中步驟S102的相關內容，在此不再贅述。

需要說明的是，上述選取用戶資料對的處理是透過姓氏和身份證號碼的前14位數字來實現的，在本申請的另一實施例中，選取用戶資料對的處理還可以透過其它資訊來

實現，例如，透過姓氏和社會保障卡號碼來實現，或者，透過身份證號碼的前14位數字和社會保障卡號碼來實現等，本申請實施例對此不做限定。

考慮到在對模型進行訓練時，需要確定用戶資料對中的兩組用戶資料對應的用戶之間的相似程度，以下提供一種相關的處理方式，具體可以參見以下步驟S404和步驟S406。

在步驟S404中，獲取第一用戶資料對所對應的用戶的生物特徵，其中，第一用戶資料對為上述多個用戶資料對中的任意用戶資料對。

其中，生物特徵可以是人體的生理特徵和行為特徵等，如指紋特徵、虹膜特徵、面部特徵、DNA等生理特徵，再如聲紋特徵、筆跡特徵和擊鍵習慣特徵等行為特徵。

在實施中，透過上述步驟S402的處理獲取到多個用戶資料對後，可以從多個用戶資料對中任意選擇一個用戶資料對（即第一用戶資料對）。用戶透過其終端設備登錄伺服器進行註冊時，可以向伺服器上傳包含該用戶上述某一項或多項生物特徵，伺服器可以將該生物特徵與該用戶的標識對應儲存，其中，用戶的標識可以是用戶註冊時填寫的用戶名或用戶的姓名等，伺服器中對應儲存的上述資訊可以如表1所示。

表1

用戶的標識	生物特徵
用戶1	生物特徵A
用戶2	生物特徵B
用戶3	生物特徵C

當伺服器選取第一用戶資料對後，可以從第一用戶資料對中分別提取其中包含的用戶的標識，然後，透過用戶的標識可以獲取相應的生物特徵，從而得到第一用戶資料對所對應的用戶的生物特徵。例如，第一用戶資料對中包含的用戶的標識為用戶2和用戶3，則透過查找如上述表格的對應關係，可以確定用戶2對應的生物特徵為生物特徵B，用戶3對應的生物特徵為生物特徵C，即第一用戶資料對所對應的用戶的生物特徵為生物特徵B和生物特徵C。

在步驟S406中，根據第一用戶資料對所對應的用戶的生物特徵，確定第一用戶資料對所對應的用戶相似度。

在實施中，透過上述步驟S404得到第一用戶資料對所對應的用戶的生物特徵後，可以分別對得到的生物特徵進行相似度計算，從而確定相應的兩個用戶之間的相似程度（即用戶相似度），其中，相似度計算可以包括多種實現方式，例如透過特徵向量之間的歐氏距離來實現等，本申請實施例對此不做限定。

需要說明的是，可以透過設置閾值來進行相似與否的判斷，例如設置閾值為70，當兩個生物特徵對應的用戶相似度大於70時，確定第一用戶資料對中的兩組用戶資料對應的用戶相似；當兩個生物特徵對應的用戶相似度小於70

時，確定第一用戶資料對中的兩組用戶資料對應的用戶不相似。

透過上述方式可以對多個用戶資料對中除第一用戶資料對外的其它用戶資料對執行上述處理過程，從而得到多個用戶資料對中每個用戶資料對所對應的用戶相似度。

上述步驟S404和步驟S406是透過用戶的生物特徵確定用戶相似度的，在實際應用中，確定用戶相似度具體可以透過多種實現方式實現，以下以生物特徵為面部特徵為例對上述步驟S404和步驟S406進行具體說明，具體可以參見以下步驟一和步驟二。

步驟一，獲取第一用戶資料對所對應的用戶的面部圖像，其中，第一用戶資料對為上述多個用戶資料對中的任意用戶資料對。

在實施中，透過上述步驟S402的處理獲取到多個用戶資料對後，可以從多個用戶資料對中任意選擇一個用戶資料對（即第一用戶資料對）。用戶透過其終端設備登錄伺服器進行註冊時，可以向伺服器上傳包含該用戶面部的圖像，伺服器可以將該圖像與該用戶的標識對應儲存，其中，用戶的標識可以是用戶註冊時填寫的用戶名或用戶的姓名等，伺服器中對應儲存的上述資訊可以如表2所示。

表2

用戶的標識	包含用戶面部的圖像
用戶1	圖像A
用戶2	圖像B
用戶3	圖像C

當伺服器選取第一用戶資料對後，可以從第一用戶資料對中分別提取其中包含的用戶的標識，然後，透過用戶的標識可以獲取相應的圖像，從而得到第一用戶資料對所對應的用戶的面部圖像。例如，第一用戶資料對中包含的用戶的標識為用戶2和用戶3，則透過查找如上述表格的對應關係，可以確定用戶2對應的包含用戶面部的圖像為圖像B，用戶3對應的包含用戶面部的圖像為圖像C，即第一用戶資料對所對應的用戶的面部圖像為圖像B和圖像C。

步驟二，對上述面部圖像進行特徵提取，得到面部圖像特徵，並根據第一用戶資料對所對應的用戶的面部特徵，確定第一用戶資料對所對應的用戶相似度。

在實施中，透過上述步驟一得到第一用戶資料對所對應的用戶的面部圖像後，可以分別對得到的面部圖像進行特徵提取，得到相應的面部圖像特徵，並基於每個面部圖像的提取特徵得到相應的特徵向量，然後，可以計算其中任意兩個面部圖像的特徵向量之間的歐式距離，透過特徵向量之間的歐式距離的數值大小，可以確定相應的兩個用戶之間的相似程度（即用戶相似度），其中，特徵向量之間的歐式距離的數值越大，用戶相似度越低；特徵向量之間的歐式距離的數值越小，用戶相似度越高。

需要說明的是，對於面部圖像而言，兩個面部圖像只有相似和非相似的區別，為此，可以透過設置閾值來進行相似與否的判斷，例如設置閾值為70，當兩個面部圖像對應的用戶相似度大於70時，確定第一用戶資料對中的兩組

用戶資料對應的用戶相似；當兩個面部圖像對應的用戶相似度小於70時，確定第一用戶資料對中的兩組用戶資料對應的用戶不相似。

例如，基於上述步驟一的示例，分別對圖像B和圖像C進行特徵提取，透過提取的特徵分別構建相應的特徵向量，得到圖像B的特徵向量和圖像C的特徵向量。計算圖像B的特徵向量和圖像C的特徵向量之間的歐式距離，透過得到的歐式距離的數值確定用戶2和用戶3之間的用戶相似度。

透過上述方式可以對多個用戶資料對中除第一用戶資料對外的其它用戶資料對執行上述處理過程，從而得到多個用戶資料對中每個用戶資料對所對應的用戶相似度。

此外，對於上述步驟S404和步驟S406的處理，以下再提供一種可選的處理方式，具體可以參見以下步驟一和步驟二。

步驟一，獲取第一用戶資料對所對應的用戶的語音資料，其中，第一用戶資料對為多個用戶資料對中的任意用戶資料對。

在實施中，透過上述步驟S402的處理獲取到多個用戶資料對後，可以從多個用戶資料對中任意選擇一個用戶資料對（即第一用戶資料對）。用戶透過其終端設備登錄伺服器進行註冊時，可以向伺服器上傳包含預定時長（如3秒或5秒等）和／或預定語音內容（如一個或多個詞的語音或一句話的語音等）的語音資料，伺服器可以將該語音

資料與該用戶的標識對應儲存。當伺服器選取第一用戶資料對後，可以從第一用戶資料對中分別提取其中包含的用戶的標識，然後，透過用戶的標識可以獲取相應的語音資料，從而得到第二用戶資料對所對應的用戶的語音資料。

步驟二，對上述語音資料進行特徵提取，得到語音特徵，並根據第一用戶資料對所對應的用戶的語音特徵，確定第一用戶資料對所對應的用戶相似度。

在實施中，透過上述步驟一得到第一用戶資料對所對應的用戶的語音資料後，可以分別對得到的語音資料進行特徵提取，並基於每個語音資料的提取特徵確定相應的兩個用戶之間的相似程度（即用戶相似度），具體處理過程可以參見上述步驟S406中的相關內容，或者，可以透過特徵的逐一比對的方式確定用戶相似度，又或者，可以對任意兩個語音資料進行語音訊譜分析，以確定用戶相似度等。透過上述方式可以對多個用戶資料對中除第一用戶資料對外的其它用戶資料對執行上述處理過程，從而得到多個用戶資料對中每個用戶資料對所對應的用戶相似度。

在步驟S408中，對上述多個用戶資料對中的每個用戶資料對進行特徵提取，得到每個用戶資料對中的兩組用戶資料之間相關聯的用戶特徵。

在實施中，可以從多個用戶資料對中任意選取一個用戶資料對（可以稱為第三用戶資料對），可以對第三用戶資料對中的兩組不同的用戶資料分別進行特徵提取，例如，第三用戶資料對中包括用戶資料1和用戶資料2，可以

對用戶資料1進行特徵提取，並對用戶資料2進行特徵提取。然後，可以對比在不同的用戶資料中提取的特徵，從而得到第三用戶資料對中的兩組用戶資料之間相關聯的用戶特徵。透過上述方式可以對多個用戶資料對中除第三用戶資料對外的其它用戶資料對執行上述處理過程，從而得到每個用戶資料對中的兩組用戶資料之間相關聯的用戶特徵。

在實際應用中，用戶特徵可以包含但不限於戶籍維度特徵、姓名維度特徵、社交特徵和興趣愛好特徵等特徵。其中，戶籍維度特徵可以包括用戶身份資訊的特徵。戶籍維度特徵主要是基於中國的戶籍管理制度，戶籍中包含的身份證資訊中包括出生日期和戶籍申報地，同時戶籍中具有父母姓名和公民住址，然而由於歷史和其它原因，部分公民登記的資訊並不與實際情況一樣，存在如提前申報生日、雙方分別隨父母姓，甚至父母離異導致戶籍分離等情況，所以戶籍維度特徵對於判定兩個用戶是否為雙胞胎起到一定的參考作用。這樣，透過用戶資料對所對應的不同用戶之間的出生日期是否一致、戶籍申報地是否一致、是否有共同父母、現住址的一致程度等特徵確定不同用戶之間的關聯。

姓名維度特徵包括用戶姓名資訊的特徵和用戶姓氏的稀缺程度的特徵。對於姓名維度特徵，基於NLP（Nature Language Processing，自然語言處理）理論和社會經驗，通常，如果兩個人的名字看起來比較像，比如張金龍和張

金虎，或者具有某種語義關聯，如張美美和張麗麗，則認為兩者之間應該具有某種關聯。在本申請實施例中，可以引入詞典來評估兩個用戶在名字上的關係，同時利用用戶註冊的個人資訊和人口統計資料統計姓氏的稀缺程度作為特徵。這樣，透過用戶資料對所對應的不同用戶之間的姓氏是否一致、姓名長度是否一致、名字近義詞程度、名字組合是否為詞和姓氏稀缺程度等特徵確定不同用戶之間的關聯。

社交特徵包括用戶的社會關係資訊的特徵。對於社交特徵，可以是基於大數據對用戶資料對的社會關係進行提煉而成，通常，雙胞胎應該具有較多的互動和重複性較高的社會關係，如共同的親戚，甚至同學等。在本申請實施例中，基於伺服器中儲存的用戶的個人資訊構成的關係網絡、通訊錄等已有資料對用戶資料對進行關聯，以得到相應的特徵。這樣，透過用戶資料對所對應的不同用戶之間的社交應用是否互相關注、是否有資金往來、通訊錄中是否包含對方的聯繫方式、通訊錄標注是否有稱謂和通訊錄的交集數量等特徵確定不同用戶之間的關聯。

此外，考慮到雙胞胎具有較多的共同愛好、購物興趣，以及可能會共同出遊等，用戶特徵還可以包括如電商、旅遊、文娛等多維度特徵，在本申請實施例中，電商、旅遊、文娛等多維度特徵的相關資料可以從預定的資料庫或某網站中獲取得到。這樣，透過用戶資料對所對應的不同用戶之間的購物記錄的交集數量、是否有過同時出

遊、是否同時入住過酒店、購物傾向的相似度和收貨位址是否一樣等特徵確定不同用戶之間的關聯。

需要說明的是，上述確定用戶相似度的處理（即包括步驟S404和步驟S406）和特徵提取的處理（即步驟S408）是按照先後循序執行的，在實際應用中，確定用戶相似度的處理和特徵提取的處理可以同時執行，也可以先執行特徵提取的處理，然後再執行確定用戶相似度的處理，本申請實施例對此不做限定。

在步驟S410中，根據每個用戶資料對中的兩組用戶資料之間相關聯的用戶特徵和每個用戶資料對所對應的用戶相似度，確定用於訓練分類模型的樣本資料。

在實施中，可以預先設置閾值，透過閾值可以從多個用戶資料對中選取用戶相似度大於該閾值的用戶資料對，可以將選取的用戶資料對中的兩組用戶資料之間相關聯的用戶特徵作為訓練分類模型的用戶特徵，可以將選取的用戶特徵和選取的用戶資料對所對應的用戶相似度確定為用於訓練分類模型的樣本資料。

上述步驟S410的處理除了可以採用上述方式外，還可以採用多種方式處理，以下還提供一種可選的處理方式，具體可以包括以下步驟一和步驟二：

步驟一，根據每個用戶資料對所對應的用戶相似度和預定的相似度閾值，從多個用戶資料對所對應的用戶特徵中選取正樣本特徵和負樣本特徵。

在實施中，基於同卵雙胞胎長相高度相似這一常識，

以及雙胞胎出生日期、出生地等相同，且通常情況下雙胞胎的姓氏也相同的社會常識，透過兩個用戶的面部圖像計算用戶相似度，從而確定兩個用戶是否為同卵雙胞胎，具體地，可以預先設置相似度閾值，如80%或70%等，可以將用戶相似度大於相似度閾值的用戶資料對確定為同卵雙胞胎的用戶資料對，可以將用戶相似度小於相似度閾值的用戶資料對確定為非同卵雙胞胎的用戶資料對。同時，由於同卵雙胞胎和異卵雙胞胎除了在長相上有所差異外，其它特徵基本一致，所以，可以將同卵雙胞胎的用戶資料對所對應的用戶特徵作為相似度分類模型的正樣本特徵，而非同卵雙胞胎（包括異卵雙胞胎和非雙胞胎）的用戶資料對所對應的用戶特徵則作為相似度分類模型的負樣本特徵。

需要說明的是，負樣本特徵並不是指其中包含的特徵全部都是異卵雙胞胎的用戶特徵，在實際應用中，異卵雙胞胎的用戶特徵也可能在負樣本特徵中的比例極少，還可能在負樣本特徵中包含有少量的正樣本特徵，而這樣並不會影響分類模型的訓練，反而會有助於提升相似度分類模型的穩固性。

此外，正樣本特徵和負樣本特徵中包含的特徵數目可以相同。例如，從多個用戶資料對中選取用戶相似度小於10%的10000個用戶資料對，從多個用戶資料對中選取用戶相似度大於10%且小於20%的10000個用戶資料對，從多個用戶資料對中選取用戶相似度大於20%且小於30%的

10000個用戶資料對，從多個用戶資料對中選取用戶相似度大於30%且小於40%的10000個用戶資料對，從多個用戶資料對中選取用戶相似度大於40%且小於50%的10000個用戶資料對，將上述50000個用戶資料對的用戶特徵作為負樣本特徵。從多個用戶資料對中選取用戶相似度大於80%且小於90%的40000個用戶資料對，從多個用戶資料對中選取用戶相似度大於90%且小於100%的10000個用戶資料對，將上述50000個用戶資料對的用戶特徵作為正樣本特徵。

步驟二，將正樣本特徵和負樣本特徵作為用於訓練分類模型的樣本資料。

在實施中，可以將用戶特徵和相應的用戶相似度的資料組合，可以將組合後的資料作為用於訓練分類模型的樣本資料。

在步驟S412中，基於樣本資料對分類模型進行訓練，得到相似度分類模型。

其中，由於分類模型的主要目的是識別出雙胞胎，因此，為了使得本申請實施例簡化可行，相似度分類模型可以為二分類器模型，具體如GBDT（Gradient Boosting Decision Tree，反覆運算決策樹）二分類器模型。

在實施中，可以分別將正樣本特徵輸入到分類模型中進行計算，得到的計算結果可以與該正樣本特徵相應的用戶相似度對比，如果兩者相匹配，則可以選擇下一個正樣本特徵或負樣本特徵輸入到分類模型中進行計算。得到的

計算結果繼續與該正樣本特徵相應的用戶相似度匹配對比。如果兩者不匹配，則可以調整分類模型中的相關參數的數值，然後再將該正樣本特徵輸入到分類模型中進行計算，得到的計算結果再與該正樣本特徵相應的用戶相似度匹配對比，即重複上述過程，直到兩者相匹配為止。透過上述方式，可以將所有的正樣本特徵和負樣本特徵輸入到分類模型中進行計算，從而達到對分類模型進行訓練的目的，可以將最終訓練得到的分類模型作為相似度分類模型。

透過上述處理過程得到了相似度分類模型，該相似度分類模型可以用於人臉識別場景中，對於具有風險的雙胞胎用戶，透過該相似度分類模型可以進行單獨的風險控制。

得到相似度分類模型後，可以應用相似度分類模型來判定待測用戶資料對所對應的待測用戶是否為雙胞胎，如圖5所示，其中的具體處理可以參見以下步驟S414~步驟S420的內容。

在步驟S414中，獲取待測用戶資料對。

上述步驟S414的步驟內容與上述實施例一中步驟S202的步驟內容相同，步驟S414的具體處理可以參見步驟S202的相關內容，在此不再贅述。

在步驟S416中，對待測用戶資料對中每組待測用戶資料進行特徵提取，得到待測用戶特徵。

其中上述步驟S416中對待測用戶資料對中每組待測用

戶資料進行特徵提取，得到待測用戶特徵的處理過程，可以參見上述步驟S408的相關內容，即從待測用戶資料中提取的特徵包括但不限於戶籍維度特徵、姓名維度特徵、社交特徵和興趣愛好特徵等，參見上述步驟S408的相關內容，在此不再贅述。

在步驟S418中，根據待測用戶特徵和預先訓練的相似度分類模型，確定待測用戶資料對中的兩組待測用戶資料對應的用戶之間的相似度。

上述步驟S418的步驟內容與上述實施例一中步驟S206的步驟內容相同，步驟S418的具體處理可以參見步驟S206的相關內容，在此不再贅述。

在步驟S420中，如果待測用戶資料對中的兩組待測用戶資料對應的用戶之間的相似度大於預定相似度分類閾值，則確定待測用戶資料對所對應的待測用戶為雙胞胎。

在實施中，由於輸出的雙胞胎名單會影響目標用戶的人臉識別功能的使用，因此，使用的過程中需要追求相似度分類模型的高準確度，在實際應用中可以設置一個較大的數值作為相似度分類閾值，例如，95%作為相似度分類閾值或97%作為相似度分類閾值等。利用訓練好的相似度分類模型對待測用戶特徵進行預測並輸出評分。其中，評分過程是計算相應的用戶資料對所對應的用戶為雙胞胎的機率，比如機率為80%，則評分為80分，機率為90%，則評分為90分，得到的分數越高，用戶資料對所對應的用戶為雙胞胎的機率越高。

本申請實施例提供一種資料相似度的確定方法，透過獲取的多個用戶資料對，且每個用戶資料對中的兩組用戶資料的資料欄位有相同的部分，以及獲取的每個用戶資料對所對應的用戶相似度，確定用於訓練預設的分類模型的樣本資料，然後，基於樣本資料對分類模型進行訓練，得到相似度分類模型，以便後續可以透過相似度分類模型確定待測用戶資料對中的兩組待測用戶資料對應的用戶之間的相似度，這樣，僅透過相同的資料欄位得到多個用戶資料對，並透過用戶相似度確定每個用戶資料對中的兩組用戶資料對應的用戶之間的關聯關係，得到用於訓練預設的分類模型的樣本資料，而不需要人工標注即可得到樣本資料，可以實現模型訓練的快速完成，提高了模型訓練效率並減少資源消耗。

實施例三

以上為本申請實施例提供的資料相似度的確定方法，基於同樣的思路，本申請實施例還提供一種模型的訓練裝置，如圖6所示。

所述模型的訓練裝置可以設置在伺服器中，該裝置包括：資料獲取模組601、相似度獲取模組602、樣本資料確定模組603和模型訓練模組604，其中：

資料獲取模組601，用於獲取多個用戶資料對，其中，所述每個用戶資料對中的兩組用戶資料的資料欄位有相同的部分；

相似度獲取模組602，用於獲取每個用戶資料對所對應的用戶相似度，所述用戶相似度為每個用戶資料對中的兩組用戶資料對應的用戶之間的相似度；

樣本資料確定模組603，用於根據所述每個用戶資料對所對應的用戶相似度和所述多個用戶資料對，確定用於訓練預設的分類模型的樣本資料；

模型訓練模組604，用於基於所述樣本資料對所述分類模型進行訓練，得到相似度分類模型。

本申請實施例中，所述相似度獲取模組602，包括：

生物特徵獲取單元，用於獲取第一用戶資料對所對應的用戶的生物特徵，其中，所述第一用戶資料對為所述多個用戶資料對中的任意用戶資料對；

相似度獲取單元，用於根據所述第一用戶資料對所對應的用戶的生物特徵，確定所述第一用戶資料對所對應的用戶相似度。

本申請實施例中，所述生物特徵包括面部圖像特徵，

所述生物特徵獲取單元，用於獲取第一用戶資料對所對應的用戶的面部圖像；對所述面部圖像進行特徵提取，得到面部圖像特徵；

相應的，所述相似度獲取單元，用於根據所述第一用戶資料對所對應的用戶的面部圖像特徵，確定所述第一用戶資料對所對應的用戶相似度。

本申請實施例中，所述生物特徵包括語音特徵，

所述生物特徵獲取單元，用於獲取第一用戶資料對所

對應的用戶的語音資料；對所述語音資料進行特徵提取，得到語音特徵；

相應的，所述相似度獲取單元，用於根據所述第一用戶資料對所對應的用戶的語音特徵，確定所述第一用戶資料對所對應的用戶相似度。

本申請實施例中，所述樣本資料確定模組603，包括：

特徵提取單元，用於對所述多個用戶資料對中的每組用戶資料對進行特徵提取，得到每個用戶資料對中的兩組用戶資料之間相關聯的用戶特徵；

樣本資料確定單元，用於根據所述每個用戶資料對中的兩組用戶資料之間相關聯的用戶特徵和所述每個用戶資料對所對應的用戶相似度，確定用於訓練分類模型的樣本資料。

本申請實施例中，所述樣本資料確定單元，用於根據每個用戶資料對所對應的用戶相似度和預定的相似度閾值，從所述多個用戶資料對所對應的用戶特徵中選取正樣本特徵和負樣本特徵；將所述正樣本特徵和負樣本特徵作為用於訓練分類模型的樣本資料。

本申請實施例中，所述用戶特徵包括戶籍維度特徵、姓名維度特徵、社交特徵和興趣愛好特徵；所述戶籍維度特徵包括用戶身份資訊的特徵，所述姓名維度特徵包括用戶姓名資訊的特徵和用戶姓氏的稀缺程度的特徵，所述社交特徵包括用戶的社會關係資訊的特徵。

本申請實施例中，所述正樣本特徵和負樣本特徵中包含的特徵數目相同。

本申請實施例中，所述相似度分類模型為二分類器模型。

本申請實施例提供一種模型的訓練裝置，透過獲取的多個用戶資料對，且每個用戶資料對中的兩組用戶資料的資料欄位有相同的部分，以及獲取的每個用戶資料對所對應的用戶相似度，確定用於訓練預設的分類模型的樣本資料，然後，基於樣本資料對分類模型進行訓練，得到相似度分類模型，以便後續可以透過相似度分類模型確定待測用戶資料對中的兩組待測用戶資料對應的用戶之間的相似度，這樣，僅透過相同的資料欄位得到多個用戶資料對，並透過用戶相似度確定每個用戶資料對中的兩組用戶資料對應的用戶之間的關聯關係，得到用於訓練預設的分類模型的樣本資料，而不需要人工標注即可得到樣本資料，可以實現模型訓練的快速完成，提高了模型訓練效率並減少資源消耗。

實施例四

以上為本申請實施例提供的模型的訓練裝置，基於同樣的思路，本申請實施例還提供一種資料相似度的確定裝置，如圖7所示。

所述資料相似度的確定裝置包括：待測資料獲取模組701、特徵提取模組702和相似度確定模組703，其中：

待測資料獲取模組701，用於獲取待測用戶資料對；

特徵提取模組702，用於對所述待測用戶資料對中每組待測用戶資料進行特徵提取，得到待測用戶特徵；

相似度確定模組703，用於根據所述待測用戶特徵和預先訓練的相似度分類模型，確定所述待測用戶資料對中的兩組待測用戶資料對應的用戶之間的相似度。

本申請實施例中，所述裝置還包括：

相似度分類別模組，用於如果所述待測用戶資料對中的兩組待測用戶資料對應的用戶之間的相似度大於預定相似度分類閾值，則確定所述待測用戶資料對所對應的待測用戶為雙胞胎。

本申請實施例提供一種資料相似度的確定裝置，透過獲取的多個用戶資料對，且每個用戶資料對中的兩組用戶資料的資料欄位有相同的部分，以及獲取的每個用戶資料對所對應的用戶相似度，確定用於訓練預設的分類模型的樣本資料，然後，基於樣本資料對分類模型進行訓練，得到相似度分類模型，以便後續可以透過相似度分類模型確定待測用戶資料對中的兩組待測用戶資料對應的用戶之間的相似度，這樣，僅透過相同的資料欄位得到多個用戶資料對，並透過用戶相似度確定每個用戶資料對中的兩組用戶資料對應的用戶之間的關聯關係，得到用於訓練預設的分類模型的樣本資料，而不需要人工標注即可得到樣本資料，可以實現模型訓練的快速完成，提高了模型訓練效率並減少資源消耗。

實施例五

基於同樣的思路，本申請實施例還提供一種模型的訓練設備，如圖8所示。

該模型的訓練設備可以為上述實施例提供的伺服器等。

模型的訓練設備可因配置或性能不同而產生比較大的差異，可以包括一個或一個以上的處理器801和記憶體802，記憶體802中可以儲存有一個或一個以上儲存應用程式或資料。其中，記憶體802可以是短暫儲存或持久儲存。儲存在記憶體802的應用程式可以包括一個或一個以上模組（圖示未示出），每個模組可以包括對模型的訓練設備中的一系列電腦可執行指令。更進一步地，處理器801可以設置為與記憶體802通信，在模型的訓練設備上執行記憶體802中的一系列電腦可執行指令。模型的訓練設備還可以包括一個或一個以上電源803，一個或一個以上有線或無線網路介面804，一個或一個以上輸入輸出介面805，一個或一個以上鍵盤806。

具體在本實施例中，模型的訓練設備包括有記憶體，以及一個或一個以上的程式，其中一個或者一個以上程式儲存於記憶體中，且一個或者一個以上程式可以包括一個或一個以上模組，且每個模組可以包括對模型的訓練設備中的一系列電腦可執行指令，且經配置以由一個或者一個以上處理器執行該一個或者一個以上套裝程式含用於進行

以下電腦可執行指令：

獲取多個用戶資料對，其中，所述每個用戶資料對中的兩組用戶資料的資料欄位有相同的部分；

獲取每個用戶資料對所對應的用戶相似度，所述用戶相似度為每個用戶資料對中的兩組用戶資料對應的用戶之間的相似度；

根據所述每個用戶資料對所對應的用戶相似度和所述多個用戶資料對，確定用於訓練預設的分類模型的樣本資料；

基於所述樣本資料對所述分類模型進行訓練，得到相似度分類模型。

可選地，所述可執行指令在被執行時，還可以使所述處理器：

獲取第一用戶資料對所對應的用戶的生物特徵，其中，所述第一用戶資料對為所述多個用戶資料對中的任意用戶資料對；

根據所述第一用戶資料對所對應的用戶的生物特徵，確定所述第一用戶資料對所對應的用戶相似度。

可選地，所述可執行指令在被執行時，還可以使所述處理器：

所述生物特徵包括面部圖像特徵，

所述獲取第一用戶資料對所對應的用戶的生物特徵，包括：

獲取第一用戶資料對所對應的用戶的面部圖像；

對所述面部圖像進行特徵提取，得到面部圖像特徵；

相應的，所述根據所述第一用戶資料對所對應的用戶的生物特徵，確定所述第一用戶資料對所對應的用戶相似度，包括：

根據所述第一用戶資料對所對應的用戶的面部圖像特徵，確定所述第一用戶資料對所對應的用戶相似度。

可選地，所述可執行指令在被執行時，還可以使所述處理器：

所述生物特徵包括語音特徵，

所述獲取第一用戶資料對所對應的用戶的生物特徵，包括：

獲取第一用戶資料對所對應的用戶的語音資料；

對所述語音資料進行特徵提取，得到語音特徵；

相應的，所述根據所述第一用戶資料對所對應的用戶的生物特徵，確定所述第一用戶資料對所對應的用戶相似度，包括：

根據所述第一用戶資料對所對應的用戶的語音特徵，確定所述第一用戶資料對所對應的用戶相似度。

可選地，所述可執行指令在被執行時，還可以使所述處理器：

對所述多個用戶資料對中的每個用戶資料對進行特徵提取，得到每個用戶資料對中的兩組用戶資料之間相關聯的用戶特徵；

根據所述每個用戶資料對中的兩組用戶資料之間相關

聯的用戶特徵和所述每個用戶資料對所對應的用戶相似度，確定用於訓練分類模型的樣本資料。

可選地，所述可執行指令在被執行時，還可以使所述處理器：

根據每個用戶資料對所對應的用戶相似度和預定的相似度閾值，從所述多個用戶資料對所對應的用戶特徵中選取正樣本特徵和負樣本特徵；

將所述正樣本特徵和負樣本特徵作為用於訓練分類模型的樣本資料。

可選地，所述用戶特徵包括戶籍維度特徵、姓名維度特徵、社交特徵和興趣愛好特徵；所述戶籍維度特徵包括用戶身份資訊的特徵，所述姓名維度特徵包括用戶姓名資訊的特徵和用戶姓氏的稀缺程度的特徵，所述社交特徵包括用戶的社會關係資訊的特徵。

可選地，所述正樣本特徵和負樣本特徵中包含的特徵數目相同。

可選地，所述相似度分類模型為二分類器模型。

本申請實施例提供一種模型的訓練設備，透過獲取的多個用戶資料對，且每個用戶資料對中的兩組用戶資料的資料欄位有相同的部分，以及獲取的每個用戶資料對所對應的用戶相似度，確定用於訓練預設的分類模型的樣本資料，然後，基於樣本資料對分類模型進行訓練，得到相似度分類模型，以便後續可以透過相似度分類模型確定待測用戶資料對中的兩組待測用戶資料對應的用戶之間的相似

度，這樣，僅透過相同的資料欄位得到多個用戶資料對，並透過用戶相似度確定每個用戶資料對中的兩組用戶資料對應的用戶之間的關聯關係，得到用於訓練預設的分類模型的樣本資料，而不需要人工標注即可得到樣本資料，可以實現模型訓練的快速完成，提高了模型訓練效率並減少資源消耗。

實施例六

基於同樣的思路，本申請實施例還提供一種資料相似度的確定設備，如圖9所示。

該資料相似度的確定設備可以為上述實施例提供的伺服器或終端設備等。

資料相似度的確定設備可因配置或性能不同而產生比較大的差異，可以包括一個或一個以上的處理器901和記憶體902，記憶體902中可以儲存有一個或一個以上儲存應用程式或資料。其中，記憶體902可以是短暫儲存或持久儲存。儲存在記憶體902的應用程式可以包括一個或一個以上模組（圖示未示出），每個模組可以包括對資料相似度的確定設備中的一系列電腦可執行指令。更進一步地，處理器901可以設置為與記憶體902通信，在資料相似度的確定設備上執行記憶體902中的一系列電腦可執行指令。資料相似度的確定設備還可以包括一個或一個以上電源903，一個或一個以上有線或無線網路介面904，一個或一個以上輸入輸出介面905，一個或一個以上鍵盤906。

具體在本實施例中，資料相似度的確定設備包括有記憶體，以及一個或一個以上的程式，其中一個或者一個以上程式儲存於記憶體中，且一個或者一個以上程式可以包括一個或一個以上模組，且每個模組可以包括對資料相似度的確定設備中的一系列電腦可執行指令，且經配置以由一個或者一個以上處理器執行該一個或者一個以上套裝程式含用於進行以下電腦可執行指令：

獲取待測用戶資料對；

對所述待測用戶資料對中每組待測用戶資料進行特徵提取，得到待測用戶特徵；

根據所述待測用戶特徵和預先訓練的相似度分類模型，確定所述待測用戶資料對中的兩組待測用戶資料對應的用戶之間的相似度。

可選地，所述可執行指令在被執行時，還可以使所述處理器：

如果所述待測用戶資料對中的兩組待測用戶資料對應的用戶之間的相似度大於預定相似度分類閾值，則確定所述待測用戶資料對所對應的待測用戶為雙胞胎。

本申請實施例提供一種資料相似度的確定設備，透過獲取的多個用戶資料對，且每個用戶資料對中的兩組用戶資料的資料欄位有相同的部分，以及獲取的每個用戶資料對所對應的用戶相似度，確定用於訓練預設的分類模型的樣本資料，然後，基於樣本資料對分類模型進行訓練，得到相似度分類模型，以便後續可以透過相似度分類模型確

定待測用戶資料對中的兩組待測用戶資料對應的用戶之間的相似度，這樣，僅透過相同的資料欄位得到多個用戶資料對，並透過用戶相似度確定每個用戶資料對中的兩組用戶資料對應的用戶之間的關聯關係，得到用於訓練預設的分類模型的樣本資料，而不需要人工標注即可得到樣本資料，可以實現模型訓練的快速完成，提高了模型訓練效率並減少資源消耗。

上述對本說明書特定實施例進行了描述。其它實施例在所附申請專利範圍的範圍內。在一些情況下，在申請專利範圍中記載的動作或步驟可以按照不同於實施例中的順序來執行並且仍然可以實現期望的結果。另外，在附圖中描繪的過程不一定要求示出的特定順序或者連續順序才能實現期望的結果。在某些實施方式中，多工處理和並行處理也是可以的或者可能是有利的。

在20世紀90年代，對於一個技術的改進可以很明顯地區分是硬體上的改進（例如，對二極體、電晶體、開關等電路結構的改進）還是軟體上的改進（對於方法流程的改進）。然而，隨著技術的發展，當今的很多方法流程的改進已經可以視為硬體電路結構的直接改進。設計人員幾乎都透過將改進的方法流程程式設計到硬體電路中來得到相應的硬體電路結構。因此，不能說一個方法流程的改進就不能用硬體實體模組來實現。例如，可程式設計邏輯裝置（Programmable Logic Device，PLD）（例如現場可程式設計閘陣列（Field Programmable Gate Array，FPGA））

就是這樣一種積體電路，其邏輯功能由用戶對裝置程式設計來確定。由設計人員自行程式設計來把一個數位系統“集成”在一片PLD上，而不需要請晶片製造廠商來設計和製作專用的積體電路晶片。而且，如今，取代手工地製作積體電路晶片，這種程式設計也多半改用“邏輯編譯器（logic compiler）”軟體來實現，它與程式開發撰寫時所用的軟體編譯器相類似，而要編譯之前的原始代碼也得用特定的程式設計語言來撰寫，此稱之為硬體描述語言（Hardware Description Language，HDL），而HDL也並非僅有一種，而是有許多種，如ABEL（Advanced Boolean Expression Language）、AHDL（Altera Hardware Description Language）、Confluence、CUPL（Cornell University Programming Language）、HDCal、JHDL（Java Hardware Description Language）、Lava、Lola、MyHDL、PALASM、RHDL（Ruby Hardware Description Language）等，目前最普遍使用的是VHDL（Very-High-Speed Integrated Circuit Hardware Description Language）與Verilog。本領域技術人員也應該清楚，只需要將方法流程用上述幾種硬體描述語言稍作邏輯程式設計並程式設計到積體電路中，就可以很容易得到實現該邏輯方法流程的硬體電路。

控制器可以按任何適當的方式實現，例如，控制器可以採取例如微處理器或處理器以及儲存可由該（微）處理器執行的電腦可讀程式碼（例如軟體或韌體）的電腦可讀

媒體、邏輯閘、開關、專用積體電路（Application Specific Integrated Circuit，ASIC）、可程式設計邏輯控制器和嵌入微控制器的形式，控制器的例子包括但不限於以下微控制器：ARC 625D、Atmel AT91SAM、Microchip PIC18F26K20以及Silicone Labs C8051F320，記憶體控制器還可以被實現為記憶體的控制邏輯的一部分。本領域技術人員也知道，除了以純電腦可讀程式碼方式實現控制器以外，完全可以透過將方法步驟進行邏輯程式設計來使得控制器以邏輯閘、開關、專用積體電路、可程式設計邏輯控制器和嵌入微控制器等的形式來實現相同功能。因此這種控制器可以被認為是一種硬體部件，而對其內包括的用於實現各種功能的裝置也可以視為硬體部件內的結構。或者甚至，可以將用於實現各種功能的裝置視為既可以是實現方法的軟體模組又可以是硬體部件內的結構。

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。具體的，電腦例如可以為個人電腦、膝上型電腦、蜂窩電話、相機電話、智慧型電話、個人數位助理、媒體播放機、導航設備、電子郵件設備、遊戲控制台、平板電腦、可穿戴設備或者這些設備中的任何設備的組合。

為了描述的方便，描述以上裝置時以功能分為各種單元分別描述。當然，在實施本申請時可以把各單元的功能在同一個或多個軟體和／或硬體中實現。

本領域內的技術人員應明白，本申請的實施例可提供為方法、系統、或電腦程式產品。因此，本申請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，本申請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。

本申請是參照根據本申請實施例的方法、設備（系統）、和電腦程式產品的流程圖和／或方框圖來描述的。應理解可由電腦程式指令實現流程圖和／或方框圖中的每一流程和／或方框、以及流程圖和／或方框圖中的流程和／或方框的結合。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設備的處理器以產生一個機器，使得透過電腦或其他可程式設計資料處理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／或方框圖一個方框或多個方框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／或方框圖一個方框或多個方框中指定的功能。

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設備上，使得在電腦或其他可程式設計設備上執

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／或方框圖一個方框或多個方框中指定的功能的步驟。

在一個典型的配置中，計算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／或非易失性記憶體等形式，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示例。

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）、電可擦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訪問的資訊。按照本文中的界定，電腦可讀媒體不包括暫存電腦可讀媒體（transitory media），如調製的資料信號和載波。

還需要說明的是，術語“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、商品或者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、商品或者設備所固有的要素。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、商品或者設備中還存在另外的相同要素。

本領域技術人員應明白，本申請的實施例可提供為方法、系統或電腦程式產品。因此，本申請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，本申請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。

本申請可以在由電腦執行的電腦可執行指令的一般上下文中描述，例如程式模組。一般地，程式模組包括執行特定任務或實現特定抽象資料類型的常式、程式、物件、元件、資料結構等等。也可以在分散式運算環境中實踐本申請，在這些分散式運算環境中，由透過通信網路而被連接的遠端處理設備來執行任務。在分散式運算環境中，程式模組可以位於包括存放裝置在內的本地和遠端電腦儲存媒體中。

本說明書中的各個實施例均採用遞進的方式描述，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，對於系統實施例而言，由於其基本相似於方法實施例，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。

以上所述僅為本申請的實施例而已，並不用於限制本申請。對於本領域技術人員來說，本申請可以有各種更改和變化。凡在本申請的精神和原理之內所作的任何修改、等同替換、改進等，均應包含在本申請的申請專利範圍之內。

【符號說明】

601：資料獲取模組

602：相似度獲取模組

603：樣本資料確定模組

604：模型訓練模組

701：待測資料獲取模組

702：特徵提取模組

703：相似度確定模組

801：處理器

802：記憶體

803：電源

804：有線或無線網路介面

805：輸入輸出介面

806：鍵盤

901：處理器

902：記憶體

903：電源

904：有線或無線網路介面

905：輸入輸出介面

906：鍵盤

201909005

【發明摘要】

【中文發明名稱】

模型的訓練方法、資料相似度的確定方法、裝置及設備

【中文】

本申請實施例公開了一種模型的訓練方法、資料相似度的確定方法、裝置及設備，該模型的訓練方法包括：獲取多個用戶資料對，其中，所述每個用戶資料對中的兩組用戶資料的資料欄位有相同的部分；獲取每個用戶資料對所對應的用戶相似度，所述用戶相似度為每個用戶資料對中的兩組用戶資料對應的用戶之間的相似度；根據所述每個用戶資料對所對應的用戶相似度和所述多個用戶資料對，確定用於訓練預設的分類模型的樣本資料；基於所述樣本資料對所述分類模型進行訓練，得到相似度分類模型。利用本申請實施例，可以實現模型的快速訓練，提高模型訓練效率並減少資源消耗。

【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

【發明申請專利範圍】

【第1項】

一種模型的訓練方法，其特徵在於，該方法包括：

獲取多個用戶資料對，其中，該每個用戶資料對中的兩組用戶資料的資料欄位有相同的部分；

獲取每個用戶資料對所對應的用戶相似度，該用戶相似度為每個用戶資料對中的兩組用戶資料對應的用戶之間的相似度；

根據該每個用戶資料對所對應的用戶相似度和該多個用戶資料對，確定用於訓練預設的分類模型的樣本資料；

基於該樣本資料對該分類模型進行訓練，得到相似度分類模型。

【第2項】

根據請求項1所述的方法，其中，所述獲取每個用戶資料對所對應的用戶相似度，包括：

獲取第一用戶資料對所對應的用戶的生物特徵，其中，該第一用戶資料對為該多個用戶資料對中的任意用戶資料對；

根據該第一用戶資料對所對應的用戶的生物特徵，確定該第一用戶資料對所對應的用戶相似度。

【第3項】

根據請求項2所述的方法，其中，該生物特徵包括面部圖像特徵，

所述獲取第一用戶資料對所對應的用戶的生物特徵，

包括：

獲取第一用戶資料對所對應的用戶的面部圖像；

對該面部圖像進行特徵提取，得到面部圖像特徵；

相應的，所述根據該第一用戶資料對所對應的用戶的生物特徵，確定該第一用戶資料對所對應的用戶相似度，包括：

根據該第一用戶資料對所對應的用戶的面部圖像特徵，確定該第一用戶資料對所對應的用戶相似度。

【第4項】

根據請求項2所述的方法，其中，該生物特徵包括語音特徵，

所述獲取第一用戶資料對所對應的用戶的生物特徵，包括：

獲取第一用戶資料對所對應的用戶的語音資料；

對該語音資料進行特徵提取，得到語音特徵；

相應的，所述根據該第一用戶資料對所對應的用戶的生物特徵，確定該第一用戶資料對所對應的用戶相似度，包括：

根據該第一用戶資料對所對應的用戶的語音特徵，確定該第一用戶資料對所對應的用戶相似度。

【第5項】

根據請求項1所述的方法，其中，所述根據該每個用戶資料對所對應的用戶相似度和該多個用戶資料對，確定用於訓練分類模型的樣本資料，包括：

對該多個用戶資料對中的每個用戶資料對進行特徵提取，得到每個用戶資料對中的兩組用戶資料之間相關聯的用戶特徵；

根據該每個用戶資料對中用戶資料之間相關聯的用戶特徵和該每個用戶資料對所對應的用戶相似度，確定用於訓練分類模型的樣本資料。

【第6項】

根據請求項5所述的方法，其中，所述根據該每個用戶資料對中的兩組用戶資料之間相關聯的用戶特徵和該每個用戶資料對所對應的用戶相似度，確定用於訓練分類模型的樣本資料，包括：

根據每個用戶資料對所對應的用戶相似度和預定的相似度閾值，從該多個用戶資料對所對應的用戶特徵中選取正樣本特徵和負樣本特徵；

將該正樣本特徵和負樣本特徵作為用於訓練分類模型的樣本資料。

【第7項】

根據請求項6所述的方法，其中，該用戶特徵包括戶籍維度特徵、姓名維度特徵、社交特徵和興趣愛好特徵；該戶籍維度特徵包括用戶身份資訊的特徵，該姓名維度特徵包括用戶姓名資訊的特徵和用戶姓氏的稀缺程度的特徵，該社交特徵包括用戶的社會關係資訊的特徵。

【第8項】

根據請求項6所述的方法，其中，該正樣本特徵和負

樣本特徵中包含的特徵數目相同。

【第9項】

根據請求項1-8中任一項所述的方法，其中，該相似度分類模型為二分類器模型。

【第10項】

一種資料相似度的確定方法，其特徵在於，該方法包括：

獲取待測用戶資料對；

對該待測用戶資料對中每組待測用戶資料進行特徵提取，得到待測用戶特徵；

根據該待測用戶特徵和預先訓練的相似度分類模型，確定該待測用戶資料對中的兩組待測用戶資料對應的用戶之間的相似度。

【第11項】

根據請求項10所述的方法，其中，該方法還包括：

如果該待測用戶資料對中的兩組待測用戶資料對應的用戶之間的相似度大於預定相似度分類閾值，則確定該待測用戶資料對所對應的待測用戶為雙胞胎。

【第12項】

一種模型的訓練裝置，其特徵在於，該裝置包括：

資料獲取模組，用於獲取多個用戶資料對，其中，該每個用戶資料對中的兩組用戶資料的資料欄位有相同的部分；

相似度獲取模組，用於獲取每個用戶資料對所對應的

用戶相似度，該用戶相似度為每個用戶資料對中的兩組用戶資料對應的用戶之間的相似度；

樣本資料確定模組，用於根據該每個用戶資料對所對應的用戶相似度和該多個用戶資料對，確定用於訓練預設的分類模型的樣本資料；

模型訓練模組，用於基於該樣本資料對該分類模型進行訓練，得到相似度分類模型。

【第13項】

根據請求項12所述的裝置，其中，該相似度獲取模組，包括：

生物特徵獲取單元，用於獲取第一用戶資料對所對應的用戶的生物特徵，其中，該第一用戶資料對為該多個用戶資料對中的任意用戶資料對；

相似度獲取單元，用於根據該第一用戶資料對所對應的用戶的生物特徵，確定該第一用戶資料對所對應的用戶相似度。

【第14項】

根據請求項13所述的裝置，其中，該生物特徵包括面部圖像特徵，

該生物特徵獲取單元，用於獲取第一用戶資料對所對應的用戶的面部圖像；對該面部圖像進行特徵提取，得到面部圖像特徵；

相應的，該相似度獲取單元，用於根據該第一用戶資料對所對應的用戶的面部圖像特徵，確定該第一用戶資料

對所對應的用戶相似度。

【第15項】

根據請求項13所述的裝置，其中，該生物特徵包括語音特徵，

該生物特徵獲取單元，用於獲取第一用戶資料對所對應的用戶的語音資料；對該語音資料進行特徵提取，得到語音特徵；

相應的，該相似度獲取單元，用於根據該第一用戶資料對所對應的用戶的語音特徵，確定該第一用戶資料對所對應的用戶相似度。

【第16項】

根據請求項12所述的裝置，其中，該樣本資料確定模組，包括：

特徵提取單元，用於對該多個用戶資料對中的每個用戶資料對進行特徵提取，得到每個用戶資料對中的兩組用戶資料之間相關聯的用戶特徵；

樣本資料確定單元，用於根據該每個用戶資料對中的兩組用戶資料之間相關聯的用戶特徵和該每個用戶資料對所對應的用戶相似度，確定用於訓練分類模型的樣本資料。

【第17項】

根據請求項16所述的裝置，其中，該樣本資料確定單元，用於根據每個用戶資料對所對應的用戶相似度和預定的相似度閾值，從該多個用戶資料對所對應的用戶特徵中

選取正樣本特徵和負樣本特徵；將該正樣本特徵和負樣本特徵作為用於訓練分類模型的樣本資料。

【第18項】

根據請求項17所述的裝置，其中，該用戶特徵包括戶籍維度特徵、姓名維度特徵、社交特徵和興趣愛好特徵；該戶籍維度特徵包括用戶身份資訊的特徵，該姓名維度特徵包括用戶姓名資訊的特徵和用戶姓氏的稀缺程度的特徵，該社交特徵包括用戶的社會關係資訊的特徵。

【第19項】

根據請求項17所述的裝置，其中，該正樣本特徵和負樣本特徵中包含的特徵數目相同。

【第20項】

根據請求項12-19中任一項所述的裝置，其中，該相似度分類模型為二分類器模型。

【第21項】

一種資料相似度的確定裝置，其特徵在於，該裝置包括：

待測資料獲取模組，用於獲取待測用戶資料對；

特徵提取模組，用於對該待測用戶資料對中每組待測用戶資料進行特徵提取，得到待測用戶特徵；

相似度確定模組，用於根據該待測用戶特徵和預先訓練的相似度分類模型，確定該待測用戶資料對中的兩組待測用戶資料對應的用戶之間的相似度。

【第22項】

根據請求項21所述的裝置，其中，該裝置還包括：

相似度分類別模組，用於如果該待測用戶資料對中的兩組待測用戶資料對應的用戶之間的相似度大於預定相似度分類閾值，則確定該待測用戶資料對所對應的待測用戶為雙胞胎。

【第23項】

一種模型的訓練設備，該設備包括：

處理器；以及

被安排成儲存電腦可執行指令的記憶體，該可執行指令在被執行時使該處理器執行以下操作：

獲取多個用戶資料對，其中，該每個用戶資料對中的兩組用戶資料的資料欄位有相同的部分；

獲取每個用戶資料對所對應的用戶相似度，該用戶相似度為每個用戶資料對中的兩組用戶資料對應的用戶之間的相似度；

根據該每個用戶資料對所對應的用戶相似度和該多個用戶資料對，確定用於訓練預設的分類模型的樣本資料；

基於該樣本資料對該分類模型進行訓練，得到相似度分類模型。

【第24項】

一種資料相似度的確定設備，該設備包括：

處理器；以及

被安排成儲存電腦可執行指令的記憶體，該可執行指令在被執行時使該處理器執行以下操作：

獲取待測用戶資料對；

對該待測用戶資料對中每組待測用戶資料進行特徵提取，得到待測用戶特徵；

根據該待測用戶特徵和預先訓練的相似度分類模型，確定該待測用戶資料對中的兩組待測用戶資料對應的用戶之間的相似度。