(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6092141号
(P6092141)
(45)発行日　平成29年3月8日(2017.3.8)　(24)登録日　平成29年2月17日(2017.2.17)

(51)Int.Cl.　F I
G06N 99/00 (2010.01)　G06N 99/00 153

請求項の数 4 (全 14 頁)

(21)出願番号　特願2014-46601(P2014-46601)
(22)出願日　平成26年3月10日(2014.3.10)
(65)公開番号　特開2015-170281(P2015-170281A)
(43)公開日　平成27年9月28日(2015.9.28)
審査請求日　平成28年2月15日(2016.2.15)

(73)特許権者　000004226
日本電信電話株式会社
東京都千代田区大手町一丁目５番１号
(74)代理人　110001519
特許業務法人太陽国際特許事務所
(72)発明者　貞光　九月
東京都千代田区大手町一丁目５番１号　日本電信電話株式会社内
(72)発明者　松尾　義博
東京都千代田区大手町一丁目５番１号　日本電信電話株式会社内
(72)発明者　浅野　久子
東京都千代田区大手町一丁目５番１号　日本電信電話株式会社内

最終頁に続く

(54)【発明の名称】データ解析装置、方法、及びプログラム

(57)【特許請求の範囲】
【請求項１】
正例及び負例の何れか一方を示すラベルが付与されている学習データに基づいて、対象データが正例であるか又は負例であるかを識別するための識別モデルを学習し、学習された識別モデルに基づいて、前記ラベルが付与されていない未知データに前記ラベルを付与し、前記未知データに前記ラベルが付与された結果に基づいて、前記識別モデルを学習することを繰り返すブートストラップ法に従って、前記識別モデルを繰り返し学習したときに、繰り返し毎に前記未知データに前記ラベルが付与された結果から得られる、前記正例のラベルが付与された未知データの各々からなる正例集合の入力を受け付ける入力部と、
前記入力部によって受け付けた繰り返し毎の前記正例集合のうち、繰り返し開始から予め定められた第１の繰り返し回数までに得られた前記正例集合を擬似正例データとして選択し、予め定められた第２の繰り返し回数から繰り返し終了までに得られた前記正例集合を擬似負例データとして選択する擬似正例負例選択部と、
前記擬似正例負例選択部によって選択された前記擬似正例データ、及び前記擬似負例データから抽出された素性に基づいて、前記識別モデルを学習するモデル学習部と、
識別対象データの各々について、前記モデル学習部によって学習された前記識別モデルと、前記識別対象データから抽出された素性とに基づいて、前記識別対象データが正例である度合いを表すスコアを算出するスコアリング部と、
前記識別対象データの各々について、前記スコアリング部によって算出されたスコアが、予め定められた第１の閾値より大きい場合には、前記識別対象データが正例であると判

定し、前記スコアリング部によって算出されたスコアが、予め定められた第２の閾値より小さい場合には、前記識別対象データが負例であると判定する判定部と、
を含むデータ解析装置。
【請求項２】
前記識別対象データを、前記正例集合とし、
前記判定部は、前記ブートストラップ法における各繰り返しにおいて、前記正例のラベルが付与された未知データの数、及び前記負例のラベルが付与された未知データの数の割合に基づいて、前記第２の閾値を変化させ、
前記識別対象データの各々について、前記スコアリング部によって算出されたスコアが、前記第１の閾値より大きい場合には、前記識別対象データが正例であると判定し、前記スコアリング部によって算出されたスコアが、前記変化した第２の閾値より小さい場合には、前記識別対象データが負例であると判定する
請求項１記載のデータ解析装置。
【請求項３】
入力部が、正例及び負例の何れか一方を示すラベルが付与されている学習データに基づいて、対象データが正例であるか又は負例であるかを識別するための識別モデルを学習し、学習された識別モデルに基づいて、前記ラベルが付与されていない未知データに前記ラベルを付与し、前記未知データに前記ラベルが付与された結果に基づいて、前記識別モデルを学習することを繰り返すブートストラップ法に従って、前記識別モデルを繰り返し学習したときに、繰り返し毎に前記未知データに前記ラベルが付与された結果から得られる、前記正例のラベルが付与された未知データの各々からなる正例集合の入力を受け付けるステップと、
擬似正例負例選択部が、前記入力部によって受け付けた繰り返し毎の前記正例集合のうち、繰り返し開始から予め定められた第１の繰り返し回数までに得られた前記正例集合を擬似正例データとして選択し、予め定められた第２の繰り返し回数から繰り返し終了までに得られた前記正例集合を擬似負例データとして選択するステップと、
モデル学習部が、前記擬似正例負例選択部によって選択された前記擬似正例データ、及び前記擬似負例データから抽出された素性に基づいて、前記識別モデルを学習するステップと、
スコアリング部が、識別対象データの各々について、前記モデル学習部によって学習された前記識別モデルと、前記識別対象データから抽出された素性とに基づいて、前記識別対象データが正例である度合いを表すスコアを算出するステップと、
判定部が、前記識別対象データの各々について、前記スコアリング部によって算出されたスコアが、予め定められた第１の閾値より大きい場合には、前記識別対象データが正例であると判定し、前記スコアリング部によって算出されたスコアが、予め定められた第２の閾値より小さい場合には、前記識別対象データが負例であると判定するステップと、
を含むデータ解析方法。
【請求項４】
コンピュータに、請求項１又は請求項２記載のデータ解析装置の各部として機能させるためのプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、データ解析装置、方法、及びプログラムに係り、特に、識別対象データが正例であるか負例であるかを判定するデータ解析装置、方法、及びプログラムに関する。
【背景技術】
【０００２】
一度学習したモデルに基づいて未知事例の識別を行い、識別結果の正例・負例を新たな学習データとして用いていく繰り返し学習の枠組みをブートストラップ法と呼ぶ。
【０００３】

例えば、少量の教師信号となる単語または文字列のセットを入力とし、未知の単語または文字列が正例か負例なのかを逐次的に識別していくブートストラップを行う手法がこれまでに数多く提案されている（例えば、非特許文献１）。

【０００４】

また、ブートストラップの過程に着目し、初期のイテレーションで正例と判定された事例と終盤のイテレーションで正例と判定された事例を分け、識別対象となる任意の事例が、これら２つのグループのうち、いずれに近いかを、単純な情報量（分布類似度）に基づいて判定することが提案されている（例えば、非特許文献２）。

【先行技術文献】

【非特許文献】

【０００５】

【非特許文献１】Patrick Pantel and Marco Pennacchiotti., “Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations.”, COLING-ACL, 2006.

【非特許文献２】Tara Mcintosh, “Unsupervised discovery of negative categories in lexicon bootstrapping”, ACL2010, p.356-365

【発明の概要】

【発明が解決しようとする課題】

【０００６】

単語や文字列に対する自動ラベリング等のタスクにおいて、その判別の基準となる正例のみが与えられる場合は多い。

【０００７】

特に、上記非特許文献１に記載の技術のようにブートストラップ法を用いる手法においては、イテレーションが進むほど、正例として識別したい対象とは異なる事例を誤って正例を識別するセマンティックドリフトと呼ばれる現象が頻発する。

【０００８】

また、正例のみが与えられる場合には、負例が存在しないために、どこまでを正例識別対象とするのかといった基準が、システムには自明でなく、そのため自動的な負例判別法が必要とされる。

【０００９】

また、上記非特許文献２に記載の技術では、単純な情報量しか用いないため、より高度な素性（例えば単語の連鎖情報や、文書全体のトピック情報等）を考慮できないので、負例検出性能が達成されない。

【００１０】

本発明は、上記の事情を鑑みてなされたもので、負例データを精度よく抽出することができるデータ解析装置を提供することを目的とする。

【課題を解決するための手段】

【００１１】

上記の目的を達成するために本発明に係るデータ解析装置は、正例及び負例の何れか一方を示すラベルが付与されている学習データに基づいて、対象データが正例であるか又は負例であるかを識別するための識別モデルを学習し、学習された識別モデルに基づいて、前記ラベルが付与されていない未知データに前記ラベルを付与し、前記未知データに前記ラベルが付与された結果に基づいて、前記識別モデルを学習することを繰り返すブートストラップ法に従って、前記識別モデルを繰り返し学習したときに、繰り返し毎に前記未知データに前記ラベルが付与された結果から得られる、前記正例のラベルが付与された未知データの各々からなる正例集合の入力を受け付ける入力部と、前記入力部によって受け付けた繰り返し毎の前記正例集合のうち、繰り返し開始から予め定められた第１の繰り返し回数までに得られた前記正例集合を擬似正例データとして選択し、予め定められた第２の繰り返し回数から繰り返し終了までに得られた前記正例集合を擬似負例データとして選択する擬似正例負例選択部と、前記擬似正例負例選択部によって選択された前記擬似正例デ

ータ、及び前記擬似負例データから抽出された素性に基づいて、前記識別モデルを学習するモデル学習部と、識別対象データの各々について、前記モデル学習部によって学習された前記識別モデルと、前記識別対象データから抽出された素性とに基づいて、前記識別対象データが正例である度合いを表すスコアを算出するスコアリング部と、前記識別対象データの各々について、前記スコアリング部によって算出されたスコアが、予め定められた第１の閾値より大きい場合には、前記識別対象データが正例であると判定し、前記スコアリング部によって算出されたスコアが、予め定められた第２の閾値より小さい場合には、前記識別対象データが負例であると判定する判定部と、を含んで構成されている。

【００１２】

本発明に係るデータ解析方法は、入力部が、正例及び負例の何れか一方を示すラベルが付与されている学習データに基づいて、対象データが正例であるか又は負例であるかを識別するための識別モデルを学習し、学習された識別モデルに基づいて、前記ラベルが付与されていない未知データに前記ラベルを付与し、前記未知データに前記ラベルが付与された結果に基づいて、前記識別モデルを学習することを繰り返すブートストラップ法に従って、前記識別モデルを繰り返し学習したときに、繰り返し毎に前記未知データに前記ラベルが付与された結果から得られる、前記正例のラベルが付与された未知データの各々からなる正例集合の入力を受け付けるステップと、擬似正例負例選択部が、前記入力部によって受け付けた繰り返し毎の前記正例集合のうち、繰り返し開始から予め定められた第１の繰り返し回数までに得られた前記正例集合を擬似正例データとして選択し、予め定められた第２の繰り返し回数から繰り返し終了までに得られた前記正例集合を擬似負例データとして選択するステップと、モデル学習部が、前記擬似正例負例選択部によって選択された前記擬似正例データ、及び前記擬似負例データから抽出された素性に基づいて、前記識別モデルを学習するステップと、スコアリング部が、識別対象データの各々について、前記モデル学習部によって学習された前記識別モデルと、前記識別対象データから抽出された素性とに基づいて、前記識別対象データが正例である度合いを表すスコアを算出するステップと、判定部が、前記識別対象データの各々について、前記スコアリング部によって算出されたスコアが、予め定められた第１の閾値より大きい場合には、前記識別対象データが正例であると判定し、前記スコアリング部によって算出されたスコアが、予め定められた第２の閾値より小さい場合には、前記識別対象データが負例であると判定するステップと、を含んで構成されている。

【００１３】

本発明は、前記識別対象データを、前記正例集合とし、前記判定部は、前記ブートストラップ法における各繰り返しにおいて、前記正例のラベルが付与された未知データの数、及び前記負例のラベルが付与された未知データの数の割合に基づいて、前記第２の閾値を変化させ、前記識別対象データの各々について、前記スコアリング部によって算出されたスコアが、前記第１の閾値より大きい場合には、前記識別対象データが正例であると判定し、前記スコアリング部によって算出されたスコアが、前記変化した第２の閾値より小さい場合には、前記識別対象データが負例であると判定するようにすることができる。

【００１４】

本発明のプログラムは、コンピュータに、上記のデータ解析装置の各部として機能させるためのプログラムである。

【発明の効果】

【００１５】

以上説明したように、本発明のデータ解析装置、方法、及びプログラムによれば、繰り返し毎の正例集合のうち、繰り返し開始から予め定められた第１の繰り返し回数までに得られた正例集合を擬似正例データとして選択し、予め定められた第２の繰り返し回数から繰り返し終了までに得られた正例集合を擬似負例データとして選択し、選択された擬似正例データ、及び擬似負例データから抽出された素性に基づいて、識別モデルを学習し、識別対象データの各々について、学習された識別モデルと、識別対象データから抽出された素性とに基づいて、識別対象データが正例である度合いを表すスコアを算出し、スコアが

予め定められた第２の閾値より小さい場合には、識別対象データが負例であると判定することにより、負例データを精度よく抽出することができる、という効果が得られる。

【図面の簡単な説明】

【００１６】

【図１】本発明の実施の形態に係るデータ解析装置の構成を示す概略図である。

【図２】本発明の第１の実施の形態に係るデータ解析装置におけるデータ解析処理ルーチンの内容を示すフローチャートである。

【図３】本発明の第２の実施の形態に係るデータ解析装置におけるデータ解析処理ルーチンの内容を示すフローチャートである。

【発明を実施するための形態】

【００１７】

本発明の実施の形態は、単語や文字列等に対し、ある基準を用いて識別を行う際のモデル学習に必要となる負例を、自動的に抽出する技術である。以下、図面を参照して本発明の実施の形態を詳細に説明する。

【００１８】

＜第１の実施の形態＞

＜システム構成＞

本発明の第１の実施の形態に係るデータ解析装置１００は、入力された、ブートストラップ法で得られた正例集合から、新負例データを抽出する。このデータ解析装置１００は、ＣＰＵと、ＲＡＭと、後述するデータ解析処理ルーチンを実行するためのプログラムを記憶したＲＯＭとを備えたコンピュータで構成され、機能的には次に示すように構成されている。図１に示すように、データ解析装置１００は、入力部１０と、演算部２０と、出力部３０とを備えている。

【００１９】

本実施の形態におけるブートストラップ法は、まず、正例及び負例の何れか一方を示すラベルが付与されている学習データに基づいて、識別モデルを学習する。そして、学習された識別モデルに基づいて、ラベルが付与されていない未知データにラベルを付与し、未知データにラベルが付与された結果に基づいて、識別モデルを学習する。そして、この未知データに対するラベル付与と、識別モデルの学習とを繰り返す。なお、識別モデルは、対象データが正例であるか又は負例であるかを識別するためのモデルである。

【００２０】

本実施の形態に係るデータ解析装置１００には、ブートストラップ法に従って識別モデルを繰り返し学習したときに、繰り返し毎に、未知データにラベルが付与された結果から得られる、正例のラベルが付与された未知データの各々を含む正例集合が入力される。

【００２１】

入力部１０は、ブートストラップ法における繰り返し毎に得られた正例集合の入力を受け付ける。また、入力部１０は、後述する素性化参照用データを受け付ける。

【００２２】

例えば、対象データが単語であって、単語が車名であるかどうかを識別したいとする。この場合、識別モデルは、単語が車名である場合には当該対象データに正例を付与し、単語が車名でない場合には当該対象データに負例を付与する。この場合、ブートストラップ法における各繰り返しにおいて得られた正例集合の一例を以下に示す。

【００２３】

iteration 1：シビック，ヴィッツ
iteration 2：プリウス，スカイライン
…
iteration 9：カブ，ヘリ，フィット
iteration10：ブルートレイン，エアバス

【００２４】

演算部２０は、正例集合データベース２００と、擬似正例負例選択部２０２と、擬似正

例データベース２０４と、擬似負例データベース２０６と、識別対象データベース２０８と、素性化参照用データベース２１０と、素性化部２１２と、素性化済み訓練データベース２１４と、素性化済み識別対象データベース２１６と、モデル学習部２１８と、識別モデルデータベース２２０と、スコアリング部２２２と、判定部２２４と、新正例データベース２２６と、新負例データベース２２８とを備えている。

【００２５】

正例集合データベース２００には、入力部１０によって受け付けた、ブートストラップ法における繰り返し毎に得られた正例集合が格納される。

【００２６】

擬似正例負例選択部２０２は、正例集合データベース２００に格納された繰り返し毎の正例集合のうち、ブートストラップ法における繰り返し開始から予め定められた初期の第１の繰り返し回数Ｂまでに得られた正例集合を擬似正例データとして選択する。

【００２７】

また、擬似正例負例選択部２０２は、正例集合データベース２００に格納された繰り返し毎の正例集合のうち、予め定められた終盤の第２の繰り返し回数Ｅから繰り返し終了までに得られた正例集合を擬似負例データとして選択する。

【００２８】

そして、擬似正例負例選択部２０２は、それ以外の繰り返し（繰り返し回数Ｂ＋１～繰り返し回数Ｅ－１）において得られた正例集合を識別対象データとして選択する。

【００２９】

擬似正例データ、及び擬似負例データは、ブートストラップ法における各繰り返しの正例集合に対して、擬似的に正例・負例のラベルを付与したものである。例えば、対象データが単語であって、単語が車名であるかどうかを識別する場合の、擬似正例データ、擬似負例データ、及び識別対象データの一例を以下に示す。

【００３０】

（例）

擬似正例データ(iteration1を取得)：シビック，ヴィッツ

擬似負例データ(iteration10を取得)：ブルートレイン，エアバス

識別対象データ(iteration 2～9) プリウス，スカイライン，…，カブ，ヘリ

【００３１】

上記の例では、ブートストラップ法における繰り返し開始から１回目の繰り返しまでに得られた正例集合を擬似正例データとして選択し、１０回目の繰り返しから繰り返し終了までに得られた正例集合を擬似負例データとして選択している。そして、それ以外の繰り返し（２回目の繰り返しから９回目の繰り返しまで）において得られた正例集合を識別対象データとして選択している。

【００３２】

擬似正例データベース２０４には、擬似正例負例選択部２０２によって選択された擬似正例データが格納される。また、擬似負例データベース２０６には、擬似正例負例選択部２０２によって選択された擬似負例データが格納される。識別対象データベース２０８には、擬似正例負例選択部２０２によって選択された識別対象データが格納される。

【００３３】

素性化参照用データベース２１０には、入力部１０によって受け付けた素性化参照用データが格納される。素性化参照用データは、正例集合データベース２００に格納された正例集合に含まれる正例データの各々を、素性化するために必要となるデータである。素性化参照用データベース２１０には、例えば、形態素済みテキストや、文書に付与されたトピック情報等が、素性化参照用データとして格納されている。

【００３４】

なお、上述したように、例えば、対象データが単語であって、単語が車名であるかどうかを識別したい場合には、上記の正例集合の例で示した「シビック」「ヴィッツ」等の単語が含まれる文書についての形態素済み文書や、当該文書に付与されたトピック情報等が

、素性化参照用データとなる。
【００３５】
素性化部２１２は、擬似正例データベース２０４に格納された擬似正例データの各々と、擬似負例データベース２０６に格納された擬似負例データの各々とについて、素性化参照用データベース２１０に格納された素性化参照用データに基づいて、擬似正例データの素性、及び擬似負例データの素性を抽出し、素性化済み訓練データとする。素性化済み訓練データは、擬似正例データと擬似負例データとを、識別モデルを学習するために、素性関数を用いて変換したデータである。
【００３６】
また、素性化部２１２は、識別対象データベース２０８に格納された識別対象データの各々について、素性化参照用データベース２１０に格納された素性化参照用データに基づいて、識別対象データの素性を抽出し、素性化済み識別対象データとする。素性化済み識別対象データは、識別対象データを、識別モデルを適用するために、素性関数を用いて変換したデータである。
【００３７】
擬似正例データと擬似負例データとを、素性関数により変換したデータの一例を以下に示す。
【００３８】
例：（最初の＋１／－１は擬似正例／擬似負例を表す）
＋１　シビック
素性ＩＤ１：（１つ後ろから「に／乗る」が連鎖する回数＝）５
素性ＩＤ２：（「車」との共起回数＝）１００
－１　ブルートレイン
素性ＩＤ１：（１つ後ろから「に／乗る」が連鎖する回数＝）３
素性ＩＤ２：（「車」との共起回数＝）０
【００３９】
なお、具体的な素性関数は従来提案されたものでよく、例えば対象データが単語である場合には、単語と共起する任意の単語の出現数や、対象単語の直後に連鎖して出現した任意の単語の出現数などがあげられる。
【００４０】
素性化済み訓練データベース２１４には、素性化部２１２によって変換された素性化済み訓練データが格納される。
【００４１】
素性化済み識別対象データベース２１６には、素性化部２１２によって変換された素性化済み識別対象データが格納される。
【００４２】
モデル学習部２１８は、素性化済み訓練データベース２１４に格納された素性化済み訓練データに基づいて、識別モデルを学習する。なお識別モデルとしては、例えば、ＳＶＭやｌｏｇｉｓｔｉｃ回帰モデルなど、各素性に対し正負それぞれに対する重み（信頼度）を記憶するモデルを用いる。また、識別モデルの学習は、従来の機械学習法によって、識別モデルを学習する。学習方法についても従来の手法を用いてよい。
【００４３】
識別モデルデータベース２２０には、モデル学習部２１８によって学習された識別モデルが格納される。
【００４４】
スコアリング部２２２は、識別モデルデータベース２２０に格納された識別モデルと、素性化済み識別対象データベース２１６に格納されている素性化済み識別対象データとに基づいて、識別対象データの各々について、当該識別対象データが正例である度合いを表すスコアを算出する。
【００４５】

判定部２２４は、識別対象データの各々について、スコアリング部２２２によって算出されたスコアが、予め定められた第１の閾値Ｐより大きい場合には、当該識別対象データが正例であると判定する。また、判定部２２４は、識別対象データの各々について、スコアリング部２２２によって算出されたスコアが、予め定められた第２の閾値Ｎより小さい場合には、当該識別対象データが負例であると判定する。

【００４６】

新正例データベース２２６には、判定部２２４によって正例であると判定された識別対象データが、新正例データとして格納される。

【００４７】

新負例データベース２２８には、判定部２２４によって負例であると判定された識別対象データが、新負例データとして格納される。

【００４８】

出力部３０は、新正例データベース２２６に格納された新正例データの各々と、新負例データベース２２８に格納された新負例データの各々とを出力する。例えば、対象データが車の名称である場合に、出力部３０が出力するデータの一例を以下に示す。

【００４９】

例：
新正例データ：プリウス，スカイライン，フィット
新負例データ：カブ，ヘリ

【００５０】

＜データ解析装置の作用＞

次に、第１の実施の形態に係るデータ解析装置１００の作用について説明する。まず、ブートストラップ法における繰り返し毎に得られた正例集合と、正例集合に含まれる正例データ毎の素性化参照用データとが、データ解析装置１００に入力されると、データ解析装置１００によって、図２に示すデータ解析処理ルーチンが実行される。

【００５１】

まず、ステップＳ１００において、入力部１０によって、ブートストラップ法における繰り返し毎に得られた正例集合の入力を受け付ける。そして、入力部１０によって、ブートストラップ法における繰り返し毎に得られた正例集合を正例集合データベース２００に格納する。また、入力部１０によって、素性化参照用データを受け付け、素性化参照用データベース２１０に格納する。

【００５２】

ステップＳ１０２において、擬似正例負例選択部２０２によって、上記ステップＳ１００で正例集合データベース２００に格納された繰り返し毎の正例集合のうち、ブートストラップ法における繰り返し開始から予め定められた初期の第１の繰り返し回数Ｂまでに得られた正例集合を擬似正例データとして選択し、擬似正例データベース２０４に格納する。

【００５３】

ステップＳ１０４において、擬似正例負例選択部２０２によって、上記ステップＳ１００で正例集合データベース２００に格納された繰り返し毎の正例集合のうち、予め定められた終盤の第２の繰り返し回数Ｅから繰り返し終了までに得られた正例集合を擬似負例データとして選択し、擬似負例データベース２０６に格納する。

【００５４】

ステップＳ１０６において、擬似正例負例選択部２０２によって、それ以外の繰り返し（繰り返し回数Ｂ＋１～繰り返し回数Ｅ－１）において得られた正例集合を識別対象データとし、識別対象データベース２０８に格納する。

【００５５】

ステップＳ１０８において、素性化部２１２によって、上記ステップＳ１０２で擬似正例データベース２０４に格納された擬似正例データの各々と、上記ステップＳ１０４で擬似負例データベース２０６に格納された擬似負例データの各々について、上記ステップＳ

１００で素性化参照用データベース２１０に格納された素性化参照用データに基づいて、擬似正例データの素性、及び擬似負例データの素性を抽出し、素性化済み訓練データとする。そして、素性化部２１２によって、得られた素性化済み訓練データを、素性化済み訓練データベース２１４に格納する。

【００５６】

ステップＳ１１０において、素性化部２１２によって、上記ステップＳ１０６で識別対象データベース２０８に格納された識別対象データの各々について、上記ステップＳ１００で素性化参照用データベース２１０に格納された素性化参照用データに基づいて、識別対象データの素性を抽出し、素性化済み識別対象データとする。そして、素性化部２１２によって、得られた素性化済み識別対象データを、素性化済み識別対象データベース２１６に格納する。

【００５７】

ステップＳ１１２において、モデル学習部２１８によって、上記ステップＳ１０８で素性化済み訓練データベース２１４に格納された素性化済み訓練データに基づいて、識別モデルを学習し、識別モデルデータベース２２０に格納する。

【００５８】

ステップＳ１１４において、スコアリング部２２２によって、上記ステップＳ１１２で識別モデルデータベース２２０に格納された識別モデルと、上記ステップＳ１１０で素性化済み識別対象データベース２１６に格納された素性化済み識別対象データとに基づいて、識別対象データの各々について、当該識別対象データが正例である度合いを表すスコアを算出する。

【００５９】

ステップＳ１１６において、判定部２２４によって、識別対象データの各々について、上記ステップＳ１１４で算出されたスコアが、予め定められた第１の閾値Ｐより大きい場合には、当該識別対象データが正例であると判定する。また、判定部２２４によって、識別対象データの各々について、上記ステップＳ１１４で算出されたスコアが、予め定められた第２の閾値Ｎより小さい場合には、当該識別対象データが負例であると判定する。そして、判定部２２４によって、正例であると判定された識別対象データを新正例データとして新正例データベース２２６に格納する。また、判定部２２４によって、負例であると判定された識別対象データを新負例データとして新負例データベース２２８に格納する。

【００６０】

ステップＳ１１８において、上記ステップＳ１１６で新正例データベース２２６に格納された新正例データの各々と、新負例データベース２２８に格納された新負例データの各々とを、結果として出力して、データ解析処理ルーチンを終了する。

【００６１】

以上説明したように、第１の実施の形態に係るデータ解析装置によれば、ブートストラップ法における繰り返し毎の正例集合のうち、ブートストラップ法における繰り返し開始から予め定められた初期の第１の繰り返し回数Ｂまでに得られた正例集合を擬似正例データとして選択し、予め定められた終盤の第２の繰り返し回数Ｅから繰り返し終了までに得られた正例集合を擬似負例データとして選択し、選択された擬似正例データ、及び擬似負例データから抽出された素性に基づいて、識別モデルを学習し、識別対象データの各々について、学習された識別モデルと、識別対象データから抽出された素性とに基づいて、識別対象データが正例である度合いを表すスコアを算出し、スコアが予め定められた第２の閾値Ｎより小さい場合には、識別対象データが負例であると判定することにより、負例データを精度よく抽出することができる。

【００６２】

また、識別モデルとして、一般の識別モデル（サポートベクタマシンや最大エントロピー法）を使うことが可能であるため、自由度の高い情報を用いることが可能となる。

【００６３】

また、ブートストラップ法における繰り返し初期と繰り返し終盤とにおける正例データ

を、擬似正例データ、及び擬似負例データとみなし、自由度の高い素性表現と識別モデルを併用することで、semantic driftを抑制し、高い精度で新負例データを抽出することができる。

【００６４】

＜第２の実施の形態＞

次に、第２の実施の形態について説明する。なお、第１の実施の形態と同様の構成となる部分については、同一符号を付して説明を省略する。

【００６５】

第２の実施の形態では、ブートストラップ法の各繰り返しにおいて、正例・負例と判定された未知データの割合を用いて、semantic driftの度合いを推測し、それに応じて第２の閾値を変化させる点が、第１の実施の形態と異なっている。

【００６６】

＜システム構成＞

第２の実施の形態における入力部１２は、ブートストラップ法における繰り返し毎に、正例のラベルが付与された未知データの数、及び負例のラベルが付与された未知データの数の割合を更に受け付ける。

【００６７】

判定部２２２４は、入力部１２によって受け付けた、繰り返し毎の上記割合に基づいて、第２の閾値Ｎを変化させる。具体的には、判定部２２２４は、ブートストラップ法における繰り返し毎の上記割合に基づいて、第２の閾値Ｎを変化させる。

【００６８】

例えば、iteration１において５１：４９の割合で正例、負例が判定され、iteration９において４８：５２の割合で正例、負例が判定された場合、iterationが進行しても負例の量が増えていない。これはsemantic driftがそれほど生じていないことを表していると考えられるため、第２の閾値Ｎを引き下げるように変化させる（負例になりづらくする）。

【００６９】

一方、例えば、iteration１において９０：１０の割合で正例、負例が判定され、iteration９において１０：９０の割合で正例、負例が判定された場合、iterationが進行するに従い負例の量が増えている。これはsemantic driftが生じていることを表していると考えられるため、第２の閾値Ｎを引き下げないこととする。

【００７０】

このときの第２の閾値Ｎの変動量は、固定値でも良いし、semantic driftの度合いを示す数値に比例した値としてもよい。第２の閾値Ｎの変動量が、semantic driftの度合いを示す数値に比例した値とする場合には、例えば、semantic driftの度合いを、最終繰り返しでの負例の数と初期繰り返しでの負例の数との比とし（４９/５２）、当該比と一定量Ｋとの積を算出し、第２の閾値Ｎの変動量とする。

【００７１】

また、判定部２２２４は、識別対象データの各々について、スコアリング部２２２によって算出されたスコアが、第１の閾値Ｐより大きい場合には、識別対象データが正例であると判定し、スコアリング部２２２によって算出されたスコアが、変化した第２の閾値Ｎより小さい場合には、識別対象データが負例であると判定する。

【００７２】

＜データ解析装置の作用＞

次に、第２の実施の形態に係るデータ解析装置の作用について説明する。まず、ブートストラップ法における繰り返し毎に得られた正例集合と、正例集合に含まれる正例データ毎の素性化参照用データと、ブートストラップ法における繰り返し毎の、正例のラベルが付与された未知データの数、及び負例のラベルが付与された未知データの数の割合とが、データ解析装置に入力されると、データ解析装置によって、図３に示すデータ解析処理ルーチンが実行される。なお、第１の実施の形態と同様の処理については、同一符号を付し

て説明を省略する。

【００７３】

まず、ステップＳ２００において、入力部１２によって、ブートストラップ法における繰り返し毎に得られた正例集合の入力を受け付ける。そして、入力部１２によって、ブートストラップ法における繰り返し毎に得られた正例集合を正例集合データベース２００に格納する。また、入力部１２によって、素性化参照用データを受け付け、素性化参照用データベース２１０に格納する。また、入力部１２によって、正例のラベルが付与された未知データの数、及び負例のラベルが付与された未知データの数の割合を、ブートストラップ法における繰り返し毎に受け付ける。

【００７４】

そして、ステップＳ２１５において、判定部２２２４によって、上記ステップＳ２００で受け付けた、ブートストラップ法における繰り返し毎の、正例のラベルが付与された未知データの数、及び負例のラベルが付与された未知データの数の割合に応じて、第２の閾値Ｎを変化させる。

【００７５】

そして、ステップＳ２１６において、判定部２２２４によって、識別対象データの各々について、上記ステップＳ１１４で算出されたスコアが、予め定められた第１の閾値Ｐより大きい場合には、当該識別対象データが正例であると判定し、上記ステップＳ１１４で算出されたスコアが、上記ステップＳ２１５で変化させた第２の閾値Ｎより小さい場合には、当該識別対象データが負例であると判定する。そして、判定部２２２４によって、正例であると判定された識別対象データを新正例データとして新正例データベース２２６に格納する。また、判定部２２２４によって、負例であると判定された識別対象データを新負例データとして新負例データベース２２８に格納する。

【００７６】

以上説明したように、第２の実施の形態に係るデータ解析装置によれば、ブートストラップ法における各繰り返しにおいて、正例のラベルが付与された未知データの数、及び負例のラベルが付与された未知データの数の割合に基づいて、第２の閾値Ｎを変化させ、識別対象データの各々について、算出されたスコアが、変化した第２の閾値Ｎより小さい場合には、当該識別対象データが負例であると判定することにより、semantic driftの度合いを考慮して、負例データを精度よく抽出することができる。

【００７７】

なお、本発明は、上述した実施形態に限定されるものではなく、この発明の要旨を逸脱しない範囲内で様々な変形や応用が可能である。

【００７８】

例えば、本実施の形態では、識別対象データを、擬似正例データ及び擬似負例データとして選択された正例集合と異なる正例集合とする場合を例に説明したが、これに限定されるものではない。例えば、識別対象データに、擬似正例データ及び擬似負例データを含めてもよい。また、識別対象データに、負例集合を含めてもよい。

【００７９】

また、本実施の形態のデータ解析装置は、正例集合データベース２００、擬似正例データベース２０４、擬似負例データベース２０６、識別対象データベース２０８、素性化参照用データベース２１０、素性化済み訓練データベース２１４、素性化済み識別対象データベース２１６、識別モデルデータベース２２０、新正例データベース２２６、及び新負例データベース２２８を備えている場合について説明したが、例えば正例集合データベース２００、擬似正例データベース２０４、擬似負例データベース２０６、識別対象データベース２０８、素性化参照用データベース２１０、素性化済み訓練データベース２１４、素性化済み識別対象データベース２１６、識別モデルデータベース２２０、新正例データベース２２６、及び新負例データベース２２８の少なくとも１つがデータ解析装置の外部装置に設けられ、データ解析装置は、外部装置と通信手段を用いて通信することにより、正例集合データベース２００、擬似正例データベース２０４、擬似負例データベース２０

６、識別対象データベース２０８、素性化参照用データベース２１０、素性化済み訓練データベース２１４、素性化済み識別対象データベース２１６、識別モデルデータベース２２０、新正例データベース２２６、及び新負例データベース２２８の少なくとも１つを参照するようにしてもよい。

【００８０】

また、上述のデータ解析装置は、内部にコンピュータシステムを有しているが、「コンピュータシステム」は、ＷＷＷシステムを利用している場合であれば、ホームページ提供環境（あるいは表示環境）も含むものとする。

【００８１】

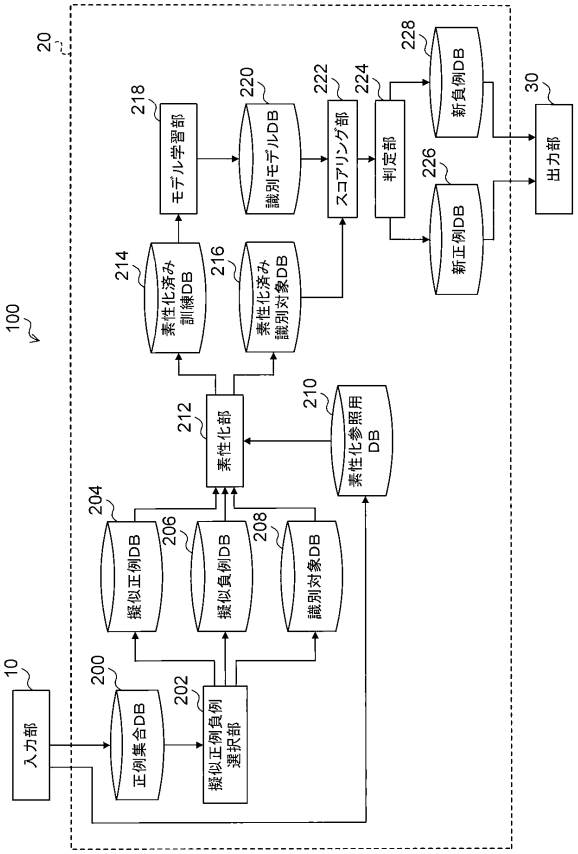
また、本願明細書中において、プログラムが予めインストールされている実施形態として説明したが、当該プログラムを、コンピュータ読み取り可能な記録媒体に格納して提供することも可能である。

【符号の説明】

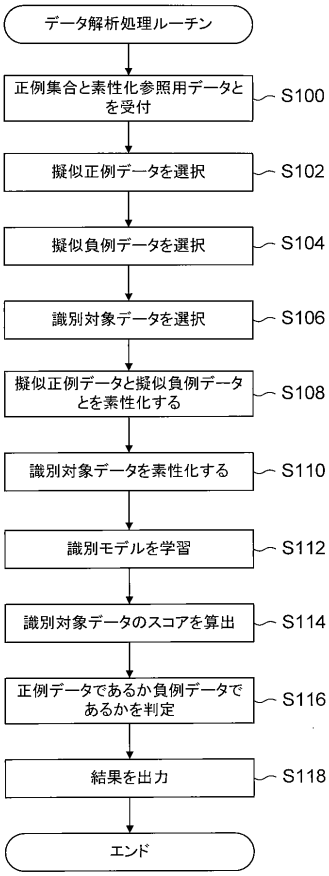
【００８２】

１０，１２　入力部
２０　演算部
３０　出力部
１００ データ解析装置
２００ 正例集合データベース
２０２ 擬似正例負例選択部
２０４ 擬似正例データベース
２０６ 擬似負例データベース
２０８ 識別対象データベース
２１０ 素性化参照用データベース
２１２ 素性化部
２１４ 素性化済み訓練データベース
２１６ 素性化済み識別対象データベース
２１８ モデル学習部
２２０ 識別モデルデータベース
２２２ スコアリング部
２２４，２２２４　判定部
２２６ 新正例データベース
２２８ 新負例データベース

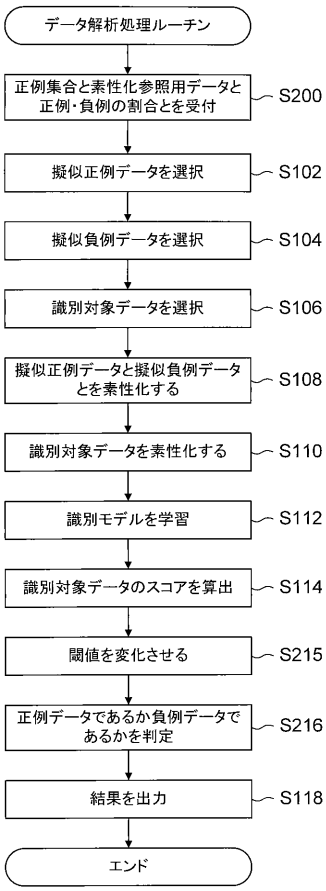
【図１】


20
100
10
入力部
200
正例集合DB
202
擬似正例負例選択部
204
擬似正例DB
206
擬似負例DB
208
識別対象DB
210
素性化参照用DB
212
素性化部
214
素性化済み訓練DB
216
素性化済み識別対象DB
218
モデル学習部
220
識別モデルDB
222
スコアリング部
224
判定部
226
新正例DB
228
新負例DB
30
出力部

【図２】


データ解析処理ルーチン
正例集合と素性化参照用データとを受付
S100
擬似正例データを選択
S102
擬似負例データを選択
S104
識別対象データを選択
S106
擬似正例データと擬似負例データとを素性化する
S108
識別対象データを素性化する
S110
識別モデルを学習
S112
識別対象データのスコアを算出
S114
正例データであるか負例データであるかを判定
S116
結果を出力
S118
エンド

【図３】


データ解析処理ルーチン
正例集合と素性化参照用データと正例・負例の割合とを受付
S200
擬似正例データを選択
S102
擬似負例データを選択
S104
識別対象データを選択
S106
擬似正例データと擬似負例データとを素性化する
S108
識別対象データを素性化する
S110
識別モデルを学習
S112
識別対象データのスコアを算出
S114
閾値を変化させる
S215
正例データであるか負例データであるかを判定
S216
結果を出力
S118
エンド

フロントページの続き

審査官　多胡　滋

(56)参考文献　貞光九月，外４名，トピック情報を用いたブートストラップ法に基づく語彙獲得，自然言語処理，日本，言語処理学会，２０１２年　７月　６日，第19巻，第2号，pp.89-106

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｎ　　９９／００