(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2009-282686
(P2009-282686A)
(43)公開日　平成21年12月3日(2009.12.3)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 17/30	(2006.01)	G06F	17/30	210D	5B075
G06N 5/04	(2006.01)	G06F	17/30	180B	5L096
G06T 7/00	(2006.01)	G06N	5/04	580A	
		G06N	5/04	580P	
		G06T	7/00	350B	

審査請求 未請求　請求項の数 10　ＯＬ　（全 16 頁）　最終頁に続く

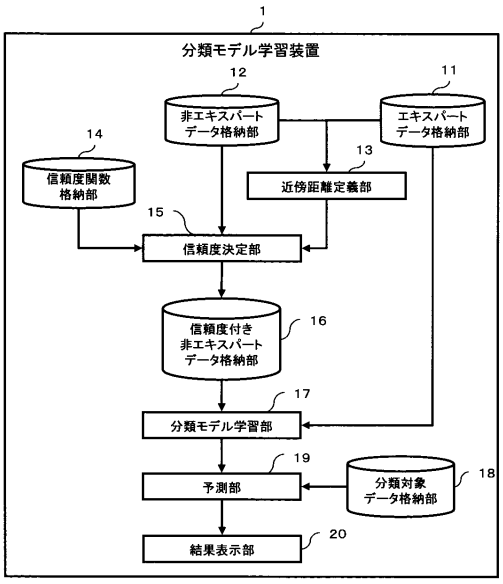
(21)出願番号　特願2008-133224 (P2008-133224)
(22)出願日　平成20年5月21日 (2008.5.21)

(71)出願人　000003078
株式会社東芝
東京都港区芝浦一丁目１番１号
(74)代理人　110000235
特許業務法人　天城国際特許事務所
(72)発明者　中田　康太
東京都港区芝浦一丁目１番１号　株式会社東芝内
(72)発明者　櫻井　茂明
東京都港区芝浦一丁目１番１号　株式会社東芝内
(72)発明者　折原　良平
東京都港区芝浦一丁目１番１号　株式会社東芝内

Ｆターム(参考)　5B075 NR12　QT10
5L096 BA06　FA66　GA30　KA04　MA07

(54)【発明の名称】分類モデル学習装置および分類モデル学習方法

(57)【要約】
【課題】質の悪い教師データが含まれていても精度の良い分類モデルを構築する。
【解決手段】ラベル付けの信頼度が所定の基準を満たすエキスパートデータおよびラベル付けの信頼度が不明な非エキスパートデータの各々が対応する座標を取得して非エキスパートデータからエキスパートデータまでの距離を各々算出し、所定の規則に当てはめて近傍距離を定義する。次に、選択した非エキスパートデータから近傍距離の範囲内にあるエキスパートデータを探索して同ラベル確率を算出し、付されたラベルが近傍距離の範囲内にあるエキスパートデータのラベルに一致する確率に基づく信頼度関数に当てはめて非エキスパートデータの信頼度を決定し、付加する。そして、エキスパートデータおよび信頼度が付加された非エキスパートデータに基づいて所望のデータにラベル付けを行う分類モデルを学習する。
【選択図】図１



【特許請求の範囲】
【請求項１】
機械学習におけるラベル付けの信頼度が所定の基準を満たす教師データをエキスパートデータとして格納するエキスパートデータ格納部と、
前記ラベル付けの信頼度が不明な教師データを非エキスパートデータとして格納する非エキスパートデータ格納部と、
前記エキスパートデータ格納部および前記非エキスパートデータ格納部に接続され、前記エキスパートデータおよび前記非エキスパートデータの各々が対応する座標を取得して前記非エキスパートデータから前記エキスパートデータまでの距離を各々算出すると共に、この算出された距離を所定の規則に当てはめて近傍距離を定義する近傍距離定義部と、
前記非エキスパートデータに付された前記ラベルが前記近傍距離の範囲内にある前記エキスパートデータに付された前記ラベルに一致する確率に基づく信頼度関数を格納する信頼度関数格納部と、
前記近傍距離定義部、前記信頼度関数格納部、前記エキスパートデータ格納部、および前記非エキスパートデータ格納部に接続され、選択した前記非エキスパートデータから前記近傍距離の範囲内にある前記エキスパートデータを探索して前記確率を算出すると共に、この算出された確率を前記信頼度関数に当てはめて前記非エキスパートデータにおける前記ラベル付けの信頼度を決定する信頼度決定部と、
前記エキスパートデータ格納部および前記信頼度決定部に接続され、前記エキスパートデータおよび前記信頼度が付加された非エキスパートデータに基づいて所望の分類対象データに前記ラベル付けを行う分類モデルを学習する分類モデル学習部と、
を有することを特徴とする分類モデル学習装置。
【請求項２】
前記信頼度関数は、前記非エキスパートデータに付された前記ラベルが前記近傍距離の範囲内にある前記エキスパートデータに付された前記ラベルに一致する確率と前記ラベル付けの信頼度との関係を定義することを特徴とする請求項１記載の分類モデル学習装置。
【請求項３】
前記近傍距離定義部は、前記非エキスパートデータから前記エキスパートデータまでの前記座標間の距離を各々算出して前記非エキスパートデータ毎に順位付けを行い、所望の順位についての距離を前記非エキスパートデータの各々から集計してその平均値を算出し、この平均値を前記近傍距離として定義することを特徴とする請求項１または請求項２記載の分類モデル学習装置。
【請求項４】
前記信頼度関数は、分類問題における所望の評価基準に応じて予め作成されていることを特徴とする請求項１乃至請求項３のいずれか一項記載の分類モデル学習装置。
【請求項５】
前記分類モデル学習部が、アンサンブル学習におけるデータ重みに対して前記信頼度決定部で付加された信頼度を反映させることにより前記分類モデルを学習することを特徴とする請求項１乃至請求項４のいずれか一項記載の分類モデル学習装置。
【請求項６】
機械学習におけるラベル付けの信頼度が所定の基準を満たしている教師データをエキスパートデータ、前記ラベル付けの信頼度が不明の教師データを非エキスパートデータとして格納するコンピュータが行う分類モデル学習方法であって、
前記エキスパートデータおよび前記非エキスパートデータの各々が対応する座標を取得して前記非エキスパートデータから前記エキスパートデータまでの距離を各々算出すると共に、この算出された距離を所定の規則に当てはめて近傍距離を定義する近傍距離定義ステップと、
前記格納された非エキスパートデータから前記信頼度の付加対象となる非エキスパートデータを選択する選択ステップと、
前記選択された非エキスパートデータから前記近傍距離の範囲内にある前記エキスパー

トデータを探索して前記非エキスパートデータに付された前記ラベルが前記エキスパートデータに付された前記ラベルに一致する確率を算出すると共に、この算出された確率を予め定義された信頼度関数に当てはめて前記非エキスパートデータの前記ラベル付けの信頼度を決定する信頼度決定ステップと、
前記決定された信頼度が付加された非エキスパートデータおよび前記エキスパートデータに基づいて所望のデータに前記ラベル付けを行う分類モデルを学習する分類モデル学習ステップと、
を有することを特徴とする分類モデル学習方法。
【請求項７】
前記信頼度関数は、前記非エキスパートデータに付された前記ラベルが前記近傍距離の範囲内にある前記エキスパートデータに付された前記ラベルに一致する確率と前記ラベル付けの信頼度との関係を定義することを特徴とする請求項６記載の分類モデル学習方法。
【請求項８】
前記近傍距離定義ステップにおいて、前記非エキスパートデータから前記エキスパートデータまでの前記座標間の距離を各々算出して前記非エキスパートデータ毎に順位付けを行い、所望の順位についての距離を前記非エキスパートデータの各々から集計してその平均値を算出し、この平均値を前記近傍距離として定義することを特徴とする請求項６または請求項７記載の分類モデル学習方法。
【請求項９】
前記信頼度関数は、分類問題における所望の評価基準に応じて予め作成されていることを特徴とする請求項６乃至請求項８のいずれか一項記載の分類モデル学習方法。
【請求項１０】
前記分類モデル学習ステップにおいて、アンサンブル学習におけるデータ重みに対して前記信頼度決定ステップにおいて付加された信頼度を反映させることにより前記分類モデルを学習することを特徴とする請求項６乃至請求項９のいずれか一項記載の分類モデル学習方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、機械学習において分類対象データにラベル付けを行う分類モデル学習装置および分類モデル学習方法に関する。
【背景技術】
【０００２】
データマイニングにおいて重要な分野の一つに機械学習が挙げられる。機械学習は分類問題に使われることが多く、分類問題において計算機は人間のつけた評価を学習することで分類モデルを構築する。この機械学習の応用は画像認識、文字認識、テキスト分類など広い分野で多くの成果を上げている。このような学習は一般的に教師あり学習と呼ばれる。
【０００３】
教師あり学習には計算機に正しい判断を教える「教師データ」、つまり人間の手によって「ラベル」が付けられたデータが必要である。計算機は教師データをもとにどのような分類をすれば良いかを学習し、新しいデータに対して自動的に判断を下せるようになる。現代ではＩＴ環境の発展により、大量で詳細なデータが機械学習に利用可能となっており、これらを教師データとして用いればより正確な分類モデルの構築に繋がると期待されている。
【０００４】
しかし、ここで「大量のデータ」から「大量の教師データ」を得る際のラベル付けが問題になっている。すなわち、得られたデータを教師データとして利用するためには、データに対して人間が判断したラベルを付与することが必要であるが、正確なラベル付けには、データが取られたドメインに対しての知識や経験などに基づく正確な判断が不可欠であ

る。
【０００５】
理想的にはこれらの条件を満たす対象分野のエキスパートがラベル付けを行うことが望ましいが、全てのデータのラベル付けをエキスパートに依頼することは非常にコストが高くなってしまう。しかし、現実的にはコストに制限があるため、大量の教師データが必要である場合には、非エキスパートがラベル付けを行った教師データを用いることになる。エキスパートによる高コストの教師データは少量になりがちなのに対し、非エキスパートによる低コストの教師データは比較的大量に獲得できるためである。その一方、非エキスパートによる教師データには判断のミスや知識の不正確さから、比較的多くの誤ったラベルが含まれてしまうことが考えられる。
【０００６】
また、一般の機械学習においては、教師データの取得に関する情報は用いられず、エキスパートによるラベル付けのような「良質の教師データ」と非エキスパートによるラベル付けのような「ノイズを含む教師データ」が混在する状況においても、全てのデータを同列に扱い、等しく学習に使用する。
【０００７】
したがって、エキスパートによる少量の教師データと非エキスパートによる大量の教師データを従来どおり同列とみなして学習に使用した場合、非エキスパートデータに含まれるノイズが学習に大きく影響し、精度の良い分類モデルが構築できないケースが考えられる。
【０００８】
一方で、分類モデルを学習する際に、一部の教師データを選択的に使用して学習を行うことや、一部の教師データに重みを置いて学習を行うことが一般的に行われている。アンサンブル学習の代表的手法の１つであるＡｄａＢｏｏｓｔもその一つである。ＡｄａＢｏｏｓｔは、学習データに対して重みを与えて学習器を生成し、その際に誤った分類をしたデータに対して重みを増して再度学習器を生成することを繰り返して複数の弱学習器を得て、それらの弱学習器の重みつき投票により分類を行う手法である（例えば特許文献１、非特許文献１参照）。
【特許文献１】特開２００２－１３３３８９号公報
【非特許文献１】Y. Freund and R.E. Shapire, “Experiments with a new boosting algorithm”, Proc. of the 13th. Int. Conf. on Machine Learning, 1996, 148-156
【発明の開示】
【発明が解決しようとする課題】
【０００９】
しかしながら、従来技術は、あくまで所定のアルゴリズムに即した形で教師データに対してデータ重みをつけるものであり、教師データの精度の差異という学習過程を開始する前の知識・情報を含んだものではなく、例えばエキスパートによる少量の教師データ（以下、「エキスパートデータ」という。）と非エキスパートによる大量の教師データ（以下、「非エキスパートデータ」という。）のような、質の異なる教師データを従来どおり同列として学習に使用した場合、質の劣る教師データに含まれるノイズが学習に大きく影響し、精度の良い分類モデルが構築できないという問題があった。
【００１０】
このような問題に対して、本出願人は特許出願２００７－２７８８９３においてエキスパートによる少量の教師データを利用することで精度の良い分類モデルの学習を行う手法を提案している。この手法は、エキスパートによる教師データを基にして非エキスパートによる教師データのラベルに信頼度を付加し、分類モデルの学習にその信頼度を反映することで分類モデルを学習するものである。この信頼度は、エキスパートデータおよび非エキスパートデータの各々を所定の規則に基づいて対応付けた座標の間の距離（例えばユークリッド距離やコサイン距離）に応じて求められている。
【００１１】

対象の非エキスパートデータから距離の近いＮ個のエキスパートデータを探索し、もしラベルが同じであればそのエキスパートデータから信頼度を得る。この信頼度は例えば距離に反比例する形で与えられ、非エキスパートデータの近くのエキスパートデータが同じラベルであれば、その非エキスパートデータは高い信頼度を得られるようになっている。これは、信頼できるデータが近くにあるほど信頼度は高いという直感的な信頼度付けを表していると言える。

【００１２】

しかしながら、上記の信頼度付け方法は、エキスパートデータには誤ラベルが含まれていないことを暗に仮定している。エキスパートデータに全て適切なラベルが与えられているならば、それらを参照して与えられた非エキスパートデータの信頼度も適切な値になることが期待できる。その反面、エキスパートデータに誤ラベルが含まれている場合には、このような信頼度の付加は必ずしも適切とは言えない。図１０および図１１は、エキスパートデータのラベル付けと非エキスパートデータのラベル付けに対する信頼度の関係を説明する図である。

【００１３】

図１０では、ある非エキスパートデータx_1の非常に近傍にエキスパートデータX_1が存在している。このX_1は非常に近傍にあるため、X_1とx_1のラベルが同じであればx_1の信頼度は高く、異なれば低くなる。ここでX_1、x_1に本来付与されるべきラベルはL_1であるとする。エキスパートデータX_1に、正確なラベルL_1が付与されているとすると、非エキスパートデータx_1にL_1が付与されている場合には信頼度は高く、異なったラベルL_2が付与されている場合には信頼度は低くなる。これは、適切な信頼度であるといえる。

【００１４】

図１１では、エキスパートデータX_1に誤ラベルL_2が付与されている場合を考える。このとき、非エキスパートデータx_1に本来付与されるべきラベルL_1が付与されていたときは信頼度が低く、反対に付与されるべきでないラベルL_2が付与されていたときに信頼度が高くなってしまう。これは明らかに適切な信頼度とは反対の傾向である。

【００１５】

すなわち、エキスパートデータ中に誤ラベルが含まれている場合、非エキスパートデータに適切でない信頼度が付加され、その信頼度を反映して生成される分類モデルの性能が劣化してしまう。現実にはエキスパートデータ中にも少量の誤ラベルが含まれると考えられるため、エキスパートデータ中の誤ラベルに頑健な信頼度付加が必要である。

【００１６】

そこで、本発明は、従来技術の問題に鑑み、質の悪い教師データが含まれている状況であっても精度の良い分類モデルの構築が可能な分類モデル学習装置および分類モデル学習方法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】

【００１７】

本発明に係る分類モデル学習装置は、機械学習におけるラベル付けの信頼度が所定の基準を満たす教師データをエキスパートデータとして格納するエキスパートデータ格納部と、前記ラベル付けの信頼度が不明な教師データを非エキスパートデータとして格納する非エキスパートデータ格納部と、前記エキスパートデータ格納部および前記非エキスパートデータ格納部に接続され、前記エキスパートデータおよび前記非エキスパートデータの各々が対応する座標を取得して前記非エキスパートデータから前記エキスパートデータまでの距離を各々算出すると共に、この算出された距離を所定の規則に当てはめて近傍距離を定義する近傍距離定義部と、前記非エキスパートデータに付された前記ラベルが前記近傍距離の範囲内にある前記エキスパートデータに付された前記ラベルに一致する確率に基づく信頼度関数を格納する信頼度関数格納部と、前記近傍距離定義部、前記信頼度関数格納部、前記エキスパートデータ格納部、および前記非エキスパートデータ格納部に接続され、選択した前記非エキスパートデータから前記近傍距離の範囲内にある前記エキスパートデータを探索して前記確率を算出すると共に、この算出された確率を前記信頼度関数に当

てはめて前記非エキスパートデータにおける前記ラベル付けの信頼度を決定する信頼度決定部と、前記エキスパートデータ格納部および前記信頼度決定部に接続され、前記エキスパートデータおよび前記信頼度が付加された非エキスパートデータに基づいて所望の分類対象データに前記ラベル付けを行う分類モデルを学習する分類モデル学習部と、を有することを特徴とする。

【００１８】

本発明に係る分類モデル学習方法は、機械学習におけるラベル付けの信頼度が所定の基準を満たしている教師データをエキスパートデータ、前記ラベル付けの信頼度が不明の教師データを非エキスパートデータとして格納するコンピュータが行う分類モデル学習方法であって、前記エキスパートデータおよび前記非エキスパートデータの各々が対応する座標を取得して前記非エキスパートデータから前記エキスパートデータまでの距離を各々算出すると共に、この算出された距離を所定の規則に当てはめて近傍距離を定義する近傍距離定義ステップと、前記格納された非エキスパートデータから前記信頼度の付加対象となる非エキスパートデータを選択する選択ステップと、前記選択された非エキスパートデータから前記近傍距離の範囲内にある前記エキスパートデータを探索して前記非エキスパートデータに付された前記ラベルが前記エキスパートデータに付された前記ラベルに一致する確率を算出すると共に、この算出された確率を予め定義された信頼度関数に当てはめて前記非エキスパートデータの前記ラベル付けの信頼度を決定する信頼度決定ステップと、前記決定された信頼度が付加された非エキスパートデータおよび前記エキスパートデータに基づいて所望のデータに前記ラベル付けを行う分類モデルを学習する分類モデル学習ステップと、を有することを特徴とする。

【発明の効果】

【００１９】

本発明によれば、質の悪い教師データが含まれている状況であっても精度の良い分類モデルの構築が可能な分類モデル学習装置および分類モデル学習方法が提供される。

【発明を実施するための最良の形態】

【００２０】

以下、本発明の実施形態について図面を用いて説明する。図１は、本発明の一実施形態に係る分類モデル学習装置１の全体構成例を示すブロック図である。同図に示されるように、本実施形態に係る分類モデル学習装置１は、エキスパートデータ格納部１１、非エキスパートデータ格納部１２、近傍距離定義部１３、信頼度関数格納部１４、信頼度決定部１５、信頼度付き非エキスパートデータ格納部１６、分類モデル学習部１７、分類対象データ格納部１８、予測部１９、および結果表示部２０から構成されている。

【００２１】

エキスパートデータ格納部１１は、エキスパートデータを格納する記憶装置である。「エキスパートデータ」とは、知識、経験の豊富な専門家が機械学習においてデータを分類するためのラベル付けを行われており、ラベル付けの精度（信頼性）が高い教師データを示すものとする。

【００２２】

非エキスパートデータ格納部１２は、非エキスパートデータを格納する記憶装置である。「非エキスパートデータ」とは、ラベル付けは行われているが、その精度（信頼性）が不明確な教師データを示すものとする。

【００２３】

近傍距離定義部１３は、非エキスパートデータからエキスパートデータまでの座標間距離を各々算出し、この座標間距離に基づいてデータ間の類似度が基準値以上の範囲を表す近傍距離を定義するプログラムである。ここでは、算出された座標間距離の中から所定の規則に基づいて複数の距離を選択し、これらの距離の平均値から近傍距離を算出するが、算出方法はこれに限られない。

【００２４】

図２は、エキスパートデータおよび非エキスパートデータを２次元で具体的に説明する

図である。同図において、丸印はエキスパートデータ、四角印は非エキスパートデータを表し、各印の色はラベルを表している。これらの座標は各データを所定の規則に基づいて変換することで得られる。例えば、電子メールの分類においては、多数の迷惑メールを解析することによって特徴語リストを予め作成しておき、この特徴語リストと受信メール本文内の単語を比較することで座標化を行う。具体的には、特徴語リストに含まれるＮ個の単語との比較結果を受信メール内に含まれる場合を１、含まれない場合を０として表すことにより、受信メールのデータをＮ次元の座標（例えば（１,０，１，…，１））に変換できる。ここでは、説明のためにメールデータを座標化したＮ次元のデータを擬似的に２次元で表しているものとする。すなわち、受信メール本文の内容が近似する場合には、座標が近似するので迷惑メールか否かのラベル付け等に用いることができる。

【００２５】

また、図２においては、近傍距離定義部１３が非エキスパートデータを選択し、この選択された非エキスパートデータから各エキスパートデータまでの距離を順次求めることが示されている。例えば、近傍距離を“非エキスパートデータから４番目に近いエキスパートデータまでの距離の平均値”とする規則が予め定められている場合には、距離r_4を非エキスパートデータ毎に求め、その平均値を算出する。

【００２６】

信頼度関数格納部１４は、分類問題に適した信頼度関数を格納する記憶装置である。この信頼度関数は、非エキスパートデータから近傍距離内にあるエキスパートデータの同ラベル確率に基づいて信頼度を定義する関数であり、この関数は種々の分類問題に対応させて予め複数作成しておくと好適である。具体的な定義方法は後述する。

【００２７】

信頼度決定部１５は、近傍距離定義部１３により定義された近傍距離に基づいて非エキスパートデータの近傍にあるエキスパートデータを探索すると共に非エキスパートデータとの同ラベル確率を算出し、この同ラベル確率を信頼度関数格納部１４から取得される信頼度関数に当てはめて非エキスパートデータの信頼度を決定するプログラムである。尚、複数の信頼度関数の内、どの関数を用いるか選択する方法としては、モデル作成時にユーザが入力装置（図示省略する）から入力した情報に基づいて選択する方法や使用する関数を予め設定しておく方法などが挙げられる。

【００２８】

信頼度付き非エキスパートデータ格納部１６は、信頼度決定部１５における処理よって信頼度が付与された非エキスパートデータ（以下、「信頼度付き非エキスパートデータ」という。）を格納する記憶装置である。

【００２９】

分類モデル学習部１７は、エキスパートデータと信頼度付き非エキスパートデータを用いて分類モデルを学習するプログラムである。

【００３０】

分類対象データ格納部１８は、新たに分類の対象となるデータ、すなわち、ラベルが付与されていないデータ（以下、「分類対象データ」という。）を格納する記憶装置である。

【００３１】

予測部１９は、分類モデル学習部１７で得られた分類モデルを用いて分類対象データ格納部１８に格納されている分類対象データにラベル付けを行うプログラムである。尚、ＡｄａＢｏｏｓｔを用いた場合、予測部１９での分類手法は、一般的なＡｄａＢｏｏｓｔにおける手法と同様であるので説明は省略する。

【００３２】

結果表示部２０は、予測部１９における予測結果を表示するディスプレイなどの表示装置である。

【００３３】

以下、分類モデル学習装置１における動作を図面に基づいて説明する。尚、本実施形態

においては、エキスパートデータおよび非エキスパートデータを２次元のデータとして具体的に説明する。図３は、近傍距離定義部１３における処理の具体例を示すフローチャートである。

【００３４】

Ｓ３０１においては、未だ選択されていない非エキスパートデータが存在するか否かを判断する。ここで、全ての非エキスパートデータが選択済みであればＳ３０５へ進む。これに対し、選択されていない非エキスパートデータが存在する場合にはＳ３０２へ進む。

【００３５】

Ｓ３０２においては、非エキスパートデータ格納部１２から未だ選択されていない非エキスパートデータを一つ選択する。

Ｓ３０３においては、選択された非エキスパートデータから全てのエキスパートデータへの距離を各々算出する。

【００３６】

Ｓ３０４においては、選択された非エキスパートデータからｋ番目に近いエキスパートデータまでの距離をバッファ領域（図示省略する）に保持する。尚、最適な整数ｋは問題によって異なるが、ここでは整数ｋをユーザが予め設定した値とする。例えば、Ｓ３０３で算出された距離の分布を解析し、各非エキスパートデータからの距離が所定の範囲内にあるように整数ｋを設定することができる。また、信頼度の付加にあたって複数の近傍エキスパートデータを考慮したい場合などには整数ｋを大きくすれば良い。

Ｓ３０５においては、保持していた全ての距離の平均をとり、その値を近傍距離として信頼度決定部１５へ出力し、処理を終了する。

【００３７】

以上の処理により、ｋ番目に近いエキスパートデータまでの平均距離が求められる。問題に適した整数ｋを設定すれば、この距離は近傍を定義する典型的な値をとると考えることができる。

【００３８】

図４は、信頼度決定部１５における処理の具体例を示すフローチャートである。Ｓ４０１においては、選択する非エキスパートデータが存在するか否かを判断する。ここで、全ての非エキスパートデータに信頼度が付与されており選択する非エキスパートデータがなければ処理を終了する。これに対し、信頼度が付与されていない非エキスパートデータが存在する場合にはＳ４０２へ進む。

【００３９】

Ｓ４０２においては、非エキスパートデータ格納部１２から未だ信頼度が付与されていない非エキスパートデータを１つ選択する。ここでは、下記の式（１）で表されるｊ番目の非エキスパートデータが選択されているとする。尚、ｘは座標、ｙはラベルを表すものとする。

【数１】

$$[\vec{x}_j, y_j] = [(1.0,2.0),1] \quad \cdots (1)$$

【００４０】

Ｓ４０３においては、選択された非エキスパートデータの近傍に含まれるエキスパートデータをエキスパートデータ格納部１１から探索して保持する。この例では、「近傍」とは近傍距離定義部１３において定義された近傍距離ｒを用いて、上記式（１）で表される非エキスパートデータを中心とした半径ｒの円の中の領域を指すものとする。したがって、近傍距離ｒが０．５ときは、下記の式（２）のエキスパートデータX_{j1}は近傍に含まれるが、式（３）のエキスパートデータX_{j2}は近傍には含まれない。

【数２】

$$[\vec{X}_{j_1}, Y_{j_1}] = [(1.1,2.3),1] \quad \cdots（２）$$

$$[\vec{X}_{j_2}, Y_{j_2}] = [(1.5,2.3),1] \quad \cdots（３）$$

【００４１】

Ｓ４０４においては、探索されたN個のエキスパートデータから同ラベル確率を算出する。この例では、対象の非エキスパートデータｘ$_j$と同ラベルの近傍エキスパートデータの数をＫ個とし、同ラベル確率Ｐ$_j$を下記の式（４）で定義する。

Ｐ$_j$＝Ｋ／Ｎ　　・・・（４）

【００４２】

Ｓ４０５においては、式（４）を入力とする信頼度関数を用いて非エキスパートデータのラベルの信頼度を算出する。信頼度関数は分類問題によって適した形が考えられる。図５乃至図７は、分類問題の評価基準に応じた信頼度関数の具体例を説明する図である。この信頼度関数の性質の直感的な理解のために、対象となっている式（１）が表す非エキスパートデータの近傍にエキスパートデータが１０例含まれており、さらにノイズのため本来は９例が同ラベルであるところ８例が同ラベルとなっている状況を考える。

【００４３】

この状況下で、例えば、非エキスパートデータのラベル付けが近傍エキスパートデータの８割以上と一致するならば、そのラベル付けに高信頼度を与えたい場合には、下記の式（５）のような信頼度関数を用いると好適である。尚、ａは関数の形を決定するパラメーターである。

【数３】

$$c_j = C(P_j) = \frac{\exp(a(P_j - 1/2))}{\exp(a(P_j - 1/2)) + \exp(-a(P_j - 1/2))} \Big/ \frac{\exp(1/2a)}{\exp(1/2a) + \exp(-1/2a)} \quad \cdots（５）$$

【００４４】

図５は、式（５）の信頼度関数を説明する図である。ここでは、横軸を同ラベル確率（Ｐ$_j$）、縦軸を信頼度（ｃ$_j$）とし、ａ＝２．０の場合に式（５）によって求められる点を結んだ曲線で示されている。同ラベル数が９例から８例に変化するときノイズによる信頼度ｃ$_j$の変化はＣ（９／１０）　０．９８からＣ（８／１０）　０．９６となり、信頼度ｃ$_j$への影響は小さい。すなわち、近傍の１０例中の同ラベルが９例、８例のいずれの場合であっても、その非エキスパートデータのラベルの信頼度は高く維持されるという結果が得られる設定になっており、直感的にも妥当な信頼度関数であると言える。

【００４５】

また、誤ラベルの混入に対して厳しい設定としたい場合には、下記の式（６）のような信頼度関数を用いると好適である。

【数４】

$$c_j = C(P_j) = \frac{\exp(aP_j) - 1.0}{\exp(a) - 1.0} \quad \cdots（６）$$

【００４６】

図６は、式（６）の信頼度関数を示す図である。ここでは、横軸を同ラベル確率（Ｐ$_j$）、縦軸を信頼度（ｃ$_j$）とし、ａ＝５．０の場合に式（６）によって求められる点を結んだ曲線で示されている。この関数を用いる場合には、一つでも誤ラベルがあると信頼度が大

幅に下がる。例えば、医療などの高い信頼度が要求される分野において特に有用である。

【００４７】

更に、誤ラベルの混入に対して寛容な設定としたい場合には、下記の式（７）のような信頼度関数を用いると好適である。

【数５】

$$c_j = C(P_j) = \frac{\log(aP_j + 1.0)}{\log(a - 1.0)} \quad \cdots (7)$$

【００４８】

図７は、式（７）の信頼度関数を示す図である。ここでは、横軸を同ラベル確率（P_j）、縦軸を信頼度（c_j）とし、a＝１０．０の場合に式（７）によって求められる点を結んだ曲線で示されている。この関数を用いる場合には、誤ラベルが多く含まれていても信頼度が大幅に下がることはなく、誤ラベルの増加に応じて信頼度が緩やかに低下する。

【００４９】

Ｓ４０６において、Ｓ４０５で得られた信頼度c_jを対象の非エキスパートデータに付加し、下記の式（８）のような形で信頼度付き非エキスパートデータ格納部１６に格納する。

【数６】

$$[(\vec{x}_1, y_1, c_1), \cdots, (\vec{x}_j, y_j, c_j), \cdots, (\vec{x}_n, y_n, c_n)] \quad \cdots (8)$$

【００５０】

前述の２次元データの例（式（１）の非エキスパートデータ）であれば、下記の式（９）の形で信頼度付き非エキスパートデータ格納部１６に格納される。

【数７】

$$[\vec{x}_j, y_j, c_j] = [(1.0, 2.0), 1, 0.96] \quad \cdots (9)$$

【００５１】

尚、エキスパートデータの信頼度は常に１としているので、エキスパートデータは擬似的に下記の式（１０）の形でエキスパートデータ格納部１１に格納されているとみなすことができる。

【数８】

$$[(\vec{X}_1, Y_1, 1), \cdots, (\vec{X}_j, Y_j, 1), \cdots, (\vec{X}_N, Y_N, 1)] \quad \cdots (10)$$

【００５２】

このように、近傍距離内における同ラベル確率を考慮した信頼度関数を用いることで、最近傍にあるエキスパートデータに誤ラベルが与えられていたとしても、他のラベルが正確であれば非エキスパートデータに適切な信頼度を付加することが可能になる。このような信頼度付けはデータ間の距離の長短のみに基づく信頼度付けよりもエキスパートデータの誤ラベルに対して頑健であると言える。

【００５３】

図８は、分類モデル学習部１７における処理の具体例を示すフローチャートである。学習器については信頼度を反映する形のものであれば、どのような学習器でも機能すると考えられるが、ここではデータ重みに対する信頼度の組み込み易さを考慮してＡｄａＢｏｏｓｔの手法に即した形で処理を行うものとする。尚、Ｂａｇｇｉｎｇなどの他の手法を用いても良い。

【００５４】

Ｓ８０１においては、読み込まれた信頼度付き非エキスパートデータとエキスパートデータに、ＡｄａＢｏｏｓｔの手法に即して均等のデータ重みｗ$_j$を付ける。本発明では、ＡｄａＢｏｏｓｔにおける従来のデータ重みｗ$_j$に加え、信頼度決定部１５で得られた信頼度ｃ$_j$が教師データに付加されているため、ここでは読み込まれたｎ個の非エキスパートデータは下記の式（１１）、Ｎ個のエキスパートデータはそれぞれ下記の式（１２）の形で処理されるものとする。

【数９】

$$[(\vec{x}_1, y_1, c_1, w_1), \cdots, (\vec{x}_j, y_j, c_j, w_j), \cdots, (\vec{x}_n, y_n, c_n, w_n)] \quad \cdots (11)$$

$$[(\vec{X}_1, Y_1, 1, w_{n+1}), \cdots, (\vec{X}_j, Y_j, 1, w_{n+j}), \cdots, (\vec{X}_N, Y_N, 1, w_{n+N})] \quad \cdots (12)$$

【００５５】

Ｓ８０２においては、非エキスパートデータに付与された信頼度ｃ$_j$をデータ重みに反映させる。ここでは、ＡｄａＢｏｏｓｔにおけるデータ重みｗ$_j$に対して信頼度ｃ$_j$を反映させたデータ重みｗ’$_j$を下記の式（１３）により設定する。

ｗ’$_j$＝ｃ$_j$ｗ$_j$　・・・（１３）

【００５６】

このように設定することにより、データ重みｗ$_j$が大きく学習に大きな影響を及ぼすと考えられる非エキスパートデータに関しても、その非エキスパートデータの信頼度ｃ$_j$が低ければデータ重みｗ’$_j$の値は小さくなり、非エキスパートデータに含まれる信頼度ｃ$_j$の低い教師データの影響を自然な形で小さくすることができる。

【００５７】

Ｓ８０３においては、Ｓ８０２で得られたデータ重みｗ’$_j$を用いて弱学習器を生成する。ＡｄａＢｏｏｓｔに用いられる弱学習器には決定木など様々なものが考えられる。

Ｓ８０４においては、ＡｄａＢｏｏｓｔのアルゴリズムに従い、データ重みと弱学習器の性能に依るコスト関数の更新を行う。

【００５８】

Ｓ８０５においては、終了条件を満たしているか否かを判定する。ここで、終了条件を満たすと判定された場合にはＳ８０６へ進む。これに対し、終了条件を満たさないと判定された場合はＳ８０２に戻る。尚、一般的なＡｄａＢｏｏｓｔの手法における終了条件は、弱学習器の数が所定数を満たすことである。例えばユーザが弱学習器を１００個作るという設定にすれば、Ｓ８０２～Ｓ８０５を１００回繰り返すことが終了条件である。

Ｓ８０６においては、生成された弱学習器を組合せることにより精度の高い分類モデルである強学習器を生成し、処理を終了する。

【００５９】

このように、教師データの精度の差異という学習過程を開始する前の知識を利用して非エキスパートデータに信頼度を付与し、分類モデルの学習に組み込むことで、エキスパートデータが少ない場合であっても精度の良い分類モデルを得ることができる。

【００６０】

図９は、予測部１９における処理の具体例を示すフローチャートである。Ｓ９０１においては、分類対象データ格納部１８における分類対象データの有無を判定する。ここで、分類対象データが有ると判定された場合には、Ｓ９０２へ進む。これに対し、分類対象データが無いと判定された場合には、処理を終了する。

【００６１】

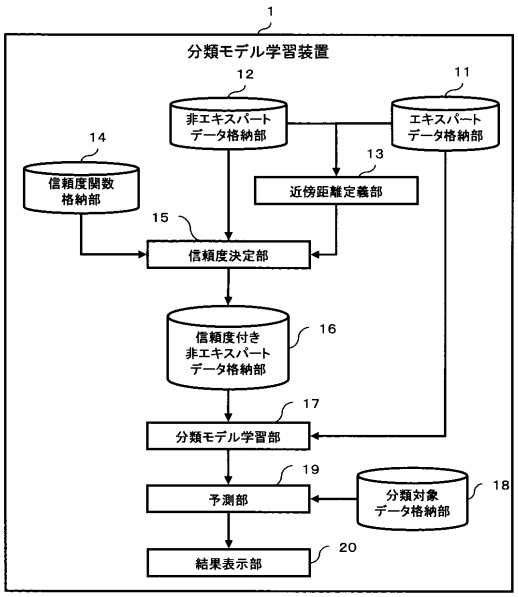
Ｓ９０２においては、分類対象データ格納部１８から分類対象データを一つ選択する。

Ｓ９０３においては、選択した分類対象データを分類モデルに当てはめてラベル付けを行い、Ｓ９０１へ戻る。Ｓ９０１～Ｓ９０３までの処理は全ての分類対象データに対してラベル付けが完了するまで繰返し行われる。

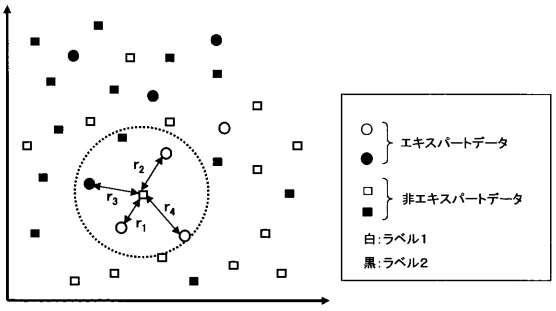
【００６２】
上記のように構成することで、高信頼度とされる教師データ（エキスパートデータ）の中にノイズが含まれる場合においても、同ラベル確率を入力とする信頼度関数と、各教師データの精度という事前知識を利用して非エキスパートデータに信頼度を付与し、分類モデルの学習に組み込むことで、精度の良い分類モデルを得ることができる。
【００６３】
尚、本発明は上記実施形態そのままに限定されるものではなく、実施段階ではその要旨を逸脱しない範囲で構成要素を変形して具体化できる。また、上記実施形態に開示されている複数の構成要素の適宜な組み合わせにより、種々の発明を形成できる。例えば、実施形態に示される全構成要素から幾つかの構成要素を削除してもよい。更に、異なる実施形態にわたる構成要素を適宜組み合わせてもよい。
【００６４】
また、上記実施形態においてはメールデータのようなテキストデータを例として説明したが、対象データの種類はこれに限定されない。すなわち、画像データや音声データなどのデータにおいても所定の規則に基づいて座標化することで、分類モデルを作成可能である。例えば、２次元のレントゲン画像データはＭ行Ｎ列に分割し、これを１行Ｍ×Ｎ列のデータに変換すればＭ×Ｎ次元の座標が得られる。この場合、画像データにおける色彩区分（例えば１６段階のグレースケールなど）を行列の成分とすると好適である。そして、経験豊富な医師によって病変の有無が判定（ラベル付け）されたレントゲン画像データをエキスパートデータ、経験の浅い医師によって判定されたデータを非エキスパートデータとし、上記実施形態と同様に信頼度付けを行うことで精度の高い分類モデルを作成できる。
【図面の簡単な説明】
【００６５】
【図１】本発明の一実施形態に係る分類モデル学習装置１の全体構成例を示すブロック図。
【図２】エキスパートデータおよび非エキスパートデータを２次元で具体的に説明する図。
【図３】近傍距離定義部１３における処理の具体例を示すフローチャート。
【図４】信頼度決定部１５における処理の具体例を示すフローチャート。
【図５】分類問題の評価基準に応じた信頼度関数の具体例を説明する図。
【図６】分類問題の評価基準に応じた信頼度関数の具体例を説明する図。
【図７】分類問題の評価基準に応じた信頼度関数の具体例を説明する図。
【図８】分類モデル学習部１７における処理の具体例を示すフローチャート。
【図９】予測部１９における処理の具体例を示すフローチャート。
【図１０】エキスパートデータのラベル付けと非エキスパートデータのラベル付けに対する信頼度の関係を説明する図。
【図１１】エキスパートデータのラベル付けと非エキスパートデータのラベル付けに対する信頼度の関係を説明する図。
【符号の説明】
【００６６】
１…分類モデル学習装置、
１１…エキスパートデータ格納部、
１２…非エキスパートデータ格納部、
１３…近傍距離定義部、
１４…信頼度関数格納部、
１５…信頼度決定部、
１６…信頼度付き非エキスパートデータ格納部、
１７…分類モデル学習部、
１８…分類対象データ格納部、

１９…予測部、
２０…結果表示部。

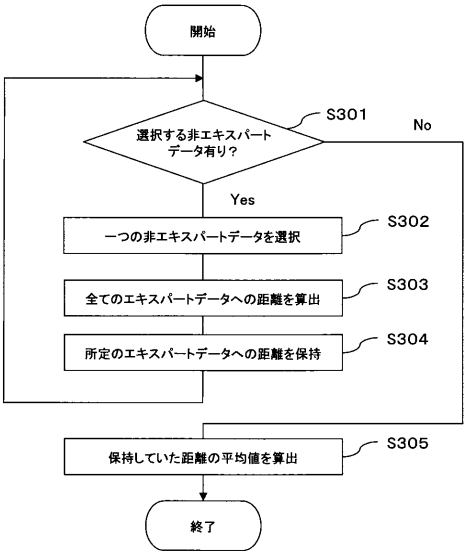
【図１】



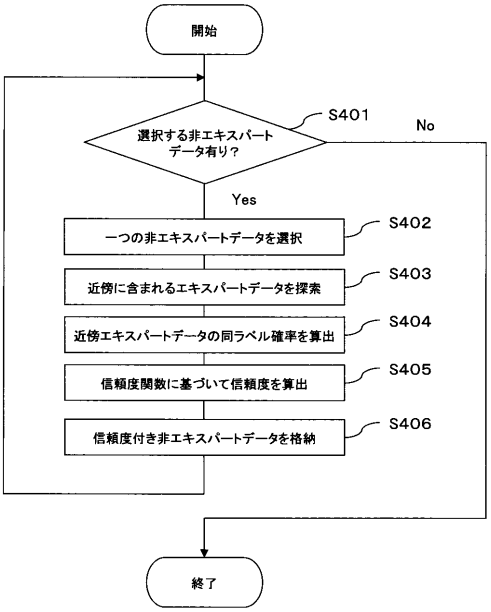
【図２】



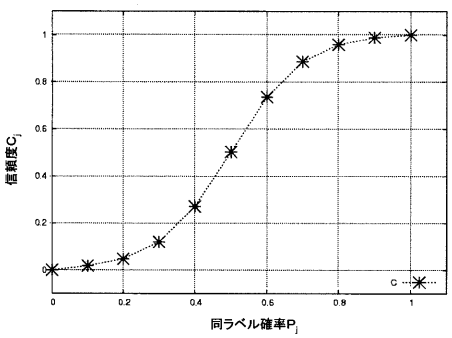
【図３】


開始
選択する非エキスパートデータ有り？
S301
No
Yes
一つの非エキスパートデータを選択
S302
全てのエキスパートデータへの距離を算出
S303
所定のエキスパートデータへの距離を保持
S304
保持していた距離の平均値を算出
S305
終了

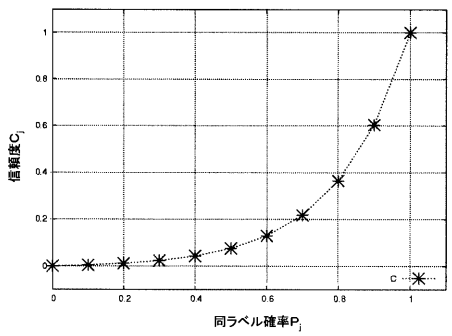
【図４】


開始
選択する非エキスパートデータ有り？
S401
No
Yes
一つの非エキスパートデータを選択
S402
近傍に含まれるエキスパートデータを探索
S403
近傍エキスパートデータの同ラベル確率を算出
S404
信頼度関数に基づいて信頼度を算出
S405
信頼度付き非エキスパートデータを格納
S406
終了

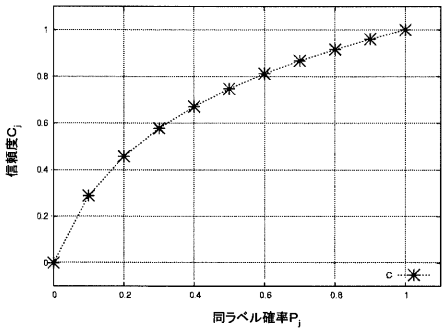
【図５】


信頼度C_j
同ラベル確率P_j
C

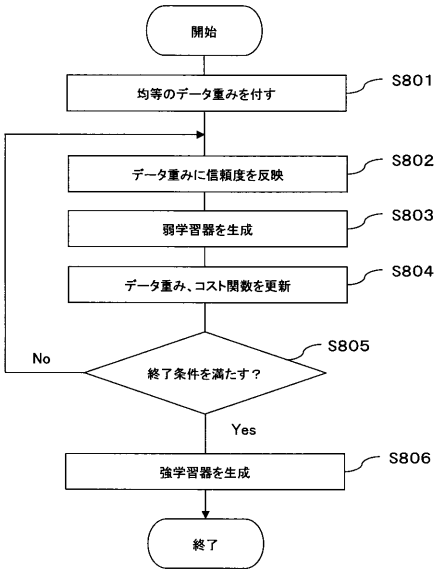
【図６】


信頼度C_j
同ラベル確率P_j
C

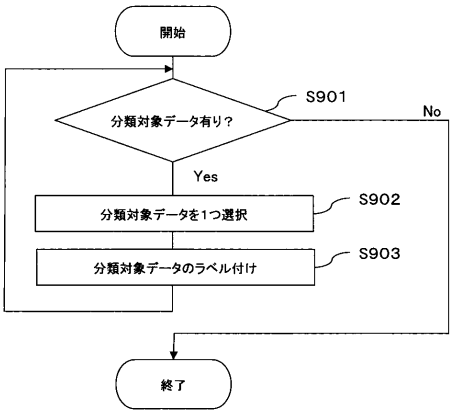
【図７】


信頼度C_j
同ラベル確率P_j
C

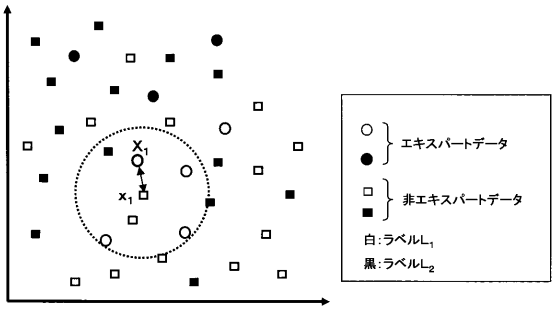
【図８】


開始
均等のデータ重みを付す
S801
データ重みに信頼度を反映
S802
弱学習器を生成
S803
データ重み、コスト関数を更新
S804
No
終了条件を満たす？
S805
Yes
強学習器を生成
S806
終了

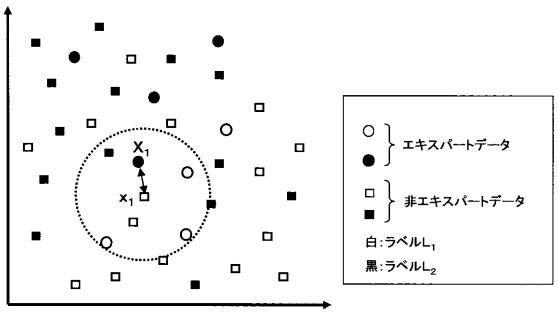
【図９】


開始
分類対象データ有り？
S901
No
Yes
分類対象データを1つ選択
S902
分類対象データのラベル付け
S903
終了

【図１０】


X1
x1
エキスパートデータ
非エキスパートデータ
白:ラベルL1
黒:ラベルL2

【図１１】


X1
x1
エキスパートデータ
非エキスパートデータ
白:ラベルL1
黒:ラベルL2

フロントページの続き

(51)Int.Cl.	FI	テーマコード（参考）
	G06N 5/04 550N	