様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　５　月　３０　日現在

機関番号：　１４６０２
研究種目：　基盤研究(C)（一般）
研究期間：　2015～2017
課題番号：　１５Ｋ００３０７

研究課題名（和文）グラフ構造に基づく情報論的半教師あり学習法の研究

研究課題名（英文）Graph-based Information Theoretic Semi-Supervised Learning

研究代表者
吉田　哲也（Yoshida, Tetsuya）

奈良女子大学・生活環境科学系・教授

研究者番号：８０２９４１６４

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：　データの量や種類の増加に効率的に対処するため，少量の領域知識を活用して性能向上を実現するための技術の確立が求められている．そこで，本研究課題では，グラフ構造に基づく情報論的半教師あり学習法の研究に取り組んだ．具体的には，データ自体の関係を相互情報量に基づいてグラフ構造として表現し，領域知識を制約とみなして正則化に活用する情報論的半教師あり学習法を定式化した．次に，この定式化に基づく最適化学習アルゴズムの開発し，開発したアルゴリズムを計算機システムとしての実装するとともに，開発した手法を実データに適用して評価し，その有効性を確認した．

研究成果の概要（英文）：In order to cope with increasing quantity and variety of data, it is important to develop information technology which enables effective use of domain knowledge. We have developed a graph-based information theoretic semi-supervised learning method. In the developed method, the relationship among data is represented as a graph based on mutual information, and domain knowledge is regarded as constraints and used for regularization. Under the framework of optimization learning, we have developed a semi-supervised learning algorith based on the representation matrix of the graph. The algorithm has been implemented as a prototype system, and experiments over the prototype system were conducted over several benchmark datasets. The results indicate the effectiveness of the developed learning method.

研究分野：　知能情報学

キーワード：　情報工学　機械学習　半教師あり学習

１．研究開始当初の背景

高速なネットワークと廉価な大容量記憶装置の普及により，科学，技術，社会に関する多くの情報が容易に入手可能になったことに伴い，情報洪水に押し流されることなく大量のデータを活用する技術が注目を集めている．データから規則性やパターンを半自動的に取り出すデータマイニングに関する研究が近年活発に行われており，スパムメールの自動フィルタリングなども実用に供されている．

従来のアプローチでは，学習性能はデータのクラスラベルなどの領域知識の量に強く依存するため，性能向上を実現するには多くの領域知識を与えなければならないという課題があった．文書や画像，音声など多種多様なデータがウェブなどを通じて利用可能になるに伴い，データを活用する際に必要となる領域知識の付与に要するコストを低減して性能向上を実現する技術の確立が求められていた．

半教師あり学習とは，大量のデータと併せて少量の領域知識を効果的に活用することで性能向上を目指す機械学習の枠組みである．2003 年の ICML ワークショップを契機として，分類学習やクラスタリング，強化学習などに対して近年様々な手法が提案されてきた．半教師あり学習の枠組みにおいては，少量の領域知識をデータ全体に活用して効果的な学習を実現するために，領域知識における関係とデータ自体における関係を数学的なモデルとして定式化して活用する必要がある．

２．研究の目的

本研究では，少量の領域知識を効果的に活用して性能向上を実現するために，グラフ構造に基づく情報論的半教師あり学習法の研究開発を行う．具体的には，相互情報量に基づく情報論的手法である情報ボトルネック法を拡張して半教師あり学習を実現することに取り組む．

機械学習法としての情報ボトルネック法の有効性は，データとクラスタの関係を相互情報量に基づく最適化問題として定式化することに起因する．そこで，本研究ではこの考え方を一歩進めて，データ自体の関係および領域知識との関係に着目し，データ自体の関係および領域知識との関係を相互情報量に基づき定式化して活用する情報論的半教師あり学習の研究開発を行う．これを実現するため，まずデータ自体の関係を相互情報量に基づいて重み付きグラフとして表現することでデータから導出される場を自然に表現する．さらに，領域知識もグラフのノードラベルや辺ラベルとして統一的にグラフ構造として表現し正則化に活用することにより，最適化に基づく半教師あり学習アルゴリズムを開発することに取り組む．

３．研究の方法

本研究では，グラフ構造に基づく情報論的半教師あり学習法の定式化と，この定式化に基づいた半教師あり学習アルゴリズムの開発を目指した．　具体的には，

(1) グラフ構造に基づく情報論的半教師あり学習の定式化，

(2) 最適化に基づく半教師あり学習アルゴリズムの開発，

(3) システム実装および実データでの評価と検証，

の 3 項目に対する研究開発を行った．それぞれの項目においては，以下の課題に取り組んだ．

(1) グラフ構造に基づく情報論的半教師あり学習の定式化

情報源符号化など，情報論的な妥当性のある処理を行う際には相互情報量に基づくアプローチが多い．本研究では，データ自体の関係を相互情報量に基づくグラフ構造として表現し，構築したグラフを活用して半教師あり学習を実現する学習法を開発することに取り組む．

この目的を実現するため，まず，研究代表者らが開発を進めてきた相互情報量に基づくアプローチを用いることによりデータ間の擬距離を定義し，個々のデータ同士の関係をグラフ構造として表現する．次に，局所最適解（定常確率分布）に対する必要条件に基づき，擬距離からデータ同士の関連度を定義する．さらに，データ同士をその関連度を重みとする有向辺で接続することにより，データ自体の関係を表現する重み付き有向グラフを定義する．

データ同士の関係に加えて，データ対の関係に対する領域知識も活用できるようにグラフ構造を拡張する．これを実現するために，クラスタ割当てにおけるデータ対の関係に対する制約に着目し，この関係もグラフ構造として統一的に表現するため，制約を表現する辺ラベルを導入して上記で定義するグラフ構造を拡張する．

上記で構築するグラフ構造に対して，領域知識が与えられたデータに対するラベルを関連するデータにグラフの辺に沿って伝播させた際の定常解を，グラフ上で隣接するノードに割り当てるラベルの相違と教師ラベルとの乖離に対する正則化項をもつ目的関数の最小化問題として定式化する．

(2) 最適化に基づく半教師あり学習アルゴリズムの開発

高速な学習を実現するため，データに割り当てるラベルを連続変数に緩和し，上記で定式化する最適化問題に対する近似問題を定式化する．次に，グラフの表現行列の固有値分解に基づくスペクトルグラフ理論に基づき，上記で定義するグラフに対応する重み付き隣接行列に対する固有ベクトルを用いて

ノードラベルを決定するスペクトル学習アルゴリズムを開発する.

さらに，学習性能がラベルの伝播に要するステップ数に依存するという課題に対処するため，再帰的なノード伝播に基づく正則化を行う際の伝播作用を推定し，これを領域知識として与えられる教師ラベルに作用させて他のノードラベルを決定する半教師あり学習アルゴリズムを開発する.

(3) システム実装および実データでの評価と検証

開発する半教師あり学習アルゴリズムの有効性を評価するために，アルゴリズム開発用計算機上に,高速な行列演算が可能な R 言語を用いてプロトタイプシステムとして実装する．次に，実装するプロトタイプシステムを機械学習の分野における標準的なベンチマークデータに適用して評価する．他手法との比較を通じて開発する手法の有効性を確認するとともに，実験結果をもとに，開発するアルゴリズムでの計算時間や挙動などを検討し，開発するアルゴリズムの更なる改良を行う.

４．研究成果

上記の３．研究の方法　で述べた項目に対しする成果は下記である.

(1) グラフ構造に基づく情報論的半教師あり学習の定式化

研究代表者らが従来から研究を進めてきたグラフ構造に基づく学習手法を発展させて，相互情報量に基づくデータ間の擬距離を定義し，個々のデータ同士の関係をグラフ構造として表現した．さらに，擬距離からデータ同士の関連度を定義し，データ同士をその関連度を重みとする有向辺で接続する重み付き有向グラフを定義した．さらに，データ対の関係に対する領域知識を表現する辺ラベルを導入した.

次に，上記で定義したグラフ構造に対して,グラフにおいて隣接するノードに対応するデータに割り当てられるラベルの相違と教師ラベルとの乖離に対する正則化項をもつ目的関数の最小化問題として，半教師あり学習を定式化した.

また，上記で定義した目的関数において,最適化基準を拡張した際に隣接行列に対する固有ベクトルが貼る部分空間の構造について，データ同士の局所性を表現する位相の影響などを考察した.

(2) 最適化に基づく半教師あり学習アルゴリズムの開発

上記で定式化した目的関数の最適化を通じて半教師あり学習を実現する際，学習の高速化のために，行列演算に基づく半教師あり学習アルゴリズムを開発した．具体的には,個々のデータに割り当てるラベルの表現を連続変数に緩和して，重み付き隣接行列に対する固有ベクトルを用いてノードラベルを決定するスペクトル学習アルゴリズムを開発した.

さらに，クラスタリングにおける教師ラベルの非対称性に着目して，同一のクラスタに割り当てる場合にはグラフの縮約を活用し,異なるクラスタに割り当てる場合には正則化を活用する半教師あり学習アルゴリズムを開発した.

(3) システム実装および実データでの評価と検証

上記で開発したアルゴリズムを,R 言語を用いてアルゴリズム開発用計算機上にプロトタイプシステムとして実装した．さらに,実装したプロトタイプシステムの評価のために，機械学習の分野において標準的なベンチマークデータとして用いられるベンチマークデータを収集して整備し，収集したデータセットに開発したシステムを適用し，他手法との比較実験を行って開発手法の有効性を確認した．比較における評価指標としてはクラスタリングの精度に対応する正規化相互情報量などを用いて開発手法の有効性を確認した.

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　1　件）

1. Yoshida, T. and Yamada, Y.: A Community Structure based approach for Network Immunization, Computational Intelligence, Vol. 33, No. 1, pp. 77-98 (2017), DOI: 10.1111/coin.12082, 査読有り,

〔学会発表〕（計　3　件）

1. 藤﨑千晶，吉田哲也: 折り紙の数理に基づくスモッキングのデザイン支援，第 31 回人工知能学会全国大会，, 2I1-1, 2017, 査読なし

2. 吉田哲也: 情報論的クラスタリングに対する局所性保存グラフモデル,情報処理学会数理モデル化と問題解決研究会, Vol.2016-MPS-110 No.8., 2016, 査読なし

3. Yoshida, T. and Yamada, Y.: Codebook Graph Coding of Descriptors, International Conference on Parallel and Distributed Processing Techniques and Applications, pp. 223-228, July 27th, 2015, Las Vegas, U.S.A., 査読有り

〔図書〕（計　1　件）
1. Yoshida, T. and Yamada, Y.：
Performance Evaluation of
Two-Dimensional Linear
Discriminant Analysis for Images,
Advances in Mathematics Research.
Volume 21,　chapter 2, 2017, ISBN:
978-1-53610-484-4, 査読なし

〔産業財産権〕
なし

○出願状況（計　0　件）

○取得状況（計　0　件）

〔その他〕
ホームページ等
なし

６．研究組織
(1)研究代表者
吉田　哲也　(Tetsuya　YOSHIDA　)
奈良女子大学・生活環境科学系・教授
研究者番号：80294164

(2)研究分担者
なし

(3)連携研究者
今井　英幸　(Hideyuki　IMAI　)
北海道大学・情報科学研究科・教授
研究者番号：10213216

(4)研究協力者
なし