様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成　24 年　6 月　11 日現在

機関番号：８２６２６
研究種目：若手研究（Ｂ）
研究期間：2010～2011
課題番号：２２７００１９１

研究課題名（和文）　不確かなクラスラベルを持つデータの解析手法に関する研究

研究課題名（英文）　Efficient analysis method for unreliable labeled data

研究代表者
渡辺　顕司（WATANABE KENJI）
独立行政法人産業技術総合研究所・フェロー・産総研特別研究員
研究者番号：５０５７１０６４

研究成果の概要(和文)：生物学系研究分野の信号情報解析などでは、計測サンプルのラベルは、計測対象が物理的・生物学的な不確定性を持っていること、およびラベルが文献情報などを用いて主観的に付与されていることから、確実にモデルで表現できる少数のサンプルに付与されたラベル以外も信頼できるとは限らない。このような場合、ラベルの確からしさ（確信度）を推定するための解析手法を確立する必要がある。そこで本研究では、数量化 IV 類とロジスティック回帰に着目したデータ解析（確信度推定）手法である Logistic label propagation (LLP)を提案した。さらに、LLP に適用し、実データ解析における計算コストを削減するために、非線形共役勾配法を用いた最適化手法に関する予備研究を行った。これら提案手法の性能評価を行ったところ、既存手法と比較して、優れた確信度推定結果を示し、計算コストの削減に成功した。

研究成果の概要（英文）：In the analysis of real data such as the biological signals, the given labels are often unreliable. Because, objects to be measured inherently contain some physical and biological uncertainty, and some labels might be incorrectly assigned by human intuition. Whereas, reliable labels would be available for a small portion of the samples. In such case, a semi-supervised learning method is effectively applied to analyze the data, estimating the label values of samples. In addition, it is favorable that the estimated label values provide us the degree of confidence of each sample. In this research, we proposed a novel method of semi-supervised learning, incorporating logistic functions into label propagation in order to accurately estimate the label values as the posterior probabilities. We call this method logistic label propagation (LLP). In addition, we proposed a novel optimization method for LR by directly using the non-linear conjugate gradient method in order to apply to LLP and to reduce the computational cost. Our proposed methods achieve the better estimation of degree of confidence and the faster computation times compared with the ordinary methods.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２０１０年度	1,100,000	330,000	1,430,000
２０１１年度	600,000	180,000	780,000
年度			
年度			
年度			
総　計	1,700,000	510,000	2,210,000

研究分野：総合領域
科研費の分科・細目：情報学、知覚情報処理・知能ロボティクス
キーワード：ロジスティック回帰、数量化 IV 類、半教師あり機械学習手法、最適化

１．研究開始当初の背景

近年、生物学分野では in vitro（試験管内）の実験から in vivo（生物個体）の実験に至るまで、計測データの数学的な基準を用いた定量化の重要性が高まりつつあり、統計的手法を用いた解析が行われてきている。そのため、このような生物学と情報工学の融合領域研究は、今後さらに重要になると考えられる。

上述のデータ解析では、文献などの事前知識によってラベル付けされた計測データを学習サンプルとして用い、識別器を学習（訓練）することにより解析手法を構築することが一般的である。学習データとして用いる生物学分野のデータを考えてみると、生命現象は確率的な振る舞いをするという内的要因と実験条件の整備や実験者の手技に関する習熟度などの外的要因によって、計測されたデータ（の分布）は大きなばらつきを持ち、対象となるデータが必ずしも実験者の意図したクラスに分類されているとは限らない。

もし、このような不確定なクラスラベルを持つデータを用いて学習を行った場合、真の現象を表現する解析結果は得られないと考えられる。このような誤ったラベルを含んだ学習サンプルを用いる場合には、学習サンプルの分布を用いて統計的にミスラベルを補正する識別空間を形成するべきである。このとき、不確定な（誤った）ラベルを持つ測定データを解析者に教示することは、誤ったラベルを持つデータの計測条件などを明確化することになるので、実験条件の改善などにも大きく役立つものであると考える。

しかし、研究開始当初、少数の確かなクラスラベルを持つサンプルを用いて、統計的基準の下、付与されたラベルの確からしさを教示できる汎化性の高い解析手法は、皆無に等しかった。

２．研究の目的

近年、機械学習手法を用いて多量計測データを定量的に解析（類別・識別）する方法が提案されている。しかし、（特に生物学分野などの）自然科学分野のデータは、不確定な現象を計測しているため、人手による制御が困難であり、文献などの事前知識に基づき付与されるクラスラベルが必ずしも真のラベルを表現していないという問題がある。

そこで本研究では、学習データにおけるクラスラベルの不確定性をも考慮に入れ、学習サンプルの分布を反映した識別空間を形成する（新たな）識別手法を提案し、生物学分野の実測データに適用することによって生物学の質的進展に資する。

３．研究の方法

本研究では、不確かなラベルをラベル無しと見做し、このラベル無しサンプルが各クラスに帰属する確からしさ（確信度）を推定する手法を提案した。このとき、確信度を事後確率と捉えると、ロジスティック関数を用いて以下のように表現できる。

$$\hat{y}_{ik} = \begin{cases} \dfrac{exp\left(\boldsymbol{w}_k{}^{\mathrm{T}}\boldsymbol{x}_i\right)}{1+\sum_{q=1}^{K-1} exp\left(\boldsymbol{w}_q{}^{\mathrm{T}}\boldsymbol{x}_i\right)} & \left(k \neq K\right) \\ \dfrac{1}{1+\sum_{q=1}^{K-1} exp\left(\boldsymbol{w}_q{}^{\mathrm{T}}\boldsymbol{x}_i\right)} & \left(k = K\right) \end{cases} \quad (1)$$

ここで、$\hat{y}_{ik}$ は i 番目の入力サンプル（ベクトル）がクラス k であることの確信度、$\boldsymbol{w}_k$ はクラス k の場合の係数ベクトル、そして、$\boldsymbol{x}_i$ は i 番目の入力ベクトルである。

ロジスティック回帰(Logistic regression: LR)は、教師あり機械学習手法と捉えられる多変量解析手法（回帰手法）であり、ラベル付き学習サンプルが少ない場合、汎化性の高い確信度を推定できるとは限らない。そこで、ラベル無しサンプルも用いた学習を行い、汎化性を向上させるため、半教師あり学習手法である数量化 IV 類(Quantization IV, Label propagation: LP)に着目し、ラベル付きサンプル集合 L における l 番目のサンプルのクラスラベル $\boldsymbol{y}_l \in [0, 1]$と、ラベル無しサンプル集合 U における u 番目のサンプルの確信度 $\hat{\boldsymbol{y}}_u$ (1)を導入し、係数 $\boldsymbol{w}$ を求めることで、確信度推定を行う手法として、Logistic label propagation (LLP)を提案した。

本研究において提案した LLP の評価関数は、予備研究として、以下の定式化を行った。

$$\hat{J}(\boldsymbol{w}) = \sum_{k=1}^{K}\left\{\frac{1}{2}\sum_{i\in U, j\in U} s_{ij}\left(\hat{y}_{ik}-\hat{y}_{jk}\right)^2 + \sum_{u\in U, l\in L} s_{ul}\left(\hat{y}_{uk}-y_{lk}\right)^2\right\} \quad (2)$$

$$\text{s.t.}\ \forall l \in L, \forall k \in \{1,\ldots,K\},\ y_{lk} \in \{0,1\}\ \text{is given}, \quad (3)$$

ここで、y_{lk} はサンプル l におけるクラス k に対するクラスラベルであり、s_{ij} はサンプル i とサンプル j の類似度である。以上の評価関数の下、$\boldsymbol{w}$ に対する最適化問題をといた予備研究版 LLP (2)は、雑誌論文③および学会発表②で発表した。

予備研究版 LLP (3)は、ラベル付きおよびラベル無しサンプルを用いた確信度推定のための多変量解析手法として、優れた射影空間を張ることが出来た。しかし、予備研究版 LLP を機械学習手法として用いて、未知サンプルの識別を行うとき、ラベル付きサンプルが十分に用意できる場合には、汎化性の観点で、LR より優れている保証はない。そこで、LR と同等以上の汎化性を持つ半教師あり機械学習手法として、LR を正則化項として導入

した LLP を雑誌論文①に発表した。この LLP における評価関数は以下のとおりである。

$$J(\boldsymbol{w})=\sum_{k=1}^{K}\left\{\frac{1}{2}\sum_{i\in U,\,j\in U}s_{ij}\left(\hat{y}_{ik}-\hat{y}_{jk}\right)^2+\sum_{u\in U,l\in L}s_{ul}\left(\hat{y}_{uk}-y_{lk}\right)^2-\sum_{l\in L}\eta_l y_{lk}\log\left(\hat{y}_{lk}\right)\right\} \tag{4}$$

このとき、制約条件は(3)と同様である。提案した LLP (4)は、未知サンプルを識別する場合の汎化性能でも、ラベル付き・ラベル無し学習サンプルのサンプル数によらず、LR と同等以上の汎化性能を示した。

以上の提案手法を多量な実データに適用する場合、手法の実用と波及を考えるならば、計算コスト（計算時間、およびメモリー使用量）は、出来うる限り削減するべきである。そこで、予備研究として、計算コスト削減が可能な LR の最適化手法に関する研究を行った。（雑誌論文②および学会発表①）。

正定値性の保障されている LR の最適化はニュートン法などを用いて行われている。一般に、ニュートン法は、最急降下法よりも早い収束が見込める。しかし、ニュートン法を用いた最適化では、ヘシアンの逆行列を算出する必要があり、メモリの使用量が入力ベクトルの次元数 m の二乗のオーダーで増大してしまい、近年の多量な実データの解析手法のための最適化手法としては、適さない場合が多い。最適化手法に関する研究では、これらの問題を克服するために、共役勾配法が研究されている。

共役勾配法は、最急降下法よりも早い収束時間・同程度のメモリ使用量が望める手法であり、最適化問題を解くうえで、広く普及している手法である。LR の最適化に関しても、線形共役勾配法を用いた手法は、すでに提案されている。しかし、線形共役勾配法は、二次関数で定義された評価関数を持つ手法には直接適用出来るが、LR のような対数尤度で表現された評価関数を持つ手法には、直接適用することが出来ない。

本研究（雑誌論文②）では、評価関数の設計によらず適用可能な非線形共役勾配法を用いた LR の最適化手法を提案した。提案した LR の最適化手法では、LR の使用者が任意に与える、直線探索問題を解くための制約条件（Wolfe condition）のパラメータも自動で最適化する手法とした（雑誌論文②参照）。この提案手法では、既存の LR の最適化手法と比較して、$O(m^2) \rightarrow O(m)$のメモリ使用量削減と、約 200 倍の計算高速化を実現した。

４．研究成果

LLP の推定確信度を主成分空間にプロットした実験結果を図 1 に示す。実験には、ベンチマークデータセットである *Iris*（1 サンプルが 4 次元の入力ベクトルであり、クラス数は 3 クラス）を用いた。図 1 における紫、緑、および青色は、それぞれクラスを示し、濃い色ほど高い確信度を示している。確信度推定には、ラベル付きデータ 30 サンプル、ラベル無しデータ 120 サンプルをランダムに選択し、用いた。

実験結果より、LR で推定した事後確率は、緑および紫のクラス間の境界領域まで高い値を示し、また、青のクラス場合、他クラスとの境界領域の方が高い値を示す結果となった。しかし、LLP では、他クラスとの境界領域ほど低く、境界領域から離れるほど高い確信度を推定しており、熟練した研究者の直感的な評価と同等以上な確信度推定を行う事を目指して設計した多変量解析手法として、好ましい結果を示した。また、識別実験において、LLP は、既存手法以上の識別率を示した（雑誌論文①および③を参照）。

次に、最適化手法に関する実験結果を表 1 に示す。比較手法は、LR の最適化において早いという報告のある IRLS-Cholesky である。実験結果より、提案した最適化手法は、*Optdigits* データセットで 10 倍程度、それ以外のベンチマークデータセットで約 200 倍程度高速となった。以上の結果から、提案した LR の最適化手法は、実用に足りる十分な計算速度を持った手法である。その他の詳細な比較結果は、雑誌論文②を参照。

本研究の特色は、サンプルデータのクラスラベルに誤りがある可能性を考慮し、ラベルを確信度という形で確率的に評価する（新しい）統計的パターン認識手法を提案する点である。さらに、分子生物学分野の研究を行ったことのある研究代表者が、実体験を持って理解した計測・解析研究のニーズを解決する手法の提案していることも、特色である。

一般のパターン認識での識別問題では、学習サンプルにおけるクラスラベルに誤りが無いと仮定する。そのため、誤ったラベルを持つデータを学習データに含めてしまった場合、真のクラスを反映した識別結果が得られない。したがって、誤ったラベルを含んだ学習サンプルを用いて識別する場合には、クラスラベルを（確信度で）確率的に評価し、識別する必要がある。

以上の誤った（不確定な）クラスラベルを持つデータの例として、生物学分野の計測データがある。このデータには、（生命現象の確率的な振る舞いによる）内的および（実験条件などによる）外的な要因で不確定なラベルが付与されている。生物学の実験現場では、ラベルの評価は文献情報や研究者の経験などの事前知識を用いて行われてきた。しかし、研究において定量性を確保するためには、クラスラベルの評価方法は統計的な手法を用いて自動化されるべきである。

このようなラベル評価の自動化に統計的

パターン認識手法を適用することは有効な手段の一つであるが、現状では、生物学および統計的パターン認識双方に十分な知見を有する研究者が少ないため、実用に即した統計的なラベル評価・サンプル解析手法の提案は非常に少ない。

したがって、双方の十分な知見を持つ研究者が適切に問題設定を行い、実験結果に示すように、実用に十分足りる（新たな）統計的評価・解析手法を提案したことは、国内外を見渡しても非常に稀有な研究成果である。

今後は、本研究で提案した LLP を LR で示した最適化手法で計算高速化し、実データの評価・解析を行っていきたい。

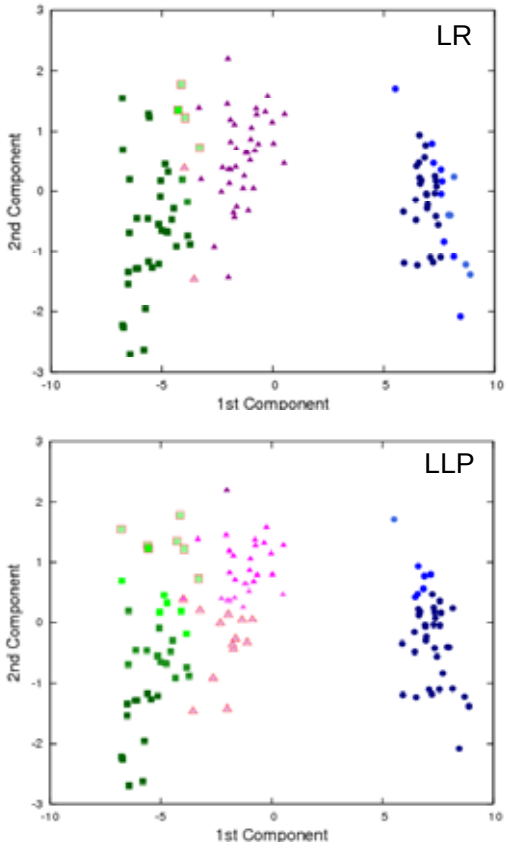

図 1. 確信度推定結果

５．主な発表論文等

表 1. 計算時間の比較結果

	Our method	IRLS-Cholesky
	***Time* (Sec.)**	***Time* (Sec.)**
Optdigits	**0.57**	5.61
Gisette	**61.59**	16371.24
Isolet	**21.03**	9033.63
Semeion	**0.32**	70.69
p53	**199.82**	36749.59

（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 3 件）

① Takumi Kobayashi, Kenji Watanabe, and Nobuyuki Otsu, Logistic Label Propagation, *Pattern Recognition Letters*, 33(5), 2012, pp.580-588, 査読あり
DOI：10.1016/ｊ.patrec.2011.12.005

② Kenji Watanabe, Takumi Kobayashi, and Nobuyuki Otsu, Efficient Optimization of Logistic Regression by Directly Use of Conjugate Gradient, *Proc. ICMLA2011*, pp. 496-500, 査読あり
DOI：10.1109/ICMLA.2011.63

③ Kenji Watanabe, Takumi Kobayashi and Nobuyuki Otsu, Logistic Label Propagation for Semi-supervised Learning, *Part I, Lecture Notes in Computer Science (LNCS)* 6443, 2010, pp.462–469, 査読あり
DOI：10.1007/978-3-642-17537-4_57

〔学会発表〕（計 2 件）

① Kenji Watanabe, Takumi Kobayashi, and Nobuyuki Otsu, Efficient Optimization of Logistic Regression by Directly Use of Conjugate Gradient, *The 10th ICMLA2011*, Honolulu, Hawaii, USA, (December, 2011)

② Kenji Watanabe, Takumi Kobayashi and Nobuyuki Otsu, Logistic Label Propagation for Semi-supervised Learning, *17th ICONIP2010*, Sydney, Australia, (November, 2010)

６．研究組織

(1)研究代表者

渡辺　顕司（WATANABE　KENJI）
独立行政法人産業技術総合研究所･フェロー・産総研特別研究員
研究者番号：５０５７１０６４

(2)研究分担者

(3)連携研究者

(4)研究協力者

大津　展之（OTSU　NOBUYUKI）
独立行政法人産業技術総合研究所･フェロー・フェロー

小林　匠（KOBAYASHI　TAKUMI）
独立行政法人産業技術総合研究所･情報技術研究部門・研究員
研究者番号：３０４４３１８８