科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　２０　日現在

機関番号：１６３０１
研究種目：基盤研究(C)
研究期間：2010～2013
課題番号：２２５００１２１

研究課題名（和文）重要な特徴を自動的に発見する系列ラベリング学習の研究

研究課題名（英文）A Study on Methods for Automatically Finding Important Features in Sequential Labeling

研究代表者
二宮　崇（NINOMIYA, TAKASHI）

愛媛大学・理工学研究科・准教授

研究者番号：２０４４４０９４

交付決定額（研究期間全体）：（直接経費）　　3,300,000円、（間接経費）　　990,000円

研究成果の概要（和文）：自然言語処理における識別モデルの素性関数は人手による試行錯誤で設計されているが、数十万から数百万に及ぶ素性関数を人手で発見・制御することは非常に困難な作業となっている。本研究は、素性関数を自動的に選択・構築しつつ目的関数を最適化するオンライン・グラフティングの効率化およびアンサンブル学習による高精度化の研究を行う。本研究の提案手法はL1正則化ロジスティック回帰の近似手法となっているが、実験により、従来法の精度を低下させることなく、学習の高速化を実現することを経験的に示した。また、確率的アルゴリズムを用いたアンサンブル学習によりオンライン・グラフティングの精度が向上することも確認した。

研究成果の概要（英文）：In natural language processing, millions of feature functions are defined for the discriminative models used in many natural language tasks. These feature functions are elaborated by human experts, but it is obviously not easy even for the human experts to find and develop such millions of feature functions by hands. This research proposes efficient methods for online grafting and ensemble methods for improving accuracy of online grafting. Online grafting is a method for automatically selecting features and optimizing the parameters in L1-regularized logistic regression. The experiments have shown that our methods significantly improved efficiency of online grafting. Though our methods are approximation techniques, deterioration of prediction performance was negligibly small. The ensemble methods using probabilistic algorithms achieved to improve the accuracy of online grafting.

研究分野：総合領域

科研費の分科・細目：情報学・知能情報学

キーワード：自然言語処理　機械学習　オンライン学習　素性選択　ロジスティック回帰　L1正則化

１．研究開始当初の背景

近年の自然言語処理における多くの解析システムは、条件付き確率場に代表される構造予測のための確率的識別モデルや、SVM 等のマージン最適化によるモデル化と学習により高精度化が実現されている。識別モデルにおいて、入力から素性ベクトルを構成するための関数は素性関数(もしくは単に素性)と呼ばれ、現在これらの素性関数は人手による試行錯誤で設計されているが、数百万次元に及ぶ自然言語処理のための素性ベクトルを人手で発見・制御することは非常に困難な作業であり、高精度化に対する大きな課題となっている。特に、自然言語処理においては、1 単語や 1 品詞に対する素性関数を合成した *n* 単語/*n* 品詞の共起 (*n*-グラム)に対する素性関数を設計することが多く、これらの組み合わせの中から有効な組み合わせの *n*-グラムを発見することは非常に難しい。

機械学習分野では、「素性選択」と呼ばれる技術分野の研究が存在し、非常に高い次元の素性ベクトル(すなわち、非常に多くの素性)を低い次元の素性ベクトル(重要な素性)へ射影する、もしくは重要な次元(素性)のみを選択する技術、または、基本となる素性関数から複雑な素性関数を合成する技術の研究が行われている。代表的な技術としては、主成分分析(PCA)や潜在意味インデキシング(LSI)、χ 二乗検定/相互情報量による選択が存在するが、これらの手法はデータから得られる情報のみを元に目的関数とは無関係に重要な素性を抽出する手法であり、目的関数を最大化する素性を抽出する手法にはなっていない。つまり、自然言語処理における各タスクにとって目的関数をより大きくする重要な素性が存在するにも関わらず、そのような素性が選択・構築されるとは限らない。目的関数を最大化しつつ、素性を選択する手法として、正則化項(パラメータのペナルティ項)に L1 ノルム(=パラメータの絶対値の和)を適用した L1 正則化もしくは lasso と呼ばれる手法が存在する。しかし、L1 正則化は最初に全ての素性集合を与え不要な素性を除去する手法であるため、指数爆発的数の合成素性集合から重要な素性を選択することは難しい。Perkins らにより 2003 年に提案されたグラフティングと呼ばれる手法は、素性選択と目的関数の最適化を同時に与える手法であり、最初は空の素性集合から始め、徐々に素性を増やすため、指数爆発的数の素性集合に対しても適用可能となる。全てのパラメータに対する勾配が計算できれば、最適解が得られることが保証されており、タスクの性質とアルゴリズムによっては全ての素性の合成に対する最適解が得られる。しかし、その一方で、グラフティングによる最適化は、訓練データに対し過学習することが知られており、非常に頻度の低い非常に複雑な合成素性関数が選択され、結果として、グラフティングによる精度の高い識別器の実現は難しい。

２．研究の目的

本研究は、素性関数を自動的に選択・構築しつつ目的関数を最適化するオンライン・グラフティングの効率化およびパラメータ平均化による高精度化の研究を行う。オンライン学習は機械学習の分野で最先端のトピックであり、SVM やロジスティック回帰等の全データを記憶する必要がある学習に比べ、逐次的にデータを解析することで学習が行われるため、学習時間が短くメモリ効率にも優れている。解析精度において難点があったが、最新の研究により SVM よりも高い精度を達成するオンライン学習を実現することがわかってきた。グラフティングのオンライン学習(オンライン・グラフティング)は Perkins らによってすでに提案されているが、素性選択は高い計算コストを要する学習であるため、本研究ではオンライン・グラフティングを改良することにより、高効率な学習の実現を目指す。また、グラフティングにおける過学習の問題に対しては、アンサンブル学習の一種であるパラメータ平均化を用いて解決する。

本研究では次の(1)～(3)の研究を行う。

(1) オンライン・グラフティングの効率化
(2) パラメータ平均化によるオンライン・グラフティングの高精度化
(3) 品詞解析における L1 正則化 CRF のパラメータ平均化

３．研究の方法

本研究では、２．であげた目的(1)～(3)の目的に対し、以下の研究を行った。

(1)オンライン・グラフティングの効率化

オンライン・グラフティングは素性選択と目的関数の最適化を同時に与える手法であるが学習に時間を要することが知られている。本研究ではオンライン・グラフティングの学習における素性選択に必要とされる再学習の回数を抑えることにより、オンライン・グラフティングの効率化を実現する手法を提案し、実験を通してその性能を検証した。

一般のオンライン学習はデータ列からデータが順次与えられることを想定した学習法であるが、オンライン・グラフティングは素性関数列から素性関数が順次与えられることを想定したオンラインアルゴリズムである。ある素性 *f* が与えられたとき、オンライン・グラフティングは素性 *f* のパラメータ *w* に対する目的関数の勾配を計算し、その勾配が一定以上であるときにその素性 *f* を選択する。勾配を正確に計算するためには素性を一つ新しく選択するたびに識別器のパラメータを最適にする必要があるが、パラメータ最適化は非常に大きな計算コストを要するため、素性を一つ選択するたびにパラメータ最適化を行っていては学習全体で多大な時

間が必要となってしまう。

本研究ではオンライン・グラフティングの学習効率化のために二つの手法を提案した。オンライン・グラフティングの従来手法においては素性選択のテストを高い精度で実現するため、常に識別器のパラメータを最適にしていたが、本研究の一つ目の提案手法では、一定回数の素性選択を行うごとにパラメータを最適化し、二つ目の提案手法では、倍数回の素性選択を行うごとにパラメータを最適化することを行った。これらの手法では常に最適化を行っているわけではないためオンライン・グラフティングの近似手法となっている。そのため効率と予測性能の間にはトレードオフが存在するが、実験によってそのトレードオフを検証する。

実験ではオンライン・グラフティングの確率モデルとしてロジスティック回帰を用いる。ロジスティック回帰は入力 x と出力 $y \in \{-1, +1\}$ に対し、次式で与えられる確率的識別モデルである。

$$p(y=+1 \mid x) = \frac{1}{1+e^{-w\cdot\phi(x)}}$$

ただし、w は重みベクトルと呼ばれるパラメータであり、φは素性ベクトルを返す関数である。今回の実験におけるオンライン・グラフティングおよび提案手法は L1 正則化ロジスティック回帰を最適解とする近似手法となるため、実験において L1 正則化ロジスティック回帰とも比較を行う。L1 正則化ロジスティック回帰はラプラス分布を事前分布とした MAP 推定によりそのパラメータが求まる。

(2)パラメータ平均化によるオンライン・グラフティングの高精度化

本研究ではオンライン・グラフティングの過学習を抑制するために、バギングやオンライン BPM などで用いられている確率的アルゴリズムを用いたアンサンブル学習の提案、実装、および実験を行った。

オンライン・グラフティングのパラメータ平均化は、オンライン・グラフティングにより学習された複数の識別器のパラメータを平均化することにより実現し、次の二つの平均化手法を試みた。

① 元の訓練データからランダムにデータを削減し、新たに複数の訓練データを作成する。得られた複数の訓練データから複数の識別器を学習し、それらのパラメータの平均を得る。

② 元の訓練データからランダムに素性を削減し、新たに複数の訓練データを作成する。得られた複数の訓練データから複数の識別器を学習し、それらのパラメータの平均を得る。

①と②のそれぞれの方法で生成した訓練データを $D^{(1)}, D^{(2)}, \ldots, D^{(L)}$ としたとき、最終的に求めるパラメータ w は以下のようになる。

$$w^{(l)} = \arg\min_{w} C(w, D^{(l)})$$

$$w = \frac{1}{k}\sum_{l=1}^{k} w^{(\sigma(l))}$$

ただし、C は目的関数、$\sigma(l)$はパラメータを精度の高い順に並び替えるソート関数であり、最終的に得られるパラメータは精度の高い k 個パラメータの平均となる。

(3)品詞解析における L1 正則化 CRF のパラメータ平均化

本研究では自然言語処理の代表的なタスクである品詞解析においてよく用いられる CRF(条件付き確率場)のパラメータ平均化を行い、その性能評価を行った。

CRF は、系列のための確率的識別モデルであり、系列 x とそのラベル列 y に対し、次式で定式化される。

$$p(y \mid x) = \frac{1}{Z_x} e^{w\cdot\phi(x,y)} = \frac{1}{Z_x}\prod_{c} e^{w\cdot\phi(c,x)}$$

ただし、w は重みベクトルと呼ばれるパラメータ、φは素性ベクトルを返す関数、Z_x は分配関数と呼ばれる正規化のための項、c は y における依存関係のあるノードである。

L1 正則化 CRF の目的関数はパラメータに対し凸であるため、最適解を唯一持つが、実際に学習を行う場合は収束条件により本当の最適解よりも少しずれた解が求まる。本研究ではパラメータ推定においてパラメータの初期値を乱数で与え、いくつかのパラメータを求め、それらの平均をとることで性能がどのように変化するのか検証した。

４．研究成果

(1)オンライン・グラフティングの効率化の実験結果

実験データには機械学習の研究で良く用いられている LIBSVM Data の a9a, w8a, IJCNN1, news20.binary を用いた。a9a はある人が年に 50,000USD 稼げるかどうか判定するタスクであり、w8a と news20.binary は文書分類のタスクである。IJCNN1 は IJCNN の競技会で用いられたデータである。提案手法は Python を用いて実装し、パラメータ最適化には Python 用の LIBLINEAR パッケージを用いた。計算環境には 1,900 コアからなる分散計算環境の InTrigger を用いた。

次の表 1、2、3、4 は実験結果を示している。

表 1: a9a に対する実験結果

	精度 (%)	累積最適化パラメータ数	実学習時間 (秒)
提案手法(定数分割)	85.24	6,970	23.42
提案手法(倍数分割)	85.25	6,772	10.45
従来法(オンライン・グラフティング)	85.19	505,822	5,952.71
L1-LR	85.19	15,252	10.78

表 2: w8a に対する実験結果

	精度 (%)	累積最適化パラメータ数	学習時間 (秒)
提案手法(定数分割)	99.07	39,807	47.03
提案手法(倍数分割)	99.04	38,209	17.26
従来法(オンライン・グラフティング)	99.05	8,833,804	56,029.97
L1-LR	99.11	90,300	24.40

表 3: IJCNN1 に対する実験結果

	精度 (%)	累積最適化パラメータ数	実学習時間 (秒)
提案手法(定数分割)	97.61	8,810	332.75
提案手法(倍数分割)	97.64	742	20.44
従来法(オンライン・グラフティング)	97.61	68,438	2,633.61
L1-LR	97.62	506	15.67

表 4: news20.binary に対する実験結果

	精度 (%)	累積最適化パラメータ数	実学習時間 (秒)
提案手法(定数分割)	94.96	221,574	20.56
提案手法(倍数分割)	94.90	37,176	11.77
従来法(オンライン・グラフティング)	95.00	16,302,937	15,994.70
L1-LR	96.22	1,355,191	12.05

従来法はオンライン・グラフティングを示しており、L1-LR は L1 正則化ロジスティック回帰を示している。提案手法は近似解を求める手法となっているため、効率と予測性能との間にはトレードオフがあったが、実験結果より、予測性能を低下させることなく、大幅に学習効率を改善できることを経験的に示した。二つの提案手法のいずれも最適化回数を削減したことで累積最適化パラメータ数(実際に最適化されたパラメータの累積数)および学習に要した実計算時間が大きく減少し、学習効率を大幅に向上させた。予測性能についてはどちらの提案手法もオンライン・グラフティングおよび L1-正則化ロジスティック回帰とほぼ同程度であり、いくつかのデータセットにおいては、オンライン・グラフティングや L1-正則化ロジスティック回帰の予測性能を上回った。

この研究成果は、the 5th International Conference on Agents and Artificial Intelligence (ICAART 2013)および日本データベース学会論文誌において発表した。

(2) オンライン・グラフティングのパラメータ平均化による高精度化の実験結果

提案手法の性能評価のために LIBSVM Data の IJCNN1 に対し実験を行った。全データは 141,691 個のデータポイントから成り、訓練データとして 70,847 個、開発用テストデータとして 35,422 個、テストデータとして 35,422 個に分割して使用した。今回の実験では 40 個の識別器を作成した。データの削減率を 50%、60%、70%、80%、90%とし、平均数(k)を 5、10、15、20、30、40 として比較評価した。

次の表 5 はその実験結果を示している。

表 5: パラメータ平均化の精度

従来法	①の手法	②の手法
93.09%	93.61%	93.78%

従来のオンライン・グラフティングでは 93.09%の精度であったが、①の手法で 93.61%の精度まで向上し、②の手法では、93.78%の精度まで向上した。

①の手法で行った実験結果の詳細を図 1 に、②の手法で行った実験結果の詳細を図 2 に示す。

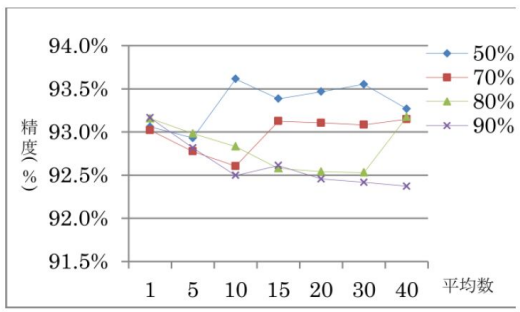

図 1: ①の実験結果

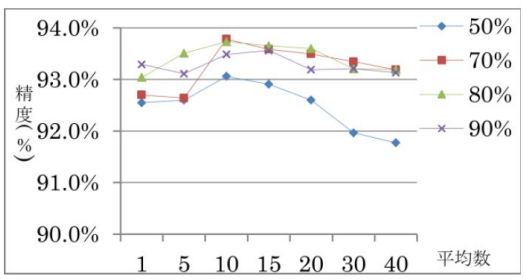

図 2: ②の実験結果

平均数 1 は平均化前のパラメータの中で最も精度が高かった識別器の精度を示している．図 1，図 2 それぞれにおいて，平均化前の最高精度と平均化後の最高精度を比較して精度の向上が確認できた。

(3) 品詞解析における L1 正則化 CRF のパラメータ平均化の実験

L1 正則化 CRF による英語品詞解析の実験

を行った。実験データには Penn Treebank 2 を用いた。Penn Treebank 2 の Wall Street Journal 部分を用い、Section00 から 02 までを訓練データとして用い、Section19 から 21 までを開発用テストデータ、Section22 から 24 までをテストデータとして用いた。

次の表 6 は品詞解析の実験結果を示している。

表 6: 品詞解析の精度評価

	従来法	提案手法
開発用テストデータ	96.0317%	96.0370%
テストデータ	95.7950%	95.7950%

実験結果から示されるように開発用テストデータでは若干の精度向上がみられるものの、テストデータでは同じ精度となった。L1 正則化ロジスティック回帰においてパラメータの初期値に乱数を与えることで性能に違いがでることがわかったが、従来法とほぼ同程度の精度となった。素性数を大きく増やしたときに平均化の効果が得られる可能性があるため、将来の課題としたい。

５．主な発表論文等

〔雑誌論文〕（計 3 件）

① Yo Ehara, Nobuyuki Shimizu, Takashi Ninomiya and Hiroshi Nakagawa: Personalized Reading Support for Second-Language Web Documents, ACM Transactions on Intelligent Systems and Technology (ACM TIST), 査読有, vol. 4, no. 2, pp. 31:1-31:19, 2013.
DOI: 10.1145/2438653.2438666

② 大井健吾, 二宮崇: 複数素性の同時テストによるオンライングラフティングの効率化. 日本データベース学会論文誌, 査読有, vol. 11, no. 3, pp. 15-20, 2013 年
http://dbsj.org/wp-content/uploads/journal/vol11/no3/dbsj-journal-11-03-015.pdf.pagespeed.ce.qlYRfTVCEe.pdf

③ Kengo Ooi and Takashi Ninomiya: Efficient Online Feature Selection Based On l1-Regularized Logistic Regression. In Proceedings of the 5th International Conference on Agents and Artificial Intelligence (ICAART 2013), 査読有, pp. 277-282, 2013.
http://www.icaart.org/Home.aspx?y=2013

〔学会発表〕（計 4 件）

① 岩田 匠, 二宮 崇: カテゴリ毎に異なるナイーブベイズ分類器を用いた大規模評判分析.平成 24 年度 電気関係学会四国支部連合大会 講演論文集, pp. 285-285 (優秀発表賞を受賞), 2012 年 9 月 29 日, 香川県高松市.

② 矢野 裕一朗, 二宮 崇: ディリクレ事前分布付き隠れマルコフモデルにおけるスムージング効果の調査.平成 24 年度 電気関係学会四国支部連合大会 講演論文集, pp. 286-286, 2012 年 9 月 29 日, 香川県高松市.

③ 大井 健吾, 二宮 崇: オンライングラフティングのパラメータ平均化による集合型学習.平成 23 年度 電気関係学会四国支部連合大会 講演論文集, pp. 258-258, 2011 年 9 月 23 日, 徳島県阿南市.

④ 大井 健吾, 二宮 崇: オンライングラフティングのアンサンブル学習. 情報処理学会 第 73 回全国大会講演論文集, vol. 1, pp. 301-302, 2011 年 3 月 4 日, 東京都目黒区.

〔図書〕（計 0 件）

〔産業財産権〕
○出願状況（計 0 件）
○取得状況（計 0 件）

〔その他〕

① 二宮 崇: 主辞駆動句構造文法のための同期文法の実現に向けて.工学ジャーナル, vol.11, pp. 178-183, 愛媛大学工学部, 2012 年.

② 二宮 崇: シーズ(研究成果)探訪 vol.81, データの自動分類とテキストの構文解析—高速化と高精度化— 自動的に特徴を学習するオンライン学習と言語学的文法に基づく構文解析, 月刊愛媛ジャーナル vol.25 no.7, p.80-82, 2011 年.

③ 二宮 崇: HPSG 構文解析とスーパータガー. 愛媛大学数学談話会にて講演, 2011 年.

６．研究組織

(1)研究代表者
二宮 崇（NINOMIYA TAKASHI）
愛媛大学・大学院理工学研究科・准教授
研究者番号：20444094

(2)研究分担者
なし

(3)連携研究者
なし