様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ６ 月 １７ 日現在

機関番号：３３９２４
研究種目：基盤研究(C)（一般）
研究期間：2013～2015
課題番号：２５３３０２７１

研究課題名（和文）Big Dataからの階層的分類技術

研究課題名（英文）Hierarchical Classification from Big Data

研究代表者
佐々木　裕（Yutaka, Sasaki）

豊田工業大学・工学（系）研究科（研究院）・教授

研究者番号：６０３９５０１９

交付決定額（研究期間全体）：（直接経費）　　2,900,000円

研究成果の概要（和文）：超大規模階層的文書分類データセットであるLSHTC3 Wikipediaデータを対象に，高速な学習手法を考案し，高い予測性能を実現した．特に，Wikipedia Medium データを対象にした学習時間を30分程度に短縮することに成功した．従来手法では，数時間～数日の学習時間を必要としていた．テスト時の分類性能が，世界最高スコアを達成していることも示した．さらに，各特長の分散ベクトル表現から新しい特徴を生成し，元の特徴ベクトルに追加することで，階層的分類の精度が44.92%にまで向上することを示した．階層的分類システムEzeをオープンソースソフトウェアとして公開した．

研究成果の概要（英文）：We constructed fast and accurate hierarchical classification systems on the basis of the LSHTC3 Wikipedia data, which are huge hierarchical classification datasets. The training time of our system on the LSHTC3 Wikipedia Medium data has been reduced to 30 minutes. Conventional methods for the same data took several hours or even several days. The predictive performance for the test data showed the world highest scores. Moreover, we generated new features based on the distributed embedding vectors which have been created from the original features. Adding the new features further improved the predictive performance over the test data to 44.92%. We made our hierarchical classification system Eze publicly available as open-source software.

研究分野：人工知能

キーワード：階層的分類　機械学習　LSHTC3　分散ベクトル表現　文書分類　Big Data

１．研究開始当初の背景

（１）この 20 年間で，機械学習技術は急速に発展してきた．特に，教師あり学習の枠組みのなかで，訓練データに基づき，分類モデルを学習し，未知データに対する分類性能を向上させるための様々な手法が提案され，速度・精度の向上のみならず適用範囲の拡大やデータ規模の拡大が進んできた．

たとえば，Support Vector Machine (SVM) は，クラス Y={+1,-1}に関する，2 クラス(binary class)分類器として提案されたが，クラス Y={1,…,r}の多クラス(multiclass)分類器に発展し，さらに，多クラス他ラベル(multiclass multilabel)分類器へと発展してきた．

（２）一方，分類問題は，教師なし学習，及び半教師あり学習の枠組みにおいても注目を集めていた．教師なし学習は，教師情報（＝正解ラベル）が与えられていない事例集合に内在する特徴を用いて，事例をグループ化し，未知事例の予測に用いるものである．主として，特定の類似度尺度に基づくクラスタリングを用いる．

半教師あり学習は，少数の教師データと大量の教師なしデータに基づき，教師なしデータの特性を少数の教師ありデータに対する学習に組み込むことにより高性能な分類を実現する．これらのどちらのアプローチも，教師情報付きデータの量が限られている状況を想定している．

医療データや Wikipedia 等のインターネット上の情報源において，大量の教師情報と大量の教師なし情報が得られる場合も多く，このような巨大な訓練データが与えられた状況において，分類問題をどのように解くかが課題となっていた．

研究開始当初の時点では，word2vec 等の単語埋め込み（Word Embedding）技術は提案されていなかったが，これは教師なし学習による特徴の拡張技術として，まさに本研究で狙っていたものであり，本研究期間の後半において活用した．

２．研究の目的

（１）本研究では，「機械学習における Big Data」（従来の機械学習技術では取り扱えない規模の巨大データ）を扱うことを目指す．具体的には，クラス階層構造を活用した高精度・高効率な超大規模階層的データ分類技術を確立することを目的とする．

これまで，分類対象クラスが構造をもたないフラットな分類学習や分類結果が構造を持つ構造学習の研究が進められてきた．本研究では，超大規模なクラス集合が階層構造を持つデータ分類問題を高精度かつ高効率に解決する仕組みを解明する．具体的には，階層構造を活用した新しい識別学習により，学習・分類の効率を向上させることより，Wikipedia から抽出された数十万ノードからなる構造的クラスへの分類を，数百万データに基づいて学習することを可能にする．

３．研究の方法

（１）Wikipedia Medium は約 45 万データを訓練データとして，約 5 万ノードからなるカテゴリの階層に対して，約 8 万件のテストデータを分類するという大規模階層化データ分類タスクである．さらに，Wikipedia Large データは，Wikipedia Medium を超える巨大なデータである．クラス構造は約 325,000 ノードからなり，訓練データ数は約 2,400,000 件となる．

機械学習技術を用いて，このような超大規模データを扱うための第１ステップとして，現状の階層化分類技術のメモリ効率，実行効率を大幅に改善する．具体的には，SGD SVM の改良手法を導入することにより，精度を落とすことなく，Wikipedia Medium データに対する学習・分類時間を改善する．

（２）次に，Wikipedia Large データを対象に学習・分類手法を改良していく．Wikipedia Medium データに対して，高速化された手法を実際に超大規模データにおいて評価し，パラメータのチューニングを行う．もし，処理速度や必要メモリサイズの制約により十分な性能が得られない場合には，GPGPU や MapReduce 手法に基づく超並列処理により数倍から数十倍の高速化を実現することを想定していたが，高速な学習アルゴリズムを導入し，階層的分類処理の全体を C++により効率よく処理する実装を実現したことにより，ファイルの入出力を含めた実際の学習・分類時間の大幅な短縮を達成できた．

（３）Wikipedia データにおいて，さらなる予測性能を向上させるためには，学習手法の改良では限界がある．そこで，特徴ベクトルを拡張することにより，性能を向上させる．

先に述べたように，研究開始当初の時点では，word2vec 等の単語埋め込み技術は提案されていなかったが，まさに本研究で狙っていた教師なし学習手法であり，本研究に導入した．

単語埋め込み技術を用いて，与えられた各特徴に対して，同じデータ内に存在する特徴に関する埋め込みベクトルを生成した．これにより，特徴間の距離が数値ベクトルにより求められた．この特徴埋め込みベクトルと元の特徴ベクトルを連結することにより新しい特徴ベクトルを作成した．

（４）研究の過程で開発したソフトウェアシステムを整理し，パッケージ化し，外部に公開する．パッケージ化されたソフトウェアは，GPL 2.0 に基づき，階層的分類システム Eze として，一般に公開した．

４．研究成果

（１）考案した高性能な階層的分類手法は以下の４つのステップからなる．

1. ボトムアップ伝播
2. トップダウン学習
3. トップダウン分類
4. 大域的枝刈り

①ボトムアップ伝播

学習フェーズにおいてカテゴリ階層構造を利用するためには，訓練データ ID が階層構造に対応づけられていなければならない．LSHTC3 訓練用データには末端のカテゴリにしか付与されていないため，学習の前処理として，訓練データ ID を階層構造に従って末端ノードからルートに向ってボトムアップに伝播する必要がある．階層構造は複数の親ノードを許すため，複数の親ノードがある場合は，分岐しながらボトムアップに階層に訓練データ ID を割当る．ここで，データ ID を末端からルートに向って伝播するのではなく，末端ノードに割当られているデータを記憶しておき，末端ノードの ID をルートに向って伝播する．

②トップダウン学習

学習はトップダウンに行う．対象のノードに伝播された末端ノード集合を対象に，各エッジについて，その下位（子）カテゴリに伝播されている末端ノード集合に割当られているデータを正例とし，その他を負例として SVM モデルを学習し，エッジに関連付ける．

学習の高速化のために，DCASVM を用いた．これは，Dual Co-ordinate Ascent SVM にミニバッチ を導入したものである．ここで $\mathrm{clip}_{[a,b]}(\cdot)$ は，引数を[a,b] のレンジに切取る関数である．Pegasos や SGD SVM は，基本的には DCASVM と同様に高速な SVM 学習アルゴリズムである．しかし，Pegasos や SGD SVM は SVM 最適化問題の主問題を解いているため，KKT 条件を停止条件として利用できず，何回の繰り返しが必要となるかが明確ではない．そのため，従来，繰り返し回数は，余裕を持って「データ数× 100」を利用していた．一方，DCASVM は双対問題を解いているため，KKT 条件による収束判定が可能となり，無駄な繰り返しを大幅に削減できることを，大規模階層的分類データを対象に明らかにした．

③トップダウン分類

テストフェーズにおいて，データ x に関するリンク (n_1,n_2) の SVM の出力値を $g_{(n1,n2)}(x)$ とするとき，Clipped Classification Score (CCS)を以下のように定義する．

$$CCS_{(n_1,n_2)}(x) = \frac{clip_{[-1,+1]}(g_{(n_1,n_2)}(x)) + 1}{2}$$

Accumulative Clipped Classification Score (ACCS) は下記のように定義される．

$$ACCS(x,m) = \prod_{e \in E} CCS_e(x)$$

ここで E は root からノード m の親ノード至るリンクの集合である．各リンクでの分類を $g_{(n1,n2)}(x)+(ACCS(x, n_2)\times 2-1) > 0$ により判断することで確信度の高い分類パスにおいて再現率を改善する．次節の枝刈りで怪しいカテゴリの割り当てを削除することで適合率を維持する．

④大域的枝刈り

ACCS 値を確信度と看做し，確信度に閾値を設け，大域的に枝刈りする．ただし，確信度が低くとも，少なくとも１つのカテゴリを各データに残す．確信度に対する閾値は，過去の研究によりテストセットへの平均カテゴリ付与数が 1.5 となる値とした．

（２）埋め込みベクトル表現により，以下のように特徴を追加する方法を考案した．データ全体に M 種類の特徴が含まれているとき，ある特徴 f_i の分散ベクトル表現を，f_i を含む事例に同時に出現（共起）する特徴を用いて作成する．これは，文中の単語の共起に基づいて単語の埋め込みベクトル表現を学習することと同じである．特徴 f_i に対応する N 次元の埋め込みベクトルをベクトル $\boldsymbol{v}_i$ とすると，埋め込み行列 $\boldsymbol{V}$ は

$$V = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_M \end{bmatrix} = \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1N} \\ v_{21} & v_{22} & \cdots & v_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ v_{M1} & v_{M2} & \cdots & v_{MN} \end{bmatrix}$$

で表される．

単純には，各事例に出現する特徴の埋め込みベクトルの平均を取ることで，事例の埋め込みベクトルを計算することができる．しかし，このような単純な方法では，事例の特徴が N 次元に圧縮されるときの歪により予測性能が低下してしまう．

そこで，オリジナルの事例の入力ベクトルを $\boldsymbol{x}$ とすると，新しい特徴ベクトル $\boldsymbol{x}'$ は

$$\boldsymbol{x}' = \boldsymbol{x}\boldsymbol{V}$$

により生成される N 次元のベクトルとする．最終的な特徴ベクトルは，オリジナルの入力ベクトルと生成されたベクトルの連結

$$(\boldsymbol{x}, \boldsymbol{x}')$$

とする．

（３）実験結果

他の学習手法との比較結果を表 1 に示す。Pegasos と SGD SVM が最も高い性能を示したが、DCASVM は、11 分の学習時間でこれらに近いスコアを残した。他の学習アルゴリズム

は、DCASVM より数倍以上の学習を要した。テストフェーズにおける分類時間はどの手法も約 10 分程度である。ファイル入出力やオーバヘッドを含めた DCASVM の学習時間は 31 分であり、ほぼ目標を達成した。LSHTC3 の 1〜2 位システムの学習時間は明らかではないが、十数時間〜数日を要すると推定される。超大規模な階層的分類データに適用することを考えると，DCASVM を用いることが，効率面と性能面を考慮した場合の解となる．

表１　学習手法による比較

学習手法	Accuracy
DCASVM	0.4445
Pegasos	0.4459
SGD-SVM	0.4457
Passive Aggressive	0.4005
ROMMA	0.3814
Logistic Regression	0.3515
LSHTC3 上位システムの性能	
(1st) arthur	0.4382
(2nd) coolvegpuff	0.4291
(3rd) TTI	0.4200

このように、DCASVM を用いることにより、大規模階層的分類の学習を 30 分程度に抑えることができた。これにより、表 2 に示すような階層の深さごとのスコアを検証することが現実的になる。この結果は、学習フェーズにおいて、各リンクに対応する SVM 分類器を学習する際のデータに対して 2 分割交差確認を行った結果である。ただし、ここでデータは上位からの分類の影響を受けておらず、各リンクでの正解データを用いた交差確認の結果である。なお、交差確認を用いて、上位から下位のノードにデータを分類しながら流していく学習の実験も行ったが、階層が深くなるほど分類誤りが蓄積していき、十分な学習データが確保できなかった。

表２　階層の深さによる性能

Depth	Accuracy	MacroF1	Micro F1
1	0.9535	0.6615	0.9762
2	0.5415	0.4656	0.7025
3	0.6378	0.4606	0.7789
4	0.6722	0.4684	0.8039
5	0.7326	0.4429	0.8457
6	0.6537	0.4443	0.7906
7	0.6912	0.4841	0.8174
8	0.6803	0.4674	0.8097
9	0.7614	0.5049	0.8646
10	0.7572	0.5282	0.8618
11	0.4737	0.5889	0.6429

特徴の埋め込みベクトルを追加することで，さらに Accuracy が向上し，44.92%を達成した．これは，現時点で我々が知る限り，Wikipedia Medium データに対する世界最高のスコアである．

Wikipedia Large データに関しても，約 50 時間の学習時間で学習が完了し，世界最高レベルの予測性能が得られることを確認した．

表３に LSHTC4 トラック１の結果を示す．

表３　学習手法による比較

System	Accuracy	MacroF1	Micro F1
TTI	0.3185	0.1920	0.3644
anttip	0.3152	0.1919	0.3038
k-NN	0.2724	0.1486	0.3015

Accuracy で高い性能を得られたが，特に，高い Micro-F1 スコアを得ることができた．

LSHTC3 Wikipedia Large データは，超大規模な階層的分類データであり，2,365,436 訓練データによる階層構造中の 874,219 エッジに対する学習時間は 3,029 分（約 50 時間）であった．テストフェーズにおける 452,167 データの 325,055 カテゴリへの分類時間は 266 分（約 4.4 時間）であった．これらの時間は，ファイル入出力やオーバヘッドを含めたすべての処理時間を含んでいる．このように，高性能な超大規模階層的分類を現実的な処理時間内で実現することに成功した．

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 0 件）

〔学会発表〕（計 5 件）

① Mohammad Golam Sohrab, Makoto Miwa, Yutaka Sasaki, IN-DEDUCTIVE and DAG-TREE Approaches for Large-Scale Extreme Multi-label Hierarchical Text Classification, 17th International Conference on Intelligent Text Processing and Computational Linguistics, 査読有, 2016.4.4, Konya (Turkey).

② Mohammad Golam Sohrab, Makoto Miwa, Yutaka Sasaki, Word Embeddings in Large-Scale Deep Architecture Learning, 第 22 回言語処理学会年次大会, 2016.3.9, B3-3, 東北大学（宮城県・仙台市）.

③ Mohammad Golam Sohrab, Makoto Miwa, Yutaka Sasaki, Centroid-Means-Embedding: An Approach to Infusing Word Embeddings into Features for Text Classification, The 19th Pacific-Asia Conference on Knowledge Discovery and Data Mining, 査読有, 2015.5.21, Ho Chi Minh City (Vietnam).

④佐々木裕, Mohammad Golam Sohrab, 三輪誠：DCASVM を用いた高性能な大規模階層的文書分類，第 21 回言語処理学会年次大会，2015.3.18, D3-4, pp.485-488，京都大学（京

都府・京都市).

⑤ 佐々木裕, Mohammad Golam Sohrab, LSHTC4 のための TTI 文書分類システム, 第 20 回言語処理学会年次大会, 2014.3.18, C1-4, pp.336-339, 北海道大学（北海道・札幌市).

〔図書〕（計 0 件）

〔産業財産権〕
○出願状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
佐々木　裕（SASAKI, Yutaka）
豊田工業大学・工学（系）研究科（研究院）・教授
研究者番号：60395019

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：