Vol.2022-MPS-138 No.36
Vol.2022-BIO-70 No.36
2022/6/28

目標ドメインにおける少量サンプルの近傍データを用いた文書分類器の精度推定手法

田原 英一[†1,a)]　湯浅 晃[†2,b)]

概要：企業への機械学習の導入において，業務システムからの大量データの抽出・加工は一般に大きな工数が必要となる．そのため，仮に大量データを用いてモデルの学習を行った場合，どのくらいの精度が得られるか少量データを用いて推定したい，というニーズが存在する．

少量データで高精度なモデル構築を行う手法として転移学習があるが，目標ドメイン側で転移仮定をモデルに組み込むモデルベース手法や，目標ドメインに合わせて元ドメインの特徴空間を変換する特徴ベース手法においては，目標ドメインのテキストデータが数十件程度と著しく少ない条件下では適用が難しいことが知られている．

そこで本稿では，事例ベース転移学習の一手法として，元ドメインと目標ドメインのサンプルをクラスタリングし，目標ドメインのサンプルが所属するクラスタの元ドメインのサンプルを，疑似的な目標ドメインのサンプルと見立てた上で文書分類器を構築し，精度を推定する手法を提案する．

実験の結果，本手法によって取得したデータを用いて学習した分類器の精度推定値は，ランダムサンプリングによりデータを取得した分類器の精度推定値よりも，真の精度に近い値となることが確認された．これにより目標ドメインの近傍データを使用することは分類器の精度推定において有効であることが確認できた．

キーワード：転移学習，ドメイン適応， 自然言語処理，機械学習

1. はじめに

企業の業務システムへの機械学習の導入において，本格導入前に大量データを用いた精度検証が実施できれば，期待する精度が得られないリスクの低減が可能だが，業務システムからの大量データの抽出・加工は一般に大きな工数が必要となるため難しい．そのため，仮に大量データを用いて学習を行った場合，どのくらいの精度が得られるか少量データを用いて推定したい，というニーズが存在する．

少量データで高精度なモデル構築を行う手法として転移学習があり，関連したドメインの共通項を転移して，目標ドメインの問題をより高精度で解くことが期待されているが，元ドメインから目標ドメインに共通項の知識を転送する際に，様々な観点が存在する．

目標ドメイン側で転移仮定をモデルに組み込むモデルベース手法や，目標ドメインに合わせて元ドメインの特徴空間を変換する特徴ベース手法においては，目標ドメインのデータが数十件程度と著しく少ない条件下では適用が難しいことが知られている．

また，元ドメインの各訓練事例を，目標ドメインとの関連性に応じて，重み付けや取捨選択する事例ベース手法も存在する．元ドメインのサンプリングデータから弱学習器を作成し，目標ドメインのデータの予測精度が高い弱学習器を採用する手法[1]においては，元ドメインからランダム抽出したサンプルを学習データとして用いるが，本稿における目的に対しては適さないと考えられる．本稿の目的は，目標ドメインのテストデータに対して高精度が得られる分類器を構築することではなく，少量の目標ドメインのサンプルをもとに，目標ドメインの真のサンプルを大量に用いた場合の分類器の精度を推定することである．元ドメインからランダム抽出したサンプルは，目標ドメインの真のサンプルとの類似性または関係性が不明であるため，目標ドメインにおいて存在するサンプルとしての妥当性が説明できない．そして，そのようなサンプルを用いた分類器の精度を目標ドメインの真のサンプルを用いた分類器の精度推定値とすることは，その根拠となる学習データが目標ドメインに存在することが説明できない点で妥当性に欠ける．

また，元ドメインと目標ドメインのデータを混ぜて分類器を作成し，目標データを誤分類すると重みを増やし，元ドメインデータを誤分類すると重みを減らす手法[2]も提案されているが，目標ドメインのデータが数十件程度と著しく小さい場合，妥当な分析結果を得るのが難しい．

加えて，テキストデータのデータ拡張では，ランダムに同義語を置換・挿入し，少量のトレーニングセットで，利用可能なすべてのデータと同じ精度を達成したことを報告する研究[3]が存在するが，元ドメインの単語の同義語をランダムに用いてデータ拡張を行った場合，生成されるデータは目標ドメインの真の単語分布にしたがったデータとかけ離れたものとなるため，そのようなデータをもとに構築された分類器の精度を目標ドメインの分類器の精度推定値とすることは妥当性に欠ける．

そこで本稿では，目標ドメインのデータが数十件程度と少量の条件下で，仮に目標ドメインのデータが大量に得られるとした場合の分類器の精度推定値を求めるという目的に対して，推定値の根拠となる分類器の学習データの選定においてより妥当な説明が可能な手法を検討する．

2. 提案法

元ドメインと目標ドメインのテキストデータをそれぞれ単語埋め込み手法等を用いて文書ベクトルに変換する．

[†1] 株式会社 NTT データ 第三金融事業本部 戦略ビジネス本部
[†2] 株式会社 NTT データ 技術革新統括本部 システム技術本部
[a)] Eiichi.Tahara@nttdata.com [b)] Akira.Yuasa@nttdata.com

Vol.2022-MPS-138 No.36
Vol.2022-BIO-70 No.36
2022/6/28

次に，元ドメインと目標ドメインの共通的な特徴空間を特定するために，元ドメインと目標ドメインのサンプルをクラスタリングする．各クラスタに所属する目標ドメインのサンプル数が，目標ドメインの全サンプル数をクラスタ数で除算した値以上になる場合，当該クラスタに所属する元ドメインのデータを疑似的な目標ドメインのサンプルと見做す．この理由は，当該クラスタのサンプルが一定以上含まれる特徴空間領域は，目標ドメインの全量データを取得した際に十分なサンプル数が得られる領域と仮定するためである．

次に，疑似的な目標ドメインのサンプルを学習データとして文書分類器を構築する．ここで得られた精度を，目標ドメインの全量データを用いたと仮定した場合の文書分類器の精度推定値とする．

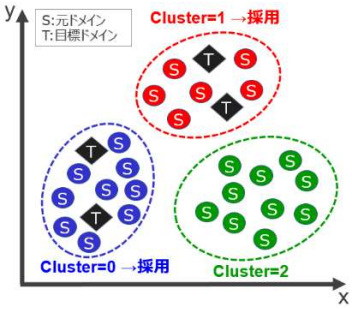

図 1 提案手法における目標ドメインの近傍データ選定方法
目標ドメインのサンプル(T)数が，目標ドメインの全サンプル数をクラスタ数で除算した値以上となるクラスタ(Cluster=0 と Cluster=1)について，所属する元ドメインのサンプル(S)を，目標ドメインのサンプルと見立てる．

3. 実験

3.1 実験データ

本実験では，筆者所属組織におけるマーケティング用のテキストデータを利用した．元ドメインと目標ドメインの内容の主な違いは関係する顧客層の異なりである．

サンプル数について，元ドメインのテキストデータが 5000 件，目標ドメインのテキストデータが 42 件で，共にランダムサンプリングにより抽出した．各サンプルは，自由記述のテキストデータであり，自社商品の購入有無の正例負例の教師ラベルが付与されている．1 件あたりの平均単語数は，元ドメイン側が 6.9 単語，目標ドメイン側が 9.4 単語となった．

3.2 実験方法

まずテキストデータの文書ベクトル化においては教師なし埋め込み手法において比較的高性能が得られている Universal Sentence Encoder[4]を採用した．

次に，元ドメインと目標ドメインの文書ベクトルを，K-means を用いたクラスタリングによって 10 分割した．

目標ドメインのサンプル数が 4.2 件以上所属するクラスタを，疑似的な目標ドメインの領域と見立て，領域内の元ドメインの文書ベクトルを抽出した．

次に，本手法で抽出したサンプルを，train データと test データに分割し，train データに対して 5 分割での交差検証により分類器の学習および精度評価を実施した．ここで分類器には，比較的単純なロジスティック回帰を採用した．

評価方法について，まず本手法によって取得したデータを用いて学習した分類器の精度推定値と，ランダムサンプリングによりデータを取得した分類器の精度推定値との間に有意差があるかどうかを確認するための評価を行った．本手法で抽出した 3550 件のサンプルと，元ドメインからランダムサンプリングした同件数の 3550 件のサンプルについて，それぞれ train データと test データの分割時の乱数 seed を 30 回変更して分類器を作成し，対応なし t 検定を用いて AUC に有意差があるかどうかを確認した．

次に，実際の目標ドメインの全量データ 5000 件で分類器を作成することによって，目標ドメインのサンプルを用いた分類器の真の精度を求め，疑似データを用いた本手法の精度推定値が，真の値に近い値となるか検証した．

3.3 評価

分類器の AUC の平均値は，ランダムサンプリングでは 0.773，提案手法では 0.791 となった．また，t 検定の結果，p 値が 0.003 となり，両者の母平均について有意差があることを確認した．

また，目標ドメインの 5000 件の全量データでの分類器の AUC は，0.812 となり，提案手法による推定値が真の値により近い値をとることを確認した．

さらに，提案手法と目標ドメインの全量データのロジスティック回帰の回帰係数を比較したところ，上位 3 位までが同じ特徴量であることを確認した．これにより，提案手法により作成されたモデルは，目標ドメインの実際のサンプルを用いたモデルと，大局的な振る舞いがある程度似ているとの解釈が可能であり，一定の説明力を備えているといえる．

4. おわりに

今回の検証対象とした 1 つの検証データセットにおいては，目標ドメインの近傍データを使用することは分類器の精度推定において有効であることが確認できた．今後は本手法の妥当性を確認するため，様々なデータセット，およびタスクを用いた実験を実施する必要がある．

参考文献

[1] 神嶌敏弘, 濱崎雅弘, and 赤穂昭太郎. “飼いならし 飼育・野生混在データからの学習,” 人工知能学会全国大会論文集 第 22 回 (2008).
[2] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu. “Boosting for transfer learning,” In Proc. of The 24th Int’l Conf. on Machine Learning, pp. 193–200 (2007).
[3] Wei, Jason, and Kai Zou. “Eda: Easy data augmentation techniques for boosting performance on text classification tasks,” arXiv preprint arXiv:1901.11196 (2019).
[4] Cer, Daniel, et al. “Universal sentence encoder,” arXiv preprint arXiv:1803.11175 (2018).