最小記述長原理に基づく転移能動学習による多クラス分類

白井　匡人† 　劉　健全††,††† 　邵　浩†††† 　三浦　孝夫†††††

† 島根大学大学院総合理工学研究科情報システム学領域　〒 690-8504 島根県松江市西川津町 1060
†† 日本電気株式会社 システムプラットフォーム研究所　〒 211-8666 神奈川県川崎市中原区下沼部 1753
††† 法政大学大学院 理工学研究科　〒 184-8584 東京都小金井市梶野町 3-7-2
†††† 上海対外経貿大学　上海市古北路 620 号
††††† 法政大学 理工学部創生科学科　〒 184-8584 東京都小金井市梶野町 3-7-2
E-mail: †shirai@cis.shimane-u.ac.jp, ††j-liu@ct.jp.nec.com, †††shaohao@suibe.edu.cn, ††††miurat@hosei.ac.jp

あらまし　本研究では，最小記述長原理に基づく転移能動学習を用いた文書分類手法を提案する．転移学習を用いた文書分類では，対象領域の未知文書を分類するために情報源領域から得られるクラス情報を転移することで分類を行う．しかし，各領域の関連性が低い場合精度が悪化する可能性がある．このため，誤った知識の転移を防ぐために各領域のクラスの対応付けを行う必要がある．本研究では，最小記述長原理に基づき対象領域のクラスに関連する情報源領域のクラスを選択する．また，能動学習により転移した知識を更新することで分類精度を改善する．

キーワード　転移能動学習，最小記述長原理，文書分類

1. 前　書　き

近年，インターネットの発達から大量のデータを容易に入手できるようになっている．ニュース記事やマイクロブログといった大規模な文書集合は，様々な情報を含む情報源として注目されている．しかし，これらの文書集合は文書数が膨大となるため，大半の情報が利用されないまま流れ去っている．このような文書集合ではデータ量が膨大となるため，人手で分類を行うことは困難であり，自動的に分類を行う手法が必要不可欠である．

文書をクラスごとに分類する文書分類では，政治，スポーツ，IT といったクラスは文書の集合として表現され，文書の様々な特徴からクラスが未知の文書を分類する．各クラスの特徴は，ラベル付き文書を基にクラスを表す特徴を学習する．このため，各クラスを区別する特徴を得るには十分なラベル付き文書が必要となる．しかし，文書のラベル付けは，人手によって行うため，コストが膨大となる．転移学習は，対象領域の解析に情報源領域から得られた情報を利用する．このため，対象領域から十分な学習データが得られない場合でも有効に解析が行える．転移学習を用いた文書分類は，情報源領域から得られるクラス情報を基に対象領域のラベル無し文書を分類する．対象領域に少量の学習データが存在する場合，情報源領域と対象領域のラベル付き文書を基にクラスの対応付けを行う．2 つの領域のラベル付き文書から各クラスの特徴を学習することで対象領域のラベル無し文書を分類する．このような転移学習の設定は，帰納転移学習と呼ばれる．帰納転移学習は，対象領域の解析結果の改善に有効であるが，各領域の関連性が低い場合，誤った情報が転移されることで精度が悪化する [1]．例えば，経済クラスの特徴を学習する場合，出現する単語が類似する市場クラスの頻度情報は役立つ可能性がある．しかし，関連性の低いスポーツクラスの単語の頻度情報は，経済クラスとは大きく異なるため役立てることができない．このため，対象領域の各クラスに対して適切な情報源領域を選択する必要がある．また，対象領域の解析を高精度化する目的で転移能動学習が提案されている [6] [7] [8] [9]．転移能動学習は，情報源領域から得られた知識をラベル無しの文書を用いて能動学習することでパラメータを更新する．

本研究では，転移能動学習に基づく文書分類を行うため，3 つの問題を論じる．第 1 の問題は，対象領域の各クラスに対して如何に適切な情報源領域を選択するか，第 2 の問題は，対象領域に適応するために必要な文書を如何に選択するか，第 3 の問題は，情報源領域と対象領域のラベル付き文書から如何にラベル無し文書を分類するかである．本研究は，適切に転移学習を行うために最小記述長原理に基づく情報源領域の選択手法を提案する．最小記述長原理は，モデルのパラメータと文書集合に対する記述長を基に対象となる集合に適したモデルを選択する．最小記述長原理を用いる理由として，パラメータに依存しない，ノイズに強い，過学習を防ぐといった利点がある．提案手法は，対象領域のクラスに関連する情報源領域のクラスを選択することで転移学習を行う．対象領域のラベル無し文書を分類するため，Nonnegative Matrix Tri-Factorization(NMTF) を用いて情報源領域から得られるクラス情報を伝搬する．また，能動学習により対象領域のラベル無し文書内で現在のラベル付き文書によって特徴付けられない文書を検出し，ラベル付き文書に加えることで分類精度を改善する．

第 2 章では転移能動学習について述べ，第 3 章では最小記述長原理について述べる．第 4 章では提案手法について述べる．第 5 章では実験により有効性を示す．第 6 章で結論とする．

2. 転移能動学習

2.1 転移能動学習

転移能動学習は，少ないラベル付きデータの基で解析を行う

ことを目的とする．ここでは，転移学習と能動学習を適応的に連携することで，各学習手法の欠点を補う．帰納転移学習では，情報源領域と対象領域のクラスと文書の同時確率 $p(x^s, y^s)$ と $p(x^t, y^t)$ が近づくように学習を行う．ここで，対象領域の学習データは少量であるため，能動学習により対象領域に適応させる必要がある．

Rai らは，情報源とラベル無しの対称領域から初期のモデルを構築する手法を提案している [10]．この問題設定では，情報源の分布が対象領域と同一でなければならず，領域間の関連性が低い場合を想定していない．Zhu らは，シグモイド関数に基づきモデルの重みを更新する手法を提案している [11]．この手法は，情報源のデータを対象領域の学習データに直接用いている．これらの手法では，各領域の関連性が低い場合に誤った情報が伝搬され，精度が悪化する可能性がある．Chattopadhyay らは，単一の凸最適化問題を解くことによって転移学習と能動学習を同時に行う枠組みを提案している [12]。 この手法は，初期の学習にラベル付けされたデータが利用できない状況に対応することができる．しかし，能動学習を行う際に情報源から学習データを 1 つずつ選択するため計算量が膨大になるという欠点がある．

能動学習は，機械自身がモデルを改善するために有効なデータを検出し，学習に用いる手法である．この学習手法は，主に静的な文書集合を対象とするプールベースの手法 [5] と動的な文書集合を対象とするストリームベースの手法に分けられる．Bouguelia らは，ストリームベースの能動学習を用いることでストリーム中で新たな特徴を学習し分類を行う手法を提案している [3]．ストリームベースの能動学習ではテスト文書が到着するたびにラベル付けを行うか判断する．一般的な能動学習では，データの検出は自動的に行うが，検出されたデータを人手によりラベル付けする．人手によるコストを減少させるためには，新たなラベル付き文書を必要としないことが望ましい．しかし，ラベル無しの文書だけを用いたモデルの更新は，誤った特徴を学習することで性能が悪化する可能性がある．本研究では，プールベースの能動学習を用いて各クラスの特徴を更新する．

2.2 問題設定

本稿は，転移能動学習を用いて対象領域のクラスが未知の文書のクラス y を推定する．各領域は，ラベル付きの情報源領域 S とラベル無しの対象領域 $U = \{u_1, u_2, ..., u_j\}$ と少量のラベル付きの対象領域 $L = \{l_1, l_2, ..., l_i\}$ とする．文書分類を行うため，対象領域のラベル付き文書集合 L と情報源領域 S を用いてクラスごとに単語の確率分布を学習する．この確率分布を基に，最小記述長原理を用いて情報源領域 S に存在する各クラス $y_1^s, y_2^s, ..., y_k^s$ と対象領域に存在するクラス $y_t^t, y_2^t, ..., y_c^t$ の対応付けを行う．また，学習データが少量しか存在しない対象領域に適応するため，ラベル無しの文書集合 U を用いて能動学習を行う．ここでは，プールベースの能動学習によりラベル無し文書 u を選択し，新たにラベル付けすることで対象領域のラベル付き文書集合 L に追加する．提案手法は，情報源領域と対象領域のラベル付き文書を基にクラス情報を伝搬することでラベル無しの文書集合 U を各クラスに分類する．

3. 最小記述長原理

最小記述長原理は，データを符号化する記述長に基づき対象となる集合に適したモデルを選択する．データ集合に対する記述長は，モデル自身のパラメータとデータへの適合度によって求まる [13], [14]．データ集合 D に対する最良のモデル h は，以下の式より求まる．

$$h_{best} = arg \min_h(-logP(h) - logP(D|h)) \tag{1}$$

ここで，$P(h)$ はモデルが発生する確率を表し，$P(D|h)$ はモデル h の基での文書集合 D が発生する確率を表す．符号化理論によれば [15]，最も効率的に h を符号化するコード列の長さは $-logP(h)$ となる．同様に $-logP(D|h)$ は，h の基での D を符号化するコード列の長さを表す．モデル h の符号化の長さは以下の式より求まる [2]．

$$-logP(h) = \Lambda(m, 0) + \sum_{i=1}^{m} |\Theta(M, m) + \Lambda(v, 0)| \tag{2}$$

ここで，M はデータの属性の数，m は非零となる属性の数，v は各属性の重みを表す．Θ，Λ は以下の式より求める．

$$\Theta(a, b) \equiv log(a) + log \begin{pmatrix} a \\ b \end{pmatrix} \tag{3}$$

$$\Lambda(a, b) \equiv log \left| 3 - (\frac{1}{2}^b) \right| + |b - a| \tag{4}$$

ここで，a は 2 進数文字列の長さ，b は 2 進数文字列中の 1 の数を表し，文字列が全て 0 の場合を考慮しないため $b > 0$ である．例えば，2 進数列”000010000000100”の記述長は，文字列の長さ $\log 16 = 4$ bits とその位置情報 $\log \begin{pmatrix} 15 \\ 2 \end{pmatrix} = 6.7$bits の合計 10.7 bits で表される．

次に，h の基での文書集合 D の符号化の長さは以下の式より求める．

$$-logP(D|h) = \Theta(|D|, \omega(h, D)) \tag{5}$$

ここで，$|D|$ は文書数，$\omega(h, D)$ はモデル h を用いて D を分類したときの誤分類した数を表す．例えば，文書集合 D に 6 つの文書がありその真のクラスが (1,0,0,1,1,0) であり，分類結果が (1,0,0,0,1,0) であるとき，送信側は受信側へ 4 番目の文書が誤分類したという情報を送れば，長さ 6 の 2 進数文字列を送信したことに相当する．また，この記述長が文書の分類結果をそのまま送信する場合である 6bits と比較して長いかどうかで文書集合を符号化するために適していないモデルを除外できる．

これらのことから，最小記述長原理は，モデルのパラメータの符号化長と文書集合に対する分類精度を基に最適なモデルを選択する．記述長 CL は以下の式より求まる．

$$CL = \Lambda(m,0) + \sum_{i=1}^{m} |\Theta(M,m) + \Lambda(v,0)| + \Theta(|D|, \omega(h,D)) \quad (6)$$

本研究は，この最小記述長原理に基づき対象領域の各クラスに対する記述長が最小となる情報源領域のクラスを抽出する．

4. Nonnegative Matrix Tri-Factorization

NMTF は，特徴空間と文書空間を共クラスタリングすることによって，高次元の行列を低次元の 3 つの行列に分解するアルゴリズムであり，文書分類やクラスタリングに用いられている [16] [17] [18]．NMTF では，特徴-文書空間の行列 $X \in R^{m \times n}$ は低次元の 3 つの行列 FAG に分解される．

$$X \approx FAG^T \quad (7)$$

行列 $F \in R^{m \times p}$ は，特徴と特徴クラスタ間の関係を表す．m は特徴数であり，p は特徴クラスタの数である．行列 $A \in R^{p \times c}$ は，特徴クラスタと文書クラスタ間の関連を表す．c は文書クラスタの数である．行列 $G \in R^{c \times n}$ は，文書クラスタと文書間の関連を表す．n は文書数である．各行列は元行列とのフロベニウスノルムが最小となるように推定する．

$$min\|X - FAG^T\| \quad (8)$$

ここで，$F \geqq 0, A \geqq 0, G \geqq 0$ である．NMTF では，文書-文書クラスタ空間の行列 G は文書が属するクラスタの情報を表す．文書クラスタ数 c をクラス数に変更し，文書が属するクラスの要素を 1，それ以外が 0 となる教師データを与えることで文書分類に対応できる．

NMTF に基づく転移学習では，情報源領域の元行列 X_s，対象領域の元行列 X_t を同時に分解し，情報源領域の持つラベルを対象領域に伝播する [20]．

$$min\|X_s - F_s A_s G_s^T\| + \|X_t - F_t A_t G_t^T\|$$
$$F_s = [F^1, F_s^2]\ ,\ A_s = [A^1, A_s^2]^T$$
$$F_t = [F^1, F_t^2]\ ,\ A_t = [A^1, A_t^2]^T \quad (9)$$

ここで，F_s，A_s，G_s は情報源領域，F_t，A_t，G_t は対象領域の分解後の各行列である．式中の上付き文字 T は転置行列を表す．特徴-特徴クラスタ空間の行列 F は，領域間の共通の特徴を表す F^1 と領域独自の特徴を表す行列 F_s^2，F_t^2 を要素に持つ．特徴クラスタ-文書クラスタ空間の行列 A は共通の特徴クラスタと文書クラスタ間の関係を表す行列 A^1 と領域独自の特徴クラスタと文書クラスタ間の関係を表す行列 A_s^2，A_t^2 を要素に持つ．情報源領域のクラス情報 G_s は学習文書から得られる各文書が属するクラスを示す．ベクトルの要素は該当するクラスに属している場合 1，属さない場合 0 の値をとる．ラベル無し文書のクラスは，学習した対象領域のクラス情報 G_t を基に決定する．

5. 提 案 手 法

提案手法は，対象領域のラベル無し文書を分類するため対象領域に関連した情報源領域から得られる情報を利用する．ここでは，誤った情報の転移を防ぐために，最小記述長原理により領域間の関連性を測定する．コサイン類似度や KL 情報量などの分布間の距離や情報量を測る類似度計算手法は，領域間の分布のパラメータにのみ着目する．このため，対象領域のラベル付き文書が少ない影響により誤った情報源領域を選択する可能性がある．提案手法は，最小記述長原理に基づき情報源領域の選択を行うことで，分布の類似と分類精度の 2 つを考慮して関連性を捉えることができる．次に選択された情報源領域と対象領域のラベル付き文書を基に NMTF を用いてラベル無し文書のクラスを推定する．最後に対象領域の文書を用いて能動学習を行う．ここでは，プールベースの能動学習を用いる．対象領域のラベル無しの文書集合から各クラスを特徴付ける文書を抽出し，ラベル付き文書集合に加える．

対象領域の各クラス y^t に対して最小記述長原理を基に情報源領域 s を以下の式より選択する．

$$|CL_s - CL_t| = \sum_{i=1}^{m} \big| |\Lambda(v_s,0) - \Lambda(v_t,0)| \big| + \Theta(|D|, \omega(y,D)) \quad (10)$$

ここで，$|\Lambda(v_s,0) - \Lambda(v_t,0)|$ は，クラス間のパラメータの差を表し，$\Theta(|D|, \omega(y,D))$ は情報源領域の確率分布に基づく分類精度を表す．これにより領域間の分布の差と分類性能を考慮して情報源領域の選択ができる．

次に対象領域のラベル無し文書を用いて能動学習を行う．能動学習では，対象領域の学習データを基に，尤度が最大となるクラスと 2 番目になるクラスの尤度差から選択するラベル無し文書 x_i を決定する．

$$Diff(x_i) = |logP(x_i|y^1) - logP(x_i|y^2)| \quad (11)$$

ここで，$logP(x_i|y^1)$ は，尤度が最大となるクラスの基で文書 x_i を生成する尤度，$logP(x_i|y^2)$ は，尤度が 2 番目となるクラスの基で x_i を生成する尤度である．対象領域の全てのラベル無し文書に対して尤度の差を求め，対象領域の学習データによって特徴付けられない文書を検出し，ラベル付き文書に加える．

提案手法は，NMTF により情報源領域と対象領域の文書集合を同時に分解することで対象領域のラベル無し文書を分類する．ここでは，選択された情報源領域のラベル付き文書 $S = \{s_1, s_2, ..., s_{|S|}\}$ を NMTF における元行列 X_s とし，対象領域のラベル付き文書と能動学習により選択した文書 $L = \{l_1, l_2, .., l_{|L|}\}$ とラベル無し文書 $U = \{u_1, u_2, ..., u_{|U|}\}$ を合わせて元行列 X_t とする．対象領域のラベル無し文書のクラスは，クラス情報を表す行列 G_t より各文書のベクトルの要素中で最も値の高くなる列に該当するクラスを推定結果とする．

6. 実　　験

実験では Reuters Corpus(RCV1) を用いて多クラス分類を行う．転移学習の有効性を確認するため，対象領域の学習データのみ使用した場合と比較する．また，最小記述長原理の有効性を確認するため，確率分布の類似のみを考慮して転移学習行った場合と比較する．

6.1 実験準備

実験に用いる Reuters Corpus は 1996 年 8 月 20 日から 1997 年 8 月 19 日までの 1 年分のニュース記事であり，1 つの記事に 128 種類からなるラベルが複数付いている．Reuter Corpus はニュース記事であることから，クラスで使われる固有名詞やそれに関連した単語などクラス固有の特徴を有する．本実験では，各文書に付与されている複数のラベルのセットを 1 つのクラスと見なし，頻度が上位となる 50 クラスを情報源領域に用いる．続く頻度が上位となる 10 個のクラスを対象領域に用いる．表 1，2 にクラスとなるラベルの組み合わせを示す．表 3 には，メインカテゴリのコーパスの識別子とクラス名を示す．他の識別子は先頭の文字に該当するメインカテゴリのサブカテゴリとなる．前処理として，実験データ中の不要語は取り除く．算用数字は*に変換し，数字列の長さにのみ着目する．文字列は全て小文字に変換する．実験データ中で出現回数が 20 以下の単語を全て除外する．単語の種類数は 11913 である．実験に用いる文書数は，学習データとして情報源領域の各クラスに 400 文書と対象領域の各クラスに 25 文書を用いる．テスト文書は，対象領域の各クラスに 500 文書の計 5000 文書とする．また，提案手法は能動学習により 50 文書を追加の学習データとして用いる．

比較手法には転移学習を行わず，対象領域の学習データのみを用いて NMTF によって分類を行う手法と確率分布の類似度によって情報源領域を選択する手法を用いる．確率分布の類似度には，コサイン類似度を用いる．コサイン類似度は以下の式より求める．

$$cos(P,Q) = \frac{\sum_i^V p_i \cdot q_i}{\sqrt{\sum_i^V p_i^2} \cdot \sqrt{\sum_i^V q_i^2}} \tag{12}$$

ここで，P,Q は各クラスに対応し，p_i，q_i は i 番目の単語の出現確率である．V は単語の種類数であるコサイン類似度は 0 〜1 の値であり，類似度が高いほど 1 に近い値になる．対象領域のクラスに対して最も類似度が最も高くなる情報源領域を選択し，転移学習を行う．

6.2 評価方法

実験の評価には f 値を用いる．f 値は再現率と適合率の調和平均であり，再現率は実際に正であるもののうち，正であると予測されたものの割合，適合率は正と予測したデータのうち，実際に正であるものの割合である．各クラスの f 値を以下の式で求める．

$$R_i = \frac{a_i}{a_i + c_i} \tag{13}$$

$$P_i = \frac{a_i}{a_i + b_i} \tag{14}$$

a_i は推定結果が正である数，c_i は正であるが負と推定された数，b_i は正であると推定した中で正解が負である数である．この 2 つの式の調和平均である f 値を次のように定義する．

$$f_i = \frac{2 \times P_i \times R_i}{P_i + R_i}$$

6.3 実験結果

表 4，5，6 に実験結果を示す．”NMTF+コサイン類似度”は，情報源領域の選択にコサイン類似度を用いる場合を示す．”NMTF+最小記述長”は提案手法と同様に最小記述長によって情報源領域を選択するが，能動学習を行わない場合を示す．表 4 より各再現率は,NMTF で 0.700，コサイン類似度で 0.729，最小記述長で 0.844，提案手法で 0.761 となる．表 5 より各適合率は，NMTF で 0.713，コサイン類似度で 0.743，最小記述長で 0.757，提案手法で 0.770 となる．表 6 より各 f 値は，NMTF で 0.706，コサイン類似度で 0.736，最小記述長で 0.751，提案手法で 0.766 となっている．全体の再現率・適合率・f 値は，いずれも転移能動学習を行う提案手法が最も高い値を示している．また，各クラスの f 値は，10 クラス中 7 クラスで提案手法が最も高精度となっている．

6.4 考　　察

表 6 より，提案手法の f 値は対象領域の学習データのみを使用する NMTF と比較して+6.0%向上している．対象領域のラベル付き文書は少量であるため，転移学習により情報源領域の学習データを利用し，ラベル情報を伝搬することで分類精度が改善する．また，最小記述長原理の有効性として，情報源領域を最小記述長によって選択する場合の f 値は，コサイン類似度と比較して+1.5%高い値を示している．コサイン類似度による情報源領域の選択では，確率分布の類似のみを考慮しているため，適切な情報源領域を選択できていない．実際に表 7 より，情報源領域として選択するクラスはコサイン類似度と最小記述長原理で異なっている．能動学習の効果として，対象領域のラベル無し文書から 50 文書を選択してラベル付けを行い，学習データとして使用することで分類精度が+1.5%改善している．表 8 より，能動学習を用いずにラベル無し文書の出現順序順に学習データとして追加した場合と比較すると，300 文書を追加した場合においても 50 文書しか使用しない提案手法の方が+0.02%高い値を示している．提案手法は，能動学習を使用することで少量の追加の学習データによって効率よく分類精度を改善できる．

これらのことから，提案手法は少量の学習データしか存在しない文書集合の多クラス分類に有効である．

7. 結　　論

本研究では，多クラス分類を行うため最小記述長原理に基づく転移能動学習の方式について論じた．提案手法を用いて分類を行った結果，提案手法は，転移能動学習を行うことで f 値が 6.0%改善する．これにより，提案手法の有効性を示した．

C15 C152 CCAT	GCAT GDIP	E11 ECAT
C15 C151 CCAT	M14 M142 MCAT	E51 E512 ECAT
M11 MCAT	C41 C411 CCAT	M12 M13 M131 MCAT
M14 M141 MCAT	C24 CCAT	GCAT GDEF GDIP
GCAT GSPO	GCAT GPOL GVOTE	E13 E131 ECAT
GCAT	C17 C171 CCAT	GCAT GDIP GPOL
C18 C181 CCAT	C33 CCAT	C17 C174 CCAT E21 E212 ECAT
C15 C151 C1511 CCAT	GCAT GCRIM	C15 C152 C18 C181 CCAT
M13 M131 MCAT	C11 CCAT	GCAT GCRIM GPOL
E21 E212 ECAT	GCAT GDIP GVIO	C11 C24 CCAT
M14 M143 MCAT	C42 CCAT E41 ECAT GCAT GJOB	E12 ECAT M13 M132 MCAT
M12 MCAT	C21 CCAT	C17 C171 C18 C181 CCAT
M13 M132 MCAT	E71 ECAT	E21 E211 ECAT
C31 CCAT	E21 E211 ECAT GCAT GPOL	GCAT GCRIM GVIO
GCAT GPOL	GCAT GPOL GVIO	C21 C24 CCAT
GCAT GVIO	C12 CCAT GCAT GCRIM	M11
C17 C172 CCAT	M14 MCAT	

表 1　情報源領域に用いるクラス

E12 ECAT M13 M131 MCAT
C13 C21 CCAT
E12 E121 ECAT
G15 GCAT
C31 C312 CCAT M14 M141 MCAT
GCAT GPOL GPRO
C18 C183 CCAT
E51 E513 ECAT
C18 C181 C183 CCAT
E41 E411 ECAT GCAT GJOB

表 2　対象領域に用いるクラス

8. 謝　　辞

本研究は NSFC(No.61603240), HSSF(No.13YJC630126) の助成を受けたものです。

Identifier	Corpus Labels
CCAT	CORPORATE/INDUSTRIAL
GCAT	GOVERNMENT/SOCIAL
MCAT	MARKETS
ECAT	ECONOMICS

表 3　ラベル識別子

文　　献

[1] Rosenstein, M.T., Marx, Z., Kaelbling, L.P. and Dietterich, T.G.: To Transfer or Not To Transfer, In NIPS'05 Workshop on Transfer Learning, volume 898, 2005

[2] Shao, H., Tong, B. and Suzuki, E.: Feature-based inductive transfer learning through minimum encoding, In Proceeding of the SIAM International Conference on Data Mining (SDM) 2011, pp. 259-270, 2011

[3] Bouguelia, M., Belaid, Y., Belaid, A.: A Stream-Based Semi-Supervised Active Learning Approach for Document Classification, In Proceeding of the 12th International Conference on Document Analysis and Recognition (ICDAR), 2013

[4] Dagan, I., and Engelson, S., P.: Committee-based sampling for training probabilistic classifiers, In Proceeding of the 23rd International Conference on Machine Learning (ICML), pp. 150-157, 2006

[5] McCallum, A., K. and Nigam, K.: Employing EM and Pool-Based Active Learning for Text Classification, In Proceeding of the International Conference on Machine Learning (ICML), pp. 359-367, 1998

[6] Shao, H., Tong, B. and Suzuki, E.: Query by committee in

	NMTF	NMTF+コサイン類似度	NMTF+最小記述長	転移能動学習
E12 ECAT M13 M131 MCAT	0.796	0.838	0.860	**0.868**
C13 C21 CCAT	0.654	0.600	0.598	**0.706**
E12 E121 ECAT	0.570	**0.612**	0.574	0.521
G15 GCAT	0.980	0.976	0.976	**0.982**
C31 C312 CCAT M14 M141 MCAT	0.736	0.770	0.782	**0.812**
GCAT GPOL GPRO	0.960	0.964	**0.966**	0.962
C18 C183 CCAT	0.438	0.544	0.584	**0.703**
E51 E513 ECAT	0.640	0.722	**0.878**	0.867
C18 C181 C183 CCAT	0.498	**0.522**	0.498	0.373
E41 E411 ECAT GCAT GJOB	0.732	0.746	0.738	**0.812**
全体	0.700	0.729	0.745	**0.761**

表 4　再　現　率

	NMTF	NMTF+コサイン類似度	NMTF+最小記述長	転移能動学習
E12 ECAT M13 M131 MCAT	0.715	**0.716**	0.708	0.673
C13 C21 CCAT	0.597	0.806	**0.810**	0.790
E12 E121 ECAT	0.814	0.838	0.852	**0.899**
G15 GCAT	0.909	0.900	0.910	**0.935**
C31 C312 CCAT M14 M141 MCAT	0.581	0.586	0.667	**0.726**
GCAT GPOL GPRO	0.854	0.868	0.869	**0.887**
C18 C183 CCAT	0.466	0.495	0.500	**0.520**
E51 E513 ECAT	0.816	0.811	0.813	**0.836**
C18 C181 C183 CCAT	0.450	0.506	0.511	**0.546**
E41 E411 ECAT GCAT GJOB	0.924	0.903	**0.927**	0.892
全体	0.713	0.743	0.757	**0.770**

表 5　適　合　率

	NMTF	NMTF+コサイン類似度	NMTF+最小記述長	転移能動学習
E12 ECAT M13 M131 MCAT	0.753	0.772	**0.777**	0.759
C13 C21 CCAT	0.624	0.688	0.688	**0.746**
E12 E121 ECAT	0.671	**0.708**	0.686	0.660
G15 GCAT	0.943	0.937	0.942	**0.958**
C31 C312 CCAT M14 M141 MCAT	0.650	0.666	0.720	**0.766**
GCAT GPOL GPRO	0.904	0.914	0.915	**0.923**
C18 C183 CCAT	0.452	0.518	0.539	**0.598**
E51 E513 ECAT	0.717	0.764	0.844	**0.851**
C18 C181 C183 CCAT	0.473	**0.514**	0.505	0.443
E41 E411 ECAT GCAT GJOB	0.817	0.817	0.822	**0.850**
全体	0.706	0.736	0.751	**0.766**

表 6　f　　値

a heterogeneous environment, In Proceeding of the 8th International Conference on Advanced Data Mining and Applications (ADMA) 2012, pp. 186C-198, 2012

[7] Shi, X., Fan, W., and Ren, J.: Actively transfer domain knowledge, In Proceeding of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD) 2008, pp. 342-357, 2008

[8] Reichart, R., Tomanek, K. and Hahn, U.: Multi-task active learning for linguistic annotations, In Annual Meeting of the Association for Computational Linguistics (ACL) 2008, pp. 861-869, 2008

[9] Raj, S., Ghosh, J. and Crawford, M. M.: An active learning approach to knowledge transfer for hyperspectral data analysis, In Proceeding of the IEEE International Conference on Geoscience and Remote Sensing Symposium (IGARSS) 2006, pp. 541-544, 2006

[10] Rai, P., Saha, A., Daume, H. and Venkatasubramanian, S.: Domain adaptation meets active learning, In Proceeding of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing, pp. 27-32, 2010

[11] Zhu, Z., Zhu, X., Ye, Y., Guo, Y.F. and Xue, X.: Transfer active learning, In Proceeding of the 20th International Conference on Information and Knowledge Management (CIKM) 2011, pp. 2169-2172, 2011

[12] Chattopadhyay, R., Fan, W., Davidson, I., Panchanathan, S. and Ye, J.: Joint transfer and batch-mode active learning, In Proceeding of the 30th International Conference on Machine Learning (ICML), pp. 253-261, 2013

[13] Wallace, C. and Patrick, J.: Coding decision trees, Journal of Machine Learning, vol. 11, no. 1, pp. 7-22, 1993.

対象領域のクラス	提案手法	コサイン類似度
E12 ECAT M13 M131 MCAT	GCAT	M13 M131 MCAT
C13 C21 CCAT	C12 CCAT GCAT GCRIM	C11 CCAT
E12 E121 ECAT	E11 ECAT	C31 CCAT
G15 GCAT	GCAT GCRIM GPOL	GCAT
C31 C312 CCAT M14 M141 MCAT	C21 C24 CCAT	M14 M141 MCAT
GCAT GPOL GPRO	GCAT GPOL	GCAT GPOL
C18 C183 CCAT	C11 C24 CCAT	C11 C24 CCAT
E51 E513 ECAT	E21 E211 ECAT	C15 C151 CCAT
C18 C181 C183 CCAT	C11 CCAT	C18 C181 CCAT
E41 E411 ECAT GCAT GJOB	E51 E512 ECAT	E11 ECAT

表 7　情報源領域として選択されたクラス

	追加 50	追加 100	追加 200	追加 300	転移能動学習 (追加 50)
再現率	0.747	0.757	0.756	0.758	0.761
適合率	0.758	0.764	0.769	0.770	0.770
f 値	0.753	0.761	0.762	0.764	0.766

表 8　学習データ数による f 値の変化

[14] Quinlan, J. R. and Rivest, R. L.: Inferring decision trees using the minimum description length principle, Information and Computation, vol. 80, no. 3, pp. 227-248, 1989.

[15] Shannon, C.: A mathematical theory of communication, Bell System Technical Journal, vol. 27, pp. 379-423, 1948.

[16] Ding, C., Li, T., Peng, W. and Park, H.: Orthogonal Nonnegative Matrix Tri-Factorizations for Clustering, In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD), pp. 126-135, 2006

[17] Wang, H., Nie, F., Huang, H. and Makedon, F.: Fast Nonnegative Matrix Tri-Factorization for Large-Scale Data Co-Clustering, In Proceedings of the Twenty-Second international joint conference on Artificial Intelligence (IJCAI), pp. 1553-1558, 2011

[18] Shirai, M., Liu, J. and Miura, T.: Transfer Learning using Latent Domain for Document Stream Classification, In Proceedings of the Second IEEE International Conference on Multimedia Big Data (BigMM), pp. 82-88, 2016

[19] Long, M., Wang, J., Ding, G., Cheng, W., Zhang, X. and Wang, W.: Dual Transfer Learning, In Proceeding of the SIAM International Conference on Data Mining (SDM), pp. 540-551, 2012

[20] Tan, B., Song, Y., Zhong, E. and Yang, Q.: Transitive Transfer Learning, In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD), 2015

[21] Lewis, D.D., et al.: RCV1 (Reuters Corpus Volume 1), 2004, www.daviddlewis.com/resources/testcollections/rcv1/

[22] Fang, M., Yin, J., Zhu, X.: Knowledge Transfer for Multi-labeler Active Learning, In Proceeding of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD), pp. 23-27, 2013