A-006

学習ベクトル量子化を用いた転移学習システム
A Novel Transfer Learning Method Using Learning Vector Quantization

中村 宗広† 梶原 祐輔‡ 木村 春彦†
Munehiro Nakamura Yusuke Kajiwara Haruhiko Kimura

1. はじめに

転移学習は，対象タスクと関連したタスクを知識として利用して，対象タスクに対するパターン分類の精度を高める手法である．転移学習の適用例として，対象タスクの事例数が乏しく，SVM やニューラルネットワークなどのパターン分類器が有効的に機能しないケースが挙げられる[1]. 本稿ではタスクの定義を事例（instance）と特徴（feature）から構成されているものとする．

転移学習では，知識の転移元を元ドメイン（source domain），知識の転移先を目標ドメイン（target domain）と呼ぶ．神嶌[2]は，転移学習のアプローチを特徴ベース（feature-based）と事例ベース（instance-based）に分けて定義している．特徴ベースのアプローチ（例えば，[3, 4]）では，元ドメインにおける特徴の中から有用なものを目標ドメインで利用したり，元ドメインの特徴空間から有用な特徴空間を目標ドメインで求めたりする．事例ベースのアプローチ（例えば，[5, 6]）では，元ドメインの事例を重み付けしたり，選択することにより，目標ドメインのタスク学習に適したパラメータの設定などが行われる．これらの既存の転移学習手法では，元ドメインと目標ドメインのタスクが類似している必要がある．そこで，元ドメインと目標ドメインのタスクが異なっている場合にも適用可能な転移学習手法が提案されている[7]．この手法では，例えば元ドメインにテキストタスク，目標ドメインに画像タスクを設定するように，異なった 2 つのタスク間での転移学習を可能とする．

本研究では，さまざまな異なるタスクから学習した知識を断片的に利用して，対象タスクに対するパターン分類の精度を高める転移学習システムを提案する．提案手法では，まずベクトル量子化を用いて個々のタスクからコードブックと呼ばれる共通の知識を抽出する．次に，元ドメインと目標ドメインでのコードブックの類似度を用いて，元ドメインのさまざまなタスクから特徴を目標ドメインの対象タスクに転移し，対象タスクの特徴空間を拡大する．なお，本研究で用いるタスクは，教師データ付きのものを対象とする．

以降，2 節では関連研究について述べる．3 節では提案する転移学習システムについて述べる．4 節では 44 種類の実データに対して本手法を適用し，その有効性を検証した結果について述べる．

2. 関連研究

提案手法では，元ドメインのタスクにおける特徴の一部を目標ドメインのタスクに移転する．つまり，目標ドメインのタスクにおいて，人工的なデータを生成することから，提案手法はオーバーサンプリングの一種であると見なすことができる．一般に，オーバーサンプリングの手法は，クラス間の特徴分布のばらつきを減らすことにより，パターン分類器の性能を高めるために用いられる．オーバーサンプリングの手法の中で代表的なものは SMOTE（Synthetic Minority Over-sampling Method）[8]である．

SMOTE では，事例数が最も少ないクラスにおいて，特徴空間の中で近傍の事例間に新たな事例が生成される．SMOTE の改良版は数多く提案されている（例えば，[9]）．いずれの方法においても人工データを生成する際に個々のタスク内の特徴空間を参照している．一方，提案手法ではさまざまな異なったタスク間の特徴空間を参照する点が既存方法と異なる．

3. 提案手法

3.1 概要

まず，図 1 に提案手法のフローを示す．まず，元ドメインのタスクからコードブックを共通知識として抽出し，ストレージに保存しておく．次に，目標データのタスクの内，学習データからコードブックを抽出し，ストレージ内のコードブックとの類似度を算出する．そして，元ドメイン内のタスクにおける特徴空間の一部を目標ドメインの学習データの特徴空間に転移する．最後に，パターン分類器を用いて特徴転移後の学習データを学習し，クラス分類を実行する．

3.2 共通知識の抽出

学習ベクトル量子化（LVQ: Learning Vector Quantization）[10]は教師あり学習器の一つである．学習ベクトル量子化（以下，LVQ）のアルゴリズムでは，各クラスの特徴空間にコードブックと呼ばれる参照点を種々のアルゴリズムにより定める．一般にコードブックは各クラスにつき複数個設定される．そして，未知の事例が入力されたときに，未知の事例と最もユークリッド距離が近いコードブックを算出し，そのコードブックに対応したクラスが未知の事例のクラスであると判断される．したがって，コードブックは任意のタスクをクラス分類するための知識と見なすことができる．

3.3 特徴空間の転移

3.2 節で得られたコードブックを用いて，目標ドメインのタスクに特徴空間を転移する方法について述べる．まず，対象タスク T における各特徴のコードブック数を n，T における特徴ペアの組み合わせ数を nc，特徴ペアの集合を T_i（i=1,2,...,nc）と定義する．そして，T_1 の特徴ペアを x と y とすれば，T_1 の x，y に対応したコードブックは T_1 (x_j, y_j) (j=1,2,...n)と表される．ここで，元ドメインの任意のタスクを R とすれば，T_1 と R_j について，各コードブックの最近傍ペアを算出し，それらのマハラノビス距離の平均が最小となる R の特徴ペアを算出する．この計算を元ドメインのすべてのタスクに適用し，T_1 と最も類似した特徴ペアを決定する．

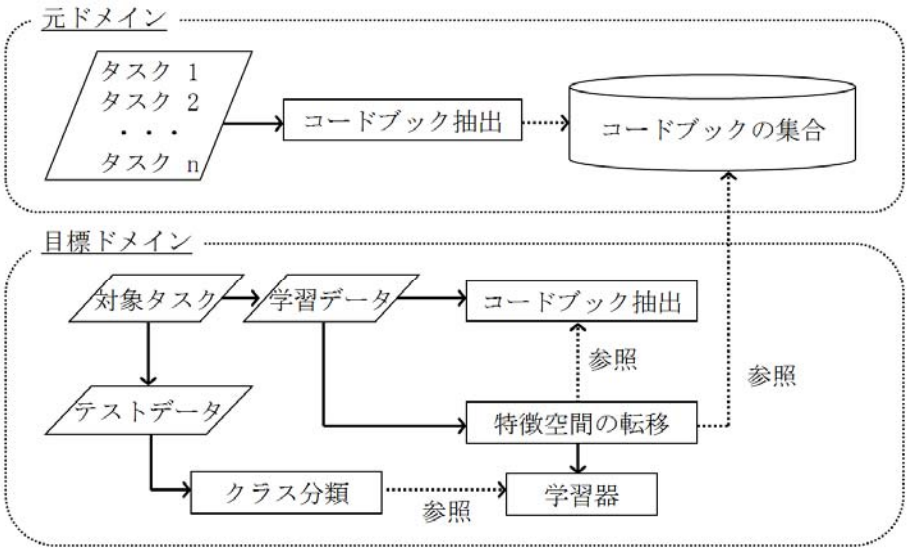

図 1　提案手法のフロー

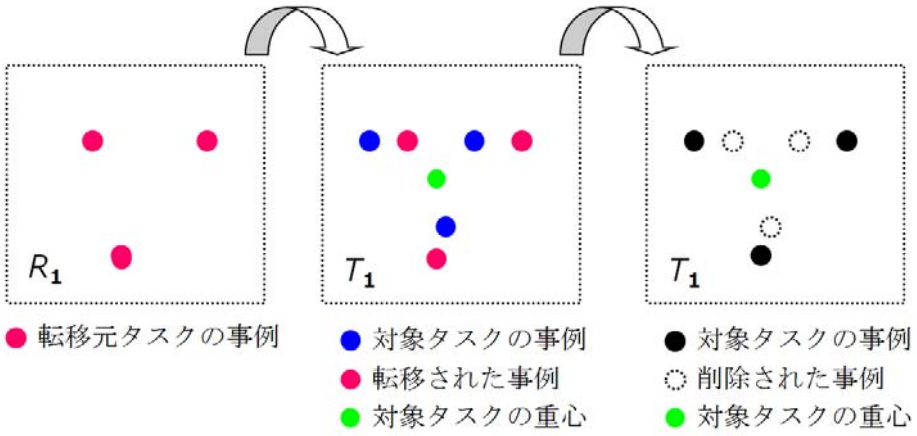

図 2　特徴空間の転移プロセス

次に，T_1 に対応した特徴ペアとして R_1 が選択されたと仮定する．図 2 に R_1 から T_1 への特徴空間の転移プロセスを示す．まず，図 2（左）の R_1 の各事例を図 2（中央）の T_1 に転移する．そして，転移された各事例と最もユークリッド距離が近い事例を算出し，図 2（右）に示すように，それぞれ T_1 の重心とのユークリッド距離が近い事例を削除する．

このように事例を選択することにより，T_1 の事例数を増加させることなく，特徴空間を拡大することができる．以上の T_1 に適用した処理を T_1～T_{nc} まで適用することにより，対象タスクに対する特徴空間の転移が完了する．

3.4 元ドメインのタスク選択

3.3 節で述べたように，本手法では元ドメイン内のタスクにおける特徴空間を目標ドメインの対象タスクに転移する．このため，転移元の特徴空間にある事例はなるべくパターン分類し易いものであることが望ましい．そこで，元ドメインに登録するタスクは，交差確認法を用いてパターン分類を実施した際に，再現率と適合率のそれぞれが任意の閾値以上のものとする．

4. 実験

提案手法の有効性を検証するために，UCI レポジトリから取得した 46 種類のさまざまなタスク（データセット）を用意した．パターン分類手法には，LIBSVM[11]のパッケージから SVM（サポートベクターマシン）を実装して実験に用いた．SVM のカーネルには PolyKernel を選択し，各種パラメータはパッケージのデフォルト設定とした．学習ベクトル量子化の手法には，文献[12]に記載されている LVQ の拡張版の内，前実験において良好な精度が得られた OLVQ3 を用いた．

表 1 は，3.4 節で述べた方法により，44 種類のタスクの中から選択された元ドメインのタスクである．交差確認法では，各タスクについて 90%を学習データ，残りの 10%をテストデータとする動作を繰り返し，再現率と適合率を算出した．このとき，交差確認法の信頼性を高めるため，

表 1　元ドメインのタスク一覧

データセット	特徴数	事例数
Balance-scale	4	625
Car	6	1728
Colic	22	368
Cylinder-bands	39	540
Dermatology	34	366
EdibleMushrooms	22	8124
FishersIrisDataset	4	150
Heart-statlog	13	270
Ionosphere	34	351
Iris	4	150
Labor	16	57
Segment	19	2310
SensorDiscrimination	12	2212
Sponge	45	76
Tic-tac-toe	9	958
VotingRecordYayNay	16	435
WineCultivars	13	153
Zoo	17	101

表 2　目標ドメインのタスク一覧

データセット	特徴数	事例数
Adalone	8	4177
Breast-cancer	9	286
Cmc	9	1473
Credit-g	20	1000
DiabetesDiagnosis	8	768
Fertility	9	100
Glass	9	214
Haberman	3	306
Hay-train	9	373
Hayes-roth-train	4	132
Hepatitis	13	303
LandformIdentification	6	300
Liver-disorders	6	345
Lung-cancer	56	32
Mfeat-morphological	6	2000
Molecular-biology_promoters	58	106
Postoperative-patient-data	8	90
Sick	29	3772
Sonar	60	208
Spect_train	22	80
Tae	5	151
ToPlayOrNotToPlay	4	14
Trains	32	10
Vehicle	18	846
Vowel	13	990
Weather	4	14

学習データおよびテストデータの選定を 1000 回繰り返し，その中でクラス内分散が元のデータと近い順に上位 10 回分の交差確認法を実行し，平均再現率と平均適合率を算出した．閾値の設定には，実験的に平均再現率 80%以上，平均適合率 80%以上とした．再現率（*Recall*）および適合率（*Precision*）は，以下の式で定義される．

$$Recall = \frac{N_{correct}}{N_{class}} \times 100 \quad (1)$$

$$Precision = \frac{N_{correct}}{N_{output}} \times 100 \quad (2)$$

N_{class} は各クラスにおける事例数，N_{output} は SVM により判定された事例数，$N_{correct}$ は SVM により正しく判定された事例数である．

次に，元ドメインの 18 種類のタスクと目標ドメインの 26 種類のタスクに提案手法を適用し，SVM を用いて目標ドメインのタスクをパターン分類した．このとき，交差確認法の設定は前述した方法と同様であり，OLVQ3 のコードブック数は各タスクのクラス数として設定した．表 2 に実験結果を示す．表 2 より，提案手法を用いた場合，26 種類中 19 種類の F 値が高くなっていることがわかる．

提案手法はさまざまな異なるタスクを対象としていることから，既存の類似したタスクを用いる転移学習手法と直接比較することは困難である．しかしながら，実験で用いたタスクのように，比較的分類し難いタスクは数多くあることが予想され，パターン分類の精度を向上させる手法として本研究は有用であると考えられる．

5. おわりに

本稿では，学習ベクトル量子化を用いてさまざまな異なるタスクから共通の知識を抽出し，対象タスクに対するパターン分類の精度を向上させる方法を提案した．提案手法に対する評価実験では，さまざまな実データを用いて提案手法の有効性を示した．今後の課題として，提案手法を適用した際に精度が低下したタスクに対して，特徴移転の方法を改良することが考えられる．

表 3　26 種類のタスクに対する再現率，適合率，G-mean，F 値

データセット	適用前				適用後			
	再現率	適合率	G-mean	F 値	再現率	適合率	G-mean	F 値
Adalone	54.5%	53.7%	54.1%	53.1%	**60.5%**	**64.9%**	**62.7%**	**59.9%**
Breast-cancer	**71.0%**	**63.6%**	**67.3%**	**59.7%**	66.0%	60.8%	63.4%	54.4%
Cmc	**48.3%**	**46.6%**	**47.5%**	**46.4%**	47.3%	46.1%	46.7%	45.6%
Credit-g	74.7%	69.6%	72.2%	66.5%	**87.8%**	**80.7%**	**84.3%**	**77.9%**
DiabetesDiagnosis	77.7%	76.8%	77.3%	72.3%	**79.7%**	**79.0%**	**79.3%**	**74.3%**
Fertility	88.0%	77.4%	82.7%	82.4%	**99.0%**	**88.4%**	**93.7%**	**94.5%**
Glass	56.1%	52.2%	54.2%	51.6%	**62.1%**	**58.5%**	**60.3%**	**56.7%**
Haberman	73.5%	**61.8%**	67.7%	50.4%	**73.9%**	61.5%	**67.8%**	**50.5%**
Hay-train	35.1%	**31.4%**	**33.3%**	28.3%	**37.1%**	29.4%	33.2%	**31.0%**
Hayes-roth-train	**53.0%**	**57.2%**	**55.1%**	**55.4%**	52.0%	55.7%	53.8%	54.3%
Hepatitis	**85.2%**	**77.4%**	**81.3%**	**76.8%**	83.2%	76.5%	79.8%	75.5%
LandformIdentification	69.7%	70.8%	70.2%	69.4%	**94.7%**	**86.9%**	**90.8%**	**88.7%**
Liver-disorders	57.7%	28.9%	43.3%	49.8%	**59.7%**	**28.4%**	**44.0%**	**51.8%**
Lung-cancer	**40.6%**	**41.7%**	**41.1%**	**42.5%**	35.5%	32.1%	33.8%	37.4%
Mfeat-morphological	**70.0%**	**70.7%**	**70.3%**	**69.9%**	66.0%	67.9%	66.9%	67.8%
Molecular-biology_promoters	30.2%	26.6%	28.4%	27.1%	**34.2%**	**33.3%**	**33.8%**	**32.2%**
Postoperative-patient-data	67.8%	**49.9%**	**58.9%**	57.5%	**80.8%**	33.8%	57.3%	**70.4%**
Sick	93.9%	**71.9%**	82.9%	50.2%	**96.9%**	71.7%	**84.3%**	**53.9%**
Sonar	76.0%	75.9%	75.9%	75.8%	**78.0%**	**77.7%**	**77.9%**	**78.4%**
Spect_train	67.5%	67.7%	67.6%	67.5%	**78.8%**	**72.6%**	**75.7%**	**78.8%**
Tae	54.3%	44.3%	48.3%	49.5%	**60.3%**	**58.8%**	**59.5%**	**59.6%**
ToPlayOrNotToPlay	**42.9%**	**27.3%**	**35.1%**	**33.3%**	26.2%	12.0%	19.1%	16.7%
Trains	70.0%	70.8%	70.4%	70.0%	**90.0%**	**92.7%**	**91.4%**	**90.0%**
Vehicle	74.8%	73.7%	74.3%	75.1%	**74.8%**	**88.1%**	**81.4%**	**75.8%**
Vowel	70.2%	69.6%	69.9%	70.2%	**80.2%**	**89.8%**	**85.0%**	**85.8%**
Weather	42.9%	27.3%	35.1%	33.3%	**59.5%**	**48.0%**	**53.8%**	**50.0%**

参考文献

[1] S. J. Pan and Q. Yang, “A Survey on Transfer Learning”, *IEEE Trans. on Knowledge and Data Engineering*, Vol.22, No.10, pp.1345-1359 (2010).

[2] 神嶌, “転移学習”, 人工知能学会 Vol.25, No.4, pp.572-580 (2010).

[3] X. Ling, W. Dai, G.-R Xue., Q. Yang, and Y. Yu, “Spectral Domain-Transfer Learning”, *In Proc. of The 14th Int’l Conf. on Knowledge Discovery and Data Mining*, pp.488-496 (2010).

[4] X. Tian, D. Tao, and Y. Rui, “Sparse Transfer Learning for Interactive Video Search Reranking”, *ACM Trans. on Multimedia Computing, Communications, and Applications*, Vol.8, No.3 (2012).

[5] P. Wu and T. G. Dietterich, “Improving SVM Accuracy by Training on Auxiliary Data Sources”, *In Proc. of The 21st Int’l Conf. on Machine Learning*, pp.871-878 (2004).

[6] 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).

[7] W. Dai, Y. Chen, G.-R. Que, Q. Yang, and Y. Yu, “Translated Learning: Transfer Learning Across Difference Feature Spaces”, *In Proc. of The 21st Ann. Conf. Neural Information Processing Systems*, pp.353-360 (2008).

[8] N. V. Chawla and K. W. Bowyer and L. O. Hall and W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique”, *Journal of Artificial Intelligence Research*, Vol.16, No.1 (2002).

[9] S. Shen, H. He, and E. A. Garcia, “RAMOBoost: Ranked Minority Oversampling in Boosting”, *IEEE Trans. on Neural NEtworks*, Vol.21, No.10 (2010).

[10] T. Kohonen, “Learning Vector Quantization”, *MIT Press* (1995).

[11] C. C. Chang and J. C. Lin, “LIBSVM: a Library for Support Vector Machines”, *ACM Trans. on Intelligent Systems and Technology*, Vol.2, No.27. pp.531-537 (2011).

[12] T. Kohonen, “LVQ PAK: the Learning Vector Quantization Program Package”, *http://www.cis.hut.fi/research/lvq_pak/* (1996).