能動学習を用いた系列データの自動分類

井上　眞乙† 　白井　匡人†† 　三浦　孝夫†

† 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
†† 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
E-mail: †maoto.inoue.7h@stu.hosei.ac.jp, ††shirai@cis.shimane-u.ac.jp, †††miurat@hosei.ac.jp

あらまし　本稿では、系列データの分類を論じる. 分類器を構築するために学習データが必要となるが，系列データの分類は容易でない．本研究では，隠れマルコフモデル（HMM）に基づく確率分類のアプローチを行う．各クラスにモデルを構築し，系列データを適用し，最大尤度のクラスを推定する．HMM は少量の学習データを必要とするが，HMM が上手く機能するための助けとなる．さらに，分類器の構築に対する能動学習のアプローチを提案する．主たるアイデアは，HMM がより良い構築に役立つ有用なデータが必要だと判断する度に，新しい学習データを自動的・自律的に要求することにある．

キーワード　系列分類，隠れマルコフモデル，能動学習

1. 前　書　き

現在，膨大な量の情報が日々増加しており，インターネットを介して簡単かつ高速にアクセスする事が可能となっている．データマイニングはこれらからの有用な知識の抽出を助長し，現代の興味深い研究テーマの 1 つとなっている．しかし，大規模な量のため作業を行う事は容易ではなく，情報は瞬時に流れ去っていく．正しい知識を抽出するためには当然，効率的かつ正確で有用な技術が必要になる．

本研究では，系列データの分類問題を論じる．単純な構造や形式を持つデータとは異なり，多くの場合は音声やプログラムコード，または文章のような系列データである．系列は要素（場合によってはラベル付き）で構成され，これらのラベルはクラスに応じて独自の役割を果たす．例えば,「ソナタ」というクラスは，4 つの主要な旋律から成る音楽構造を持つ：主題の提示，展開，再現，コーダ．この形式は 18 世紀後期の音楽に広く現れている（初期の古典派の時代）[Wiki]．ある音楽の各要素がいずれかの旋律に分類されると，音楽は「ソナタ」の形をしていると言える．従って，音楽の要素とそのラベルの並びはクラス「ソナタ」を構成する．

系列データの問題に取り組むため，HMM，MEMM，CRF など，これまでに提案されている確率論に基づく高度な手法がいくつか存在している [6]．隠れマルコフモデル（HMM）は少量の学習データで済むが，これらのデータは HMM が上手く機能する助けになる．最大エントロピーマルコフモデル（MEMM）と条件付き確率場（CRF）は，系列モデルに対して強力ではあるが，事前にある程度の量の学習データが必要である．では最初に，どのような学習データを考えることが出来るだろうか．どのようにラベルを（部分/全体の）系列に当てることが可能か．特定分野の知識をどの程度まで考慮すべきか？また，ラベル付に対する主観性を避ける事が可能か？

学習データの取得が容易ではないことは明らかであり，多くの時間とコストが掛かり，大量の学習データを取得する事はほぼ出来ない．本研究では，能動学習（Active Learning HMM, ALHMM と呼ぶ）を用いた HMM に基づく系列分類の新しい手法を提案する．本研究の貢献は以下の 3 点である：

(1)　分類器に必要な量の学習データを構築する事が可能．

(2)　正確な分類器を構築するための十分な少量のデータを選択する．

(3)　モデルの構築のための ALHMM の効率的なアルゴリズムを実装する．

本論文は次のように編成されている．2 章では分類と学習データについて，3 章では HMM と能動学習について述べる．4 章では ALHMM のアプローチを提案し，5 章では実験結果を示し，6 章では結論を述べる．

2. 分類と学習データ

分類とは，事前に与えられた別個のグループの一つを情報に割り当てることをいう．このようなグループをクラス（ラベル付き）と呼び，分類する法則を分類器と呼ぶ．しかし，膨大な量の情報が存在し，それらを分類するのに多くの時間/コストが掛かってしまう．さらに深刻なことに，分類結果は主観に依存してしまう可能性がある．

これらの問題を克服するため，情報を自動的に分類する手法である分類器を導入する．このような手法は，高速で安価で効率的な処理を提供する．情報の中身を調べることで，情報に特徴を付けて識別し，分類するという幾つかの機能がある．例えばクラス「野球」は,「ピッチャー」,「イチロー」,「ホームラン」などの特徴語で識別出来る．これらの特徴を見つけ，その特徴が情報に現れているかどうかを調べる事により，クラス「野球」に正しく分類する事が期待される．

データを分類するための特徴をどのように抽出する事が可能か？一般には，分類器を構築するための学習データが用意されており [8]，このようなアプローチは教師あり学習と呼ばれている．学習データは，データを調べ，クラスを決定付けるための知識を提供する．決定木，k-NN 及びいくつかのベイズ手法

が提案されている [8]．しかし，主に 2 つの問題がある．第 1 に，分類器は多くの学習データを必要とし，第 2 に単純で整理されたデータを前提としている．基本的に学習データが多いほどより良い分類器が得られる．膨大な量の学習データが必要であるが，大きな変化にすばやく追随する事は難しく（バースト現象と呼ばれる），系列などの構造化された情報の扱いが困難である．

系列分類に関して，N-gram に基づく文書分類などの手法がいくつか提案されている．連続する N 語を 1 つにまとめることにより，頻度，共起及びそれらの関係を調べる事が可能となる．しかし，系列が大規模になると N-gram は疎な分布を持つ傾向にあり，計算が複雑になる [8]．

ここでは，系列データの分類に対する隠れマルコフモデル（HMM）に基づく確率論的アプローチをとる．この目的のために，各クラスに対し分類器を構築し，系列データを適用する．系列の概念が論じられていないため既知の分類手法を系列に適用する事は容易ではない．本稿では，確率的手法と最尤原理によりクラスを推定する．

3. 隠れマルコフモデルと能動学習

本章では，HMM と能動学習の背景と動機づけについて検討する．

3.1 マルコフモデル

確率過程とは，確率変数が一組の状態に関連付けられ，時間の経過と共にモデルがランダムに変化する数学的な手法を意味し，確率的有限オートマトン（出力観測値付）と考えられる．このアプローチは，様相がランダムに変化する現象及びシステムの数学的モデルに広く使用されている．一般的な例として，音声認識などが挙げられる．マルコフモデル（MM）は未来の状態が現在の状態のみに依存し，それ以前に発生した状態/出来事の影響は受けないという仮定を伴う確率論の枠組みの 1 つである．これはマルコフ性と呼ばれる．一般的にこの仮定は，単純で効率的な方法で予測モデリングや確率的推測を容易にし，扱いにくいシステムへの推論や計算を可能にする [9]．

しかし，マルコフモデルには 2 つの深刻な欠点がある．どの状態にいるのかを見極める方法と，学習データを構築する方法が不明な事である．学習データによってモデルを構築し，状態遷移と出力観測値を求める必要があるため，真の問題は，多くの学習データを正確に，かつ効率的に取得する方法である．

3.2 隠れマルコフモデルと分類

隠れマルコフモデルとは，状態が部分的にしか観測できないマルコフモデルの一種である [3]．出力観測値はシステムの状態に関連付けられるが，状態を正確に決定するには不十分である．では，どのような潜在意味が，最も適した解釈に対する観測値と関連した状態であるのか？隠れマルコフモデルは 3 つの問題に対する解決策を持つ．観測系列 $O = o_1...o_N$ と仮定する．第 1 に，観測系列 O の確率を得る事が可能である．実際には，マルコフモデルを仮定すると，O を生成する全ての経路を探索し，その確率の合計をとる．第 2 に，観測系列 O を生成する可能性が最も高い経路を取得する事が可能である．ビタビアルゴリズムにより，動的計画法を用いて効率的に最尤経路を推定する．また，フォワードアルゴリズムは観測系列の確率を計算する．

1 つの重要な適用は，最尤原理法（MLP）を利用し，HMM を系列データの分類器としてみなすことが出来る．つまり，各クラスにマルコフモデルを構築し系列の尤度（ビタビによる）を得て，最大尤度の系列をもつクラスを決定する．例えば，ソナタマルコフモデルと音楽が与えられたとする．出力観測値として考慮したこの音楽を区切り，主題・展開・再現・コーダの状態遷移を推測する．同様にロンド，バラードに対しても値を取得する．次に，この音楽を最大の尤度である音楽形式に分類する．

HMM の第 3 の貢献は，マルコフモデルの推定が可能なことである．つまり状態遷移確率 $P(s_j|s_i)$，状態 s_i における観測値 o_k の確率 $P(o_k|s_i)$ 及び開始時の初期状態確率 $P(s_i)$ を推定する事が可能である．基本的に HMM を推定するために EM アルゴリズムを適用する．始めから EM アルゴリズムによる調整過程を繰り返すことにより，データの尤度を最大化していく．事前に学習データを必要としない．隠れマルコフモデルにはいくつか知られたアルゴリズムが存在し，特に Baum Welch（BW）アルゴリズムは，(学習されていない）データ集合から効率的にマルコフモデルを推定する．

しかし，十分なサイズの「学習データ」は，(繰り返し過程における）高速な収束と精度の良い結果を得るための助けとなり，局所的最小値を回避するのに役立つ．学習データまたは不良な初期値が無いという条件が無ければ，これらの問題を回避する事は困難である．本研究では，この側面に取り組み，関連するパラメータを改善する．

3.3 能 動 学 習

能動学習とは，高い分類精度を達成するため，学習に有用なデータを自動・自律的に選定する枠組みである．ここでいう「有用」とは，分類器の精度を向上させるためのデータを意味する．幾つかのデータから始め，出来るだけ少量の学習データでの推測を可能にする．ここで注意しなければならないのは，どのような条件でデータを選定するかである．

能動学習には少なくとも 2 つの利点がある．第 1 に，大規模な学習データを扱うことなく効率的に高精度の分類を達成する事が可能である．第 2 に，結果が注釈者の主観に依存しない，つまり，学習データを構築するため合理的で統一された条件を定めることが出来る．

これまで，未知データのラベル付を可能にする 3 つの手法が提案され，能動学習に関する研究がいくつか行われてきた．第 1 の手法は，プールベース型能動学習（Pool-Based Active Learning）である [10]．プール内の全ての実例を評価するための情報尺度に従い，大規模なデータのプールからいくつかデータをしきりに選択する．

第 2 は，ストリームベース型能動学習（Stream-Based Active Learning）と呼ばれ，新しいデータが入るたびに学習データに

入れるべきかを尋ねる [4],[5]．ここではラベルの無い実例を取得する事は安価であると想定し，ラベル付けを要求するかどうかの決定を求められる．この決定には，情報量の測定やクラスタリングなどの他の機械学習手法を使用するなど，いくつかの方法で組むことが出来る．

最後は，クエリ生成型能動学習法（Membership Query Syntheses）である [2]．これは入力ストリーム内の未知データに対しラベルの付与を要求する．任意のデータにラベルを付ける注釈者を仮定し，定理が難しい場合がある．ここでは，HMM を改善するためのアプローチとする．

能動学習の手法には，未知データの情報性を評価する事が含まれている．不確実性サンプリング (Uncertaity Sampling), Query By Committee, 推定されるモデルを変えるなど，方法を策定する提案は多く存在する [9]．とりわけ,「不確実性サンプリング (Uncertainty Sampling)」手法，周辺確率（Margin Sampling）[12]，エントロピーに基づく方法（Entropy-Based Approach），最小/最大確信度法（Least/Largest Confident）[6] に基づく 3 つの方法に注目する．

周辺確率（Margin Sampling）とは，最大の確率と 2 番目に大きい確率との差が最小となるデータを選択する．エントロピーを用いる方法では，クラス内で最小のエントロピーを維持できる新しいデータを選択する．最後に最小/最大確信度法は，クラスの最小所属確率の最大（または最大所属確率の最小）を持つ新しいデータを提供する．

一方，学習データの構築には少なくとも 2 つの欠点がある．第 1 に，いわゆるサンプリングバイアス問題，すなわち，いくつかの条件に従いサンプルを取得する場合，常に最良とは限らない場合がある．可能な解決策として例えば，サンプリングのために関心のあるトピックに重きを置く重点サンプリングがある．第 2 の欠点は，パラメータの条件を途中で変更する場合，同一のデータを学習のため何度も利用する事（再利用性）である．

4. HMM に対する能動学習

本章では，HMM と能動学習，または HMM に対し能動学習を組み合わせる新しいアプローチを提案する．HMM の繰り返し推定の間ではデータの選択を決定するのに十分な所属確率が分からないため，このアプローチは常に改善するとは限らないことに注意したい．

Anderson らは，能動学習の観点から HMM パラメータ，状態遷移，最良な状態の推定について述べている [1]．学習データを使用してモデルのどの部分を改善すべきかを教える損失関数が含まれている．不確実性サンプリングと比較し，少量の学習データでも有効であり，効率的な状況が臨める事を示した．しかし，彼らはモデル尤度（及び分類）については言及していない．

ここでは，能動学習を用いた HMM の ALHMM と呼ばれるモデル計算を提案する．到来の素朴な HMM においては，学習データ（ラベル付きデータ）は想定せず，系列の尤度のみを調査する．EM アルゴリズムに基づき収束するまでの値を最大にする．ここでは，どんなに大きな尤度に関わらず所属確率を区別するだけで十分である．

本稿では，クエリ生成型能動学習の手法により，分類するテスト系列データの集合と，能動学習のための開発データと呼ばれる各クラスへのモデル開発のための系列データの集合が存在すると仮定する．

ALHMM では，HMM 推定の過程を以下のように設定する．

（Step0） 新しい各系列データに対し（1）〜（4）を行う．

（Step1） 現在のパラメータから各クラスに対する新しい HMM パラメータを推定する．

（Step2） テスト系列データ 1 件に最尤原理（MLP）を適用し，最大の尤度と 2 番目に大きい尤度の差を比較し，閾値 q を上回るかどうかを調べる．

（Step3） 以下であれば，新たな系列データとクラスを選択し学習を行い，そのクラスのモデルの新しい HMM パラメータを推定する．新しい開発系列データにはクラスラベルが付いているが，状態の情報は含まれていないため，クラスの尤度を改善するのに役立つ．

（Step4） 再度，最尤原理（MLP）を適用し，Step2 の系列データのクラスを決定（分類）する．

ALHMM の Step（3）では，(a) 系列の長さが閾値 σ 以下であり，(b) 系列が最小確信度または最大確信度を満たす場合に限り，1 件の開発系列データを選択する．閾値 σ は，全ての開発データの平均サイズとする．一般的に，系列が短いほど尤度が高くなる可能性があることに注意すべきである．これは，新しい開発データによってクラスの平均尤度が改善される可能性があることを意味する．最小確信度法を採用すると，曖昧なクラスの所属確率を上げることができるが，最大確信度法では小規模クラスを回避し，クラスサイズを洗練するのに役立つ．

さらに，Step（4）で再度閾値 q 以下であった場合，その系列データの分類を止め，閾値 q の値を下げて次のテスト系列データのため Step（1）へ戻る．また，閾値 q 以上であるデータが続いた場合は，閾値 q を上げて次のデータの分類を行う．本研究では初期の閾値 q を 0.5 に設定し，上げ幅・下げ幅を変更する．

5. 実　　験

5.1 実 験 手 順

隠れマルコフモデル（以下，HMM と呼ぶ.），最大確信度法で選定する能動学習を用いる隠れマルコフモデル（以下，Largest-ALHMM と呼ぶ.）を用い，HMM をベースラインとし提案手法である ALHMM を評価する．各々により所属尤度を推定し，適合率および再現率を比較する．さらに，HMM では学習データ数 15 件，21 件，27 件，30 件，36 件，42 件，45 件の HMM の精度及び f 値を求める．実験では「DAILY YOMIURI 記事データ集 2007 年版」の 1 月 1 日の先頭から，学習データとしてスポーツ，経済，科学クラスにそれぞれ同じデータ数を割り当てる．HMM，ALHMM のどちらの場合においても扱うテストデータの中身は同じものとする．

「DAILY YOMIURI 記事データ集 2007 年版」コーパスは予め GoTagger により前処理をする．各文章の単語に品詞付けをし，名詞と動詞のみの自立語を抽出する．全 465 件中の記事データから，テストデータとしてスポーツ，経済，科学記事に 50 件ずつ用いる．開発データにはスポーツ，経済，科学記事に 90 件ずつ割り当てる．

本研究では，ALHMM を学習データ全 15 件から開始する．能動学習で選定されるデータは開発データ 270 件より選定されるものとし，学習データ数に上限を設けない．

図 1 に品詞付けを行った「DAILY YOMIURI 記事データ集 2007 年版」の記事データの一部を示す．

```
1 winter_NN Japan_NNP teams_NNS begin_VBP quest_NN league_NN titles_NNS Japan_NNP
  Series_NNP rings_NNS today_NN spring_NN training_NN opens_VBZ Kyushu_NNP Okinawa
  _NNP Australia_NNP↓
2 chance_NN address_VB needs_NNS BayStars_NNP loaded_VBD arms_NNS ignored_VBN bats
  _NNS missed_VBD boat_NN↓
3 Hingis_NNP showed_VBD developed_VBN game_NN thumping_VBG Australia_NNP Nicole_NN
  P Pratt_NNP advance_VB quarterfinals_NNS dollars_NNS Toray_NNP Pan_NNP Pacific_N
  NP Open_NNP↓
4 coach_NN Top_NNP League_NNP team_NN Mitsubishi_NNP Juko_NNP Scott_NNP Pierce_NNP
  is_VBZ going_VBG get_VB opportunities_NNS play_VB↓
5 Ai_VBP Sugiyama_NNP confirmed_VBD Thursday_NNP was_VBD looking_VBG doubles_NNS p
  artner_NN replace_VB Daniela_NNP Hantuchova_NNP has_VBZ played_VBN May_NNP↓
6 New_NNP York_NNP Yankees_NNP have_VBP come_VBN Asia_NNP build_VB game_NN end_NN
  turn_VB profit_NN↓
7 Sugiyama_NNP made_VBD adjustment_NN needed_VBD rallied_VBD defeat_VB Russia_NNP
  Maria_NNP Kirilenko_NNP advance_VB Thursday_NNP quarterfinals_NNS dollars_NNS To
  ray_NNP Pan_NNP Pacific_NNP Open_NNP↓
```

図 1　品詞付済みの記事データ

各開発データのサイズを求め，閾値以下である記事を表 1 に示す．

表 1　閾値以下の開発データ

閾値	15.93									
正解ラベル	スポーツ									
記事番号	3	5	6	7	11	14	15	16	17	18
サイズ	14	9	12	13	8	15	13	7	12	12
記事番号	22	24	25	26	27	28	32	33	34	35
サイズ	7	11	5	12	5	13	10	7	11	9
記事番号	36	37	38	39	41	42	44	45	46	48
サイズ	15	8	10	10	13	11	9	12	15	14
記事番号	49	50	54	55	60	61	62	64	65	66
サイズ	15	15	15	15	13	10	12	8	8	14
記事番号	67	68	69	70	71	75	76	78	79	80
サイズ	12	12	10	15	6	15	13	13	3	6
記事番号	81	82	83	84	86	87	88	89	90	
サイズ	12	12	7	9	15	10	10	7	12	
正解ラベル	科学									
記事番号	91	92	93	97	98	102	104	105	109	112
サイズ	9	12	13	13	9	15	6	8	15	10
記事番号	113	114	116	117	121	124	128	129	132	133
サイズ	15	8	15	11	13	9	5	6	14	5
記事番号	138	141	143	147	149	151	153	156	157	160
サイズ	6	7	14	15	15	14	11	11	13	8
記事番号	162	163	166	167	171	172	173	176	177	178
サイズ	13	13	13	12	15	8	12	12	14	9
記事番号	180									
サイズ	13									
正解ラベル	経済									
記事番号	181	182	184	190	192	200	201	204	205	213
サイズ	10	13	15	13	9	13	11	7	13	11
記事番号	215	219	225	230	233	235	236	238	241	243
サイズ	10	12	15	14	9	8	15	7	15	14
記事番号	244	249	251	253	255	256	260	261	262	263
サイズ	15	13	15	15	11	8	14	15	14	14

各記事には所属するクラスラベルが付与されているが，開発データから学習データに追加する場合には，この正解ラベルを用いて該当のクラスのモデルの再計算を行う．

閾値として平均サイズである 15.93 を仮定する．サイズがこの条件を満たす開発データは，スポーツ記事から 59 件，科学記事から 41 件，経済記事から 30 件あり，計 130 件が能動学習での選定対象となる．

分類器の評価には適合率及び再現率を用いる．適合率は MicroPrecision で評価する．

5.2　実 験 結 果

能動学習を施した回数及び選定されたデータを表 2 に示す．

表 2　能動学習を施した回数及び選定データ

能動学習を施した回数	設定した閾値 q	閾値以下のテストデータ（記事番号）	選定された開発データ
初期（1 回目）	0.5	2	科学記事（記事番号：102）
2 回目	0.5 → 0.1	2, 3	科学記事（記事番号：116）
3 回目	0.1 → 0.01	3, 7	経済記事（記事番号：236）
4 回目	0.01 → 0.0001, 0.0001 → 0.01	7, 20	経済記事（記事番号：251）
5 回目	0.01 → 0.0001, 0.0001 → 0.01	20, 32	スポーツ記事（記事番号：54）
6 回目	0.01 → 0.0001, 0.0001 → 0.01	32, 46	経済記事（記事番号：230）
7 回目	0.01 → 0.0001, 0.0001 → 0.01	46, 62	経済記事（記事番号：244）
8 回目	0.01 → 0.0001, 0.0001 → 0.01	62, 80	スポーツ記事（記事番号：14）
9 回目	0.01 → 0.0001, 0.0001 → 0.01	80, 91	科学記事（記事番号：109）
10 回目	0.01 → 0.0001, 0.0001 → 0.01	91, 102	経済記事（記事番号：241）
11 回目	0.01 → 0.0001, 0.0001 → 0.01 → 0.1 → 0.5	102, 133	科学記事（記事番号：147）
12 回目	0.5 → 0.1	133, 134	経済記事（記事番号：261）
13 回目	0.1 → 0.01, 0.01 → 0.1	134, なし	なし

HMM での精度は表 3 で示すように，全 15 件の場合では適合率が 47.3 %，再現率は 33.3 %である．学習データ全 45 件の場合では適合率が 75.3 %，再現率が 60.7 %である．図 2 に学習データ全 45 件の場合の各クラスのマルコフモデルを示す．状態名は，状態から出力される頻出語に基づき推定している．

Largest-ALHMM の精度の結果を表 4 に示す．ここで能動学習により開発データから選定された記事データは，科学記事から 4 件（記事番号：102,109,116,147），スポーツ記事から 2 件（記事番号：14,54），経済記事から 6 件（記事番号：230,236,241,244,251,261）であり，学習データは最終的に全 27 件となる．適合率は 71.3 %，再現率は 54.7 %である．図 3 に各クラスのマルコフモデルを示す．

表 3　HMM（ベースライン）における精度

学習データ全15件

適合率(MicroPrecision)：47.3(%)，再現率：33.3(%)

結果 / 正解	スポーツ（件）	科学（件）	経済（件）	?	記事数（件）
スポーツ	31	2	2	15	50
適合率(%)	62.0				
科学	21	6	10	13	50
適合率(%)		12.0	20.0		
経済	18	11	13	8	50
適合率(%)		22.0	26.0		

学習データ全45件

適合率(MicroPrecision)：75.3(%)，再現率：60.7(%)

結果 / 正解	スポーツ（件）	科学（件）	経済（件）	?	記事数（件）
スポーツ	30	9	11	0	50
適合率(%)	60.0				
科学	10	26	14	0	50
適合率(%)		52.0	18.0		
経済	7	8	35	0	50
適合率(%)		16.0	70.0		

表 4　Largest-ALHMM における精度

適合率(MicroPrecision)：71.3(%)，再現率：54.7(%)

結果 / 正解	スポーツ（件）	科学（件）	経済（件）	?	記事数（件）
スポーツ	33	3	14	0	50
適合率(%)	66.0				
科学	19	11	20	0	50
適合率(%)		22.0			
経済	7	5	38	0	50
適合率(%)			76.0		

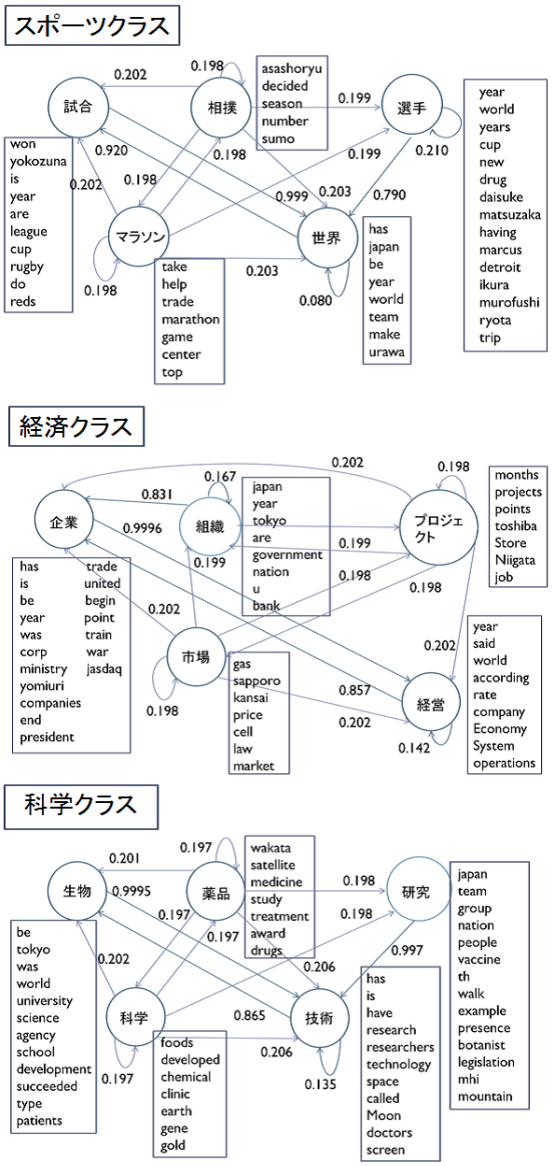

図 2　HMM（ベースライン）における各クラスのマルコフモデル（学習データ全 45 件）

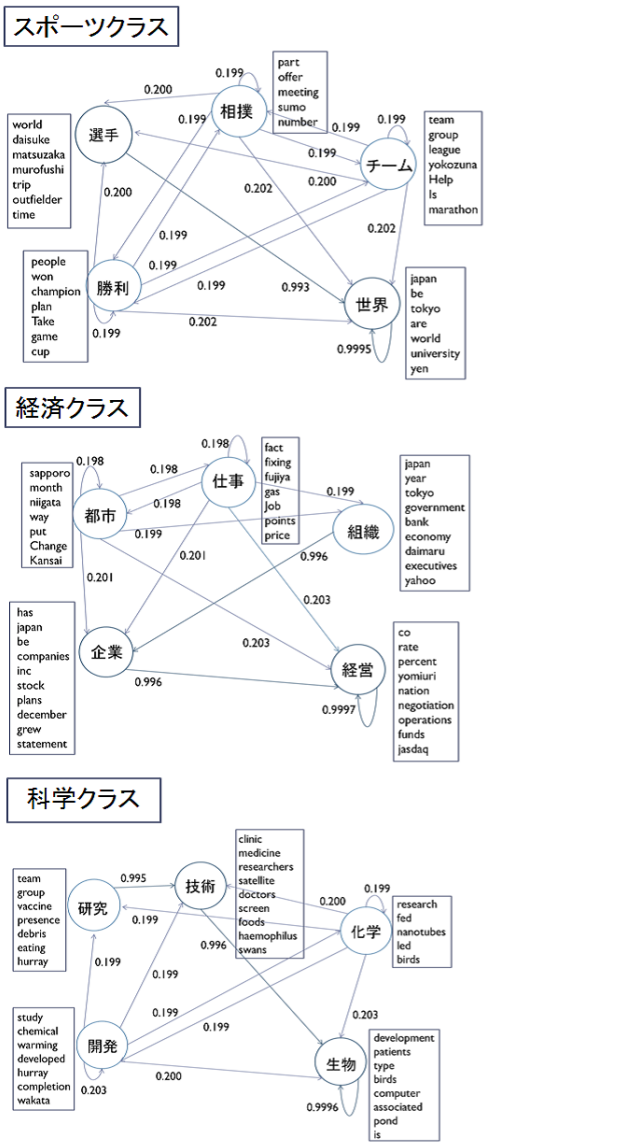

図 3　Largest-ALHMM における各クラスのマルコフモデル

Largest-ALHMM において，全 15 件の HMM 場合よりスポーツクラス記事の正しい割り当てが 4.0 %，科学クラス記事では 10.0 %，経済クラス記事では 50.0 %大幅に改善している．また，適合率が 24.0 %，再現率が 21.4 %向上している．さらに，全 45 件の HMM の場合の適合率 75.3 %，再現率 60.7 %と比べると，ほぼ変わらない精度となっている．

Largest-ALHMM におけるマルコフモデルについては，どちらも各クラスの状態遷移確率が増加している．また，ALHMM と 45 件の HMM の場合の科学，経済クラスのマルコフモデルには頻出経路及び構造に差が見られなく，類似した振る舞いをしていると見れる．しかしスポーツクラスにおいて違いがよく現れている．

頻出経路を比較すると，全 45 件の場合では「選手」→「世界」⇔「試合」の 3 通りの経路というモデルを構築している．

ALHMM では,「選手」→「世界」の 2 経路で構築されている．どの方法でも，共通する頻出語「matsuzaka」「murofushi」から推定する状態名「選手」が初期状態となっている．

さらに，学習データ量別の HMM と ALHMM との精度及び f 値の比較を表 5 に示す．

表 5　HMM と ALHMM との精度及び f 値比較

HMM	適合率	再現率	f 値
全 15 件	0.473333333	0.333333333	0.391184573
全 21 件	0.593333333	0.473333333	0.526583333
全 27 件	0.62	0.42	0.500769231
全 30 件	0.646666667	0.413333333	0.504318658
全 36 件	0.7	0.486666667	0.574157303
全 42 件	0.706666667	0.46	0.557257143
全 45 件	0.753333333	0.606666667	0.672091503
ALHMM	適合率	再現率	f 値
全 27 件	0.713333333	0.546666667	0.618977072

表 5 より，学習データ量が増加するほど適合率，f 値が伸びていることが分かる．ALHMM での適合率は 71.3 %，f 値は 0.618977 であるが，4 割程少ない学習データ量で全 42 件の HMM 以上の性能を示している．

5.3 考　　察

能動学習により Largest-ALHMM の適合率，再現率が向上し，全体として全 15 件の場合の HMM よりも改善している．また，27 件という少量の学習データ数で全 42 件の HMM 以上の性能を示している．この原因として，各テストデータの所属クラスの尤度に違いが出て，無所属の記事が本来の正解クラスに分類されたからであると考える．例えば，HMM（全 15 件）ではどのクラスにも属さなかった記事（本来経済クラスの記事）が経済クラスに属すようになっている．表 6 に HMM 全 15 件（能動学習適用前）と Largest-ALHMM（能動学習適用後）のテストデータの所属尤度の一部を示す．

表 6　テストデータの所属尤度の変化

能動学習適用前（HMM 全 15 件）					能動学習適用後（Largest-ALHMM）				
正解クラス	番号	スポーツ尤度	科学尤度	経済尤度	正解クラス	番号	スポーツ尤度	科学尤度	経済尤度
経済	101	4.00817E-47	9.69406E-52	4.71131E-49	経済	101	1.75021E-50	1.3714E-50	2.4086E-48
	102	9.72296E-34	9.75243E-34	9.75243E-34		102	9.78194E-34	8.81927E-34	9.85679E-34
	103	1.16909E-26	3.81107E-26	2.83387E-26		103	9.80153E-28	9.82713E-28	2.46568E-26
	104	1.16675E-32	9.75243E-34	9.75243E-34		104	9.78194E-34	8.81927E-34	2.4642E-32
	105	2.39061E-64	3.75806E-68	3.12208E-67		105	9.6652E-70	3.96753E-67	1.69713E-66
	106	1.66E-75	3.75E-77	1.73E-77		106	9.63623E-79	2.83055E-77	1.413E-76
	107	9.59E-76	9.62E-76	1.73E-74		107	9.64588E-76	9.76443E-76	1.17786E-74
	108	9.74E-28	9.77E-28	9.77E-28		108	9.80153E-28	9.84689E-28	9.86271E-28
	109	9.76E-22	9.79E-22	9.79E-22		109	9.82117E-22	9.85478E-22	9.86863E-22
	110	1.16E-56	9.67E-58	3.13E-55		110	1.74671E-56	9.80758E-58	2.45829E-56

さらに，ALHMM と全 45 件の HMM のスポーツクラスのマルコフモデルのみ，頻出経路及び構造に差が見られた．全 45 件では「選手」→「世界」⇔「試合」の 3 経路であるが，ALHMM では「選手」→「世界」の 2 経路である．これらが類似したマルコフモデルであるかを確かめるため，各状態の出力語の確率分布の差異を検定する．各方法での状態「世界」と「試合」のそれぞれから出力される語を全て列挙し，多項分布の MAP 推定を行うことにより出現確率を求める．検定には適合性検定を用いる．検定結果を表 7 に示す．

表 7　適合性検定による各分布の差異

理論値：HMM45 世界，標本値：HMM45 試合														
ラベル	is	year	are	league	cup	do	reds	has	japan	be	world	team	make	urawa
理論値	0.02173913	0.195652174	0.02173913	0.02173913	0.02173913	0.02173913	0.02173913	0.02173913	0.086956522	0.086956522	0.130434783	0.152173913	0.086956522	0.065217391	0.065217391
標本値	0.083333333	0.25	0.083333333	0.083333333	0.138888889	0.083333333	0.083333333	0.083333333	0.027777778	0.027777778	0.027777778	0.027777778	0.027777778	0.027777778	0.027777778
LOG	1.662757832	0.708515322	1.662757832	1.662757832	1.662757832	1.662757832	1.662757832	1.662757832	1.06969784	1.06969784	0.884606581	0.817659792	1.06969784	1.185630577	1.185630577
	1.079181246	0.602059991	1.079181246	1.079181246	0.857332496	1.079181246	1.079181246	1.556302501	1.556302501	1.556302501	1.556302501	1.556302501	1.556302501	1.556302501	1.556302501
	0.204817337	0.015995049	0.204817337	0.204817337	0.20014098	0.204817337	0.204817337	0.215568285	0.215568285	0.51002945	0.667261687	0.215568285	0.115881401	0.115881401	
X^2	3.533981507														
理論値：HMM45 世界，標本値：ALHMM 世界															
ラベル	japan	be	tokyo	are	world	university	yen	has	year	team	make	urawa			
理論値	0.090909091	0.136363636	0.022727273	0.022727273	0.159090909	0.022727273	0.022727273	0.090909091	0.204545455	0.090909091	0.068181818	0.068181818			
標本値	0.107142857	0.107142857	0.107142857	0.107142857	0.142857143	0.142857143	0.107142857	0.035714286	0.035714286	0.035714286	0.035714286	0.035714286			
LOG	1.041392685	0.865301426	1.643452676	1.643452676	0.798354636	1.643452676	1.643452676	1.041392685	0.689210167	1.041392685	1.166331422	1.166331422			
	0.970036777	0.970036777	0.970036777	0.970036777	0.84509804	0.84509804	0.970036777	1.447158031	1.447158031	1.447158031	1.447158031	1.447158031			
	0.004889285	0.012677078	0.27593674	0.27593674	0.002736811	0.387929839	0.27593674	0.15810128	0.833541048	0.15810128	0.067616788	0.067616788			
X^2	2.520914417														

「理論値 : HMM45 世界，標本値 : HMM45 試合」での場合，語数は 14，X 二乗値は 3.533982 である．この時，有意水準を 5 %とし自由度を 13 とすると X 二乗分布表により 22.36 である．

「理論値 : HMM45 世界，標本値 : ALHMM 世界」での場合，語数は 12，X 二乗値は 2.520914417 である．この時，有意水準を 5 %とし自由度を 11 とすると X 二乗分布表により 19.68 である．

どちらの分布も各臨界値以下であるため，ALHMM と全 45 件の HMM での状態「世界」と「試合」は本来類似した振る舞いをしていると見てよい．

6. 結　　論

本研究では，能動学習による有用なデータの選定方法を提案し，隠れマルコフモデルを用いた系列データの自動分類を行った．この能動学習を隠れマルコフモデルに組み込むことで，自律的にデータを選定する ALHMM を提案した．各クラスの特徴を抽出し分類器を構築したことにより未知データのクラス推定が可能となり，各クラスのマルコフモデルを図式化することで，系列データにはクラスを特徴づける構造が存在することが分かった．全 15 件の場合の HMM に比べて Largest-ALHMM の適合率は 24.0 %，再現率は 21.4 %向上した．

学習データ量別の HMM との比較では，全 42 件の場合以上の高い適合率・再現率，f 値を示した．従って，Largest-A L HMMでの少ない学習量 27 件で得た性能が、HMMでは学習データ量 42 件以上に相当し、極めて高い性能を示すことを本実験で確認した．

文　　献

[1] Anderson, B., Moore, A, ”Active Learning for Hidden Markov Models : Objective Functions and Algorithms”, ICML, 2005.

[2] Angluin, D., ”Queries and concept learning”, Machine Learning, 2, pp.319—342, 1988.

[3] Bilmes, J.A., ”A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models”, ICSI, 1998

[4] Bouguelia, M.R., Belad, Y., and Belad, A., ”A Stream-Based Semi-Supervised Active Learning Approach for Document Classification”, ICDAR, 2013

[5] D. Cohn, L. Atlas, and R. Ladner. , ”Improving generalization with active learning”, Machine Learning, 15(2), pp.201—221, 1994.

[6] Dredze, M. and Crammer, K., ”Active Learning with Confidence”, ACL-08, pp.233—236, 2008

[7] Dasgupta, S. and Langford, J., ”A tutorial on active learn-

ing", ICML, 2009

[8] Han, J., Kamber, M. and Pei, J., "Data Mining: Concepts and Techniques" (3rd ed.), Morgan Kaufmann, 2011

[9] Jurafsky, D., Martin, J.H., "Hidden Markov Model", Speech and Language Processing, 2016

[10] Lewis, D. and W. Gale, W., "A sequential algorithm for training text classifiers", ACM SIGIR Conference on Research and Development in Information Retrieval, 1994, pp. 3—12

[11] Settles, B., "Active Learning Literature Survey", Computer Sciences Technical Report 1648 University of Wisconsin Madison, 2010

[12] Zhou, J. and Sun, S., Improved Margin Sampling for Active Learning Communications in Computer and Information Science 483, 2014, pp.120-129