Vol.2014-MPS-101 No.10
2014/12/9

機械学習を用いた問題解答のための推論システムの開発

岩片悠里[†1]　川嵜美波[†1]　江口由記[†1]　石川由羽[†1]　高田雅美[†1]　城和貴[†1]

本稿では，問題文の解答を推測するシステムの開発について述べる．本システムでは，問題文の言語解析及び特徴ベクトルの抽出によって学習データの作成を行った後，機械学習を行い問題文のパターン分類を行うことによってテストデータから問題文の答えを推測することができる．機械学習には，教師あり学習を用いるべきである．多クラス判別を行う教師あり学習の代表的な手法に Learning Vector Quantization がある．Learning Vector Quantization では事前にクラス数を求めておく必要がある．クラス数を求める方法として，カーネル主成分分析及びギャップ統計量がある．ただし，カーネル主成分分析は視覚的に推定する方法であり，ギャップ統計量は，教師なし学習のための推定方法であるため，本システムには適さない．一方，Affinity Propagation はクラス数を事前に求める必要がないが，教師なし学習の手法である．そこで，本稿では，Affinity Propagation および Learning Vector Quantization を複合させ，クラス数の推定を行う教師あり学習の手法を提案する．

Development of Reasoning System for Answering Question using Machine Learning

YURI IWAKATA[†1]　MINAMI KAWASAKI[†1]
YUKI EGUCHI[†1]　YUU ISHIKAWA[†1]
MASAMI TAKATA[†1]　KAZUKI JOE[†1]

In this paper, we develop a system to guess the answer of the questions. In the system, once training data is created with language analysis and feature vectors of question statements, then a learning machine is performed for pattern classification. By using the system, the answer can be obtained. LVQ(Learning Vector Quantization) is known as a representative method of supervised learning for multi class classification. LVQ needs the number of classes as initialization. Kernel principal component analysis and gap statistic are methods for getting number of classes. Since kernel principal component analysis is a method for estimating visually and gap statistic is a method for unsupervised learning, there is not suitable for the proposed system. AP(Affinity Propagation),which is a method of unsupervised learning, is not necessary to determine the number of classes in advance. Hence, we propose a method for performed supervised learning to estimate number of classes through AP is improved using LVQ.

1. はじめに

現代社会では，情報技術の発達により，情報量が爆発的に増加している．この情報に含まれている知識を効率良く再利用するために，オントロジ[1]と呼ばれる知識ベースの構築がなされている．オントロジとは，コンピュータと人間が相互理解できるように対象世界を知識体系化したもののことである．構築されたオントロジを様々なシステムにおいて共有及び再利用する場合，その汎用性を評価する必要がある．しかしながら，規模の大きなオントロジの汎用性を評価するには，膨大な時間や労力がかかってしまう．そこで，オントロジの汎用性を自動評価するシステムが必要となってくる．

現在，我々は，専門書の知識を基にオントロジを自動構築している．研究の第一段階として，専門書には，最も簡単な専門書である小学校の教科書を使用している．本稿ではこの研究で用いる，教科書に準拠した問題集の問題文から答えを推測するシステムを提案する．本システムでは，問題文を言語解析し，その特徴量を抽出，機械学習により得られた結果をもとに答えの導出を行う．機械学習の学習データは，教科書に準拠した問題集の問題文と答えをもとに作成する．そのため，本システムで用いる機械学習では，教師あり学習を行うべきである．多クラス判別を行う教師あり学習の代表的な手法に Learning Vector Quantization (LVQ)[2] がある．LVQ では事前にクラス数を求めておく必要がある．クラス数を求める方法として，カーネル主成分分析[3]およびギャップ統計量[4]がある．カーネル主成分分析は，クラス数を視覚的に推定する方法であるため，本システムには適さない．ギャップ統計量は，教師なし学習のための機械的な推定方法である．同じ答えを持つ問題文毎にギャップ統計量を用いてクラス数を推定し，各々のクラス数の合計値を全体のクラス数とすることもできるが，正確な推測方法とは言えない．2007 年に提案された機械学習の手法に Affinity Propagation (AP)[5] がある．この手法では，クラス数を求める必要がない．ただし，教師なし学習を目的とした手法であるため，本システムに直接用いることはできない．そこで，AP と LVQ を複合させることでギャップ統計量の処理を無くし，プログラム内でクラス数の最適値を求める教師あり学習の手法を開発する．評価実験では，提案手法と，AP および LVQ をそれぞれ比較する．

†1 奈良女子大学
Nara Women's University

本稿では，2 章で機械学習の手法である AP および LVQ について述べる．3 章では，本稿の提案システムについて説明を行い，4 章にて，AP と LVQ それぞれと比較することによる本提案手法の評価実験について述べる．

2. 機械学習

本システムでは，機械学習に LVQ と AP を複合したものを用いる．本章では，LVQ および AP について述べる．

2.1 Learning Vector Quantization

LVQ は，Kohonen によって提案された教師あり学習を行う手法である．この学習法では，入力空間のパターン分類を行うための代表ベクトルの導出を行う．LVQ には，LVQ1, LVQ2，LVQ3，OLVQ の 4 種類が存在する．LVQ2 および LVQ3 は，一度に更新する代表ベクトルを LVQ1 の 2 倍にすることによって学習の効率性を高めた手法である．OLVQ は，代表ベクトルを更新する際に必要となる学習係数の最適化を行うことによって，学習の収束性を高めた手法である．本稿では，これら LVQ のうち最も単純なアルゴリズムで挙動を詳しく調べやすい LVQ1 を用いている．以下，LVQ1 について述べる．

学習データは，n個の入力$\boldsymbol{x_i}$およびラベルy_iから成る組である．学習データは，$\{(\boldsymbol{x_0}, y_0), \cdots, (\boldsymbol{x_n}, y_n)\}$のように表される．代表ベクトルも同様に$\{(\boldsymbol{m_0}, y_0), \cdots, (\boldsymbol{m_n}, y_n)\}$のように表される．ここで，$\boldsymbol{x_i} = (x_{i1}, \cdots, x_{ip}), \boldsymbol{m_j} = (m_{j1}, \cdots, m_{jp})$である．$p$は，ベクトルの次元数を表す．

LVQ の処理手順は以下のとおりである．

(1-1) 時刻tにおいて任意のデータ点$(\boldsymbol{x}(t), y(t))$を取得
(1-2) 最も近い代表ベクトル$\boldsymbol{m}_i(t)$を取得
(1-3) 代表ベクトルの更新

$(\boldsymbol{x}(t), y(t))$は，時刻$t$に選ばれたデータ点のことである．(1-1)~(1-3)の処理手順を条件を満たすまで反復を繰り返す．条件は，次の 2 点である．

- 代表ベクトルおよびそのクラスに属するデータ点に一定回数変化がない場合
- 反復回数が任意の最大反復回数に達する場合

手順(1-1)では，時刻tにおいて任意のデータ点$(\boldsymbol{x}(t), y(t))$を 1 つ取得する．手順(1-2)では，手順(1-1)で取得したデータ点と最も近い距離を持つ代表ベクトル$\boldsymbol{m}_i(t)$を取得する．手順(1-3)では，式(1)を用いて代表ベクトルの更新を行う．

$$\boldsymbol{m}_c(t+1) = \begin{cases} \boldsymbol{m}_c(t) + \alpha(t)\big(\boldsymbol{x}(t) - \boldsymbol{m}_c(t)\big), & y(t) = l_c(t), \\ \boldsymbol{m}_c(t) - \alpha(t)\big(\boldsymbol{x}(t) - \boldsymbol{m}_c(t)\big), & y(t) \neq l_c(t), \end{cases} \quad (1)$$
$$\boldsymbol{m}_i(t+1) = \boldsymbol{m}_i(t), \quad i \neq c$$

$\alpha(t)$は$0 < \alpha(t) < 1$を満たす学習係数であり，一般的に時間とともに小さく変化させる．反復終了後の代表ベクトルから入力空間のパターン分類を行う．

2.2 Affinity Propagation

AP は，2007 年に Brendan J.Frey と Delbert Dueck によって提案された機械学習の手法である．この手法では，初期値依存や学習順序への依存はない．また，クラス数を事前に与える必要はない．

AP では，学習データの中で，クラスの中心になるデータ点のことを exemplar と呼んでいる．初期の段階では，全てのデータ点を exemplar の候補とし，AP を行うことによって exemplar 候補の中から exemplar を求める． similarity はデータ点間の関係性を表し，データ点間の関係性が密接であれば，値を大きくし，密接でなければ，値を小さく設定しておく．通常，similarity には，負のユークリッド距離を設定する．

AP では，データ点と exemplar 候補との間で responsibility および availability という 2 種類のメッセージの交換を行う．ここで言うメッセージとは，評価値のことを指す．responsibility は，データ点iから exemplar 候補kに送られるメッセージのことで$r(i,k)$と表記し，exemplar 候補kがデータ点iにとって exemplar として適切である度合いを表す．availability は，exemplar 候補kから
データ点iに送られるメッセージのことで$a(i,k)$と表記し，データ点iが exemplar として exemplar 候補kを選ぶことが適切である度合いを表す．

処理の流れは，以下のとおりである．

(2-1) similarity を計算
(2-2) $a(i,k) = 0$に初期化
(2-3) 条件を満たすまで(2-3-i)~(2-3-iii)を反復
  (2-3-i) responsibility を算出
  (2-3-ii) availability を算出
  (2-3-iii) $r(i,k) + a(i,k)$の最大値から，exemplar 候補およびそれに属するデータ点を更新
(2-4) exemplar 候補を exemplar として終了

手順(2-1)では，similarity の計算を行う．similarity の計算は，式(2)のとおりである．ただし，$i = k$の場合，$s(i,i)$には，全データ点間の similarity の値の中央値を設定する．

$$s(i,k) = -\|x_i - x_k\|^2 \quad (2)$$

手順(2-2)では，availability の初期化を行う．availability が計算によって導出される前に，$r(i,k)$の導出過程で availability を使用するため，availability の初期値が必要となる．availability の初期値は，$a(i,k) = 0$である．

手順(2-3)では，(2-3-i)~(2-3-iii)の処理を次の条件が満たされるまで，反復させる．条件は，次の 2 点である．

- exemplar およびそのクラスに属するデータ点に一定回数変化がない場合
- 反復回数が任意の最大反復回数に達する場合

これらのどちらかが満たされたとき，手順(2-3)の反復は終了する．次に手順(2-3)の反復内容について説明する．手順(2-3-i)では，responsibility の算出を行う．算出方法は，式(3)のとおりである．k'は，exemplar 候補kと競合している exemplar 候補である．

$$r(i,k) = s(i,k) - \max_{k' s.t. k' \neq k} \{a(i,k') + s(i,k')\} \quad (3)$$

手順(2-3-ii)では，availability の算出を行う．算出方法は，式(4)および式(5)のとおりである．i'は，exemplar 候補kを最も高く評価しているデータ点である．

$$a(i,k) = min\left\{0, r(k,k) + \sum_{i' s.t. i' \notin \{i,k\}} max\{0, r(i',k)\}\right\} \quad (4)$$

$$a(k,k) = \sum_{i' s.t. i' \neq k} max\{0, r(i',k)\} \quad (5)$$

手順(2-3-ii) (2-3-iii)では，exemplar 候補およびそのクラスに属するデータ点の更新を行う．exemplar 候補およびそのクラスに属するデータ点は，次のように導出される．データ点iに対して$r(i,k) + a(i,k)$の値が最大となる exemplar 候補kが exemplar として決定される．その際，データ点iは，exemplar kのクラスに属するデータ点となる．

式(3)，(4)，(5)の計算結果は，振動する可能性があるため，次の式を用いて振動を防ぐ必要がある．λは減衰係数を表し，通常，$\lambda = 0.5$とされる．$r'(i,k)$は，$r(i,k)$の計算前のメッセージである．同様に$a'(i,k)$は，$a(i,k)$の計算前のメッセージである．

$$r(i,k) = \lambda \cdot r'(i,k) + (1-\lambda) \cdot r(i,k) \quad (6)$$

$$a(i,k) = \lambda \cdot a'(i,k) + (1-\lambda) \cdot a(i,k) \quad (7)$$

手順(2-4)では，反復終了後の exemplar 候補を exemplar として処理を終える．

3. 機械学習を用いた推論システム

本章では，本提案システムについての説明を行う．システムは，学習データから機械学習を行いクラスの境界線を取得するものおよび，テストデータから機械学習の結果を利用して問題の答えを推測するものの 2 つに分けられる．前者では，入力ファイルから問題文および答えを取得し，言語解析を行い，その特徴量を抽出し学習データの作成を行う．得られた学習データを用いて機械学習を行い，問題文のパターン毎にクラス分けを行い，クラスの境界線に関する情報を取得する．後者では，前者とは別の入力ファイルから問題文を取得し，言語解析を行い，その特徴量を抽出し，テストデータの作成を行う．得られたテストデータおよび機械学習の結果から得られたクラスの境界線をもとに問題の答えを推測する．

3.1 学習データからクラスの境界線を取得

本節では，学習データから機械学習を行いクラスの境界線を取得する手順について説明を行う．処理手順は，以下の通りである．

(3-1) 学習データとなる問題文および答えの取得
(3-2) 言語解析
(3-3) 特徴ベクトルの抽出
(3-4) 機械学習

手順(3-1)では，ファイルから問題文および答えの取得を行う．システムは，それらを題問毎にグループで分けて取得する．

手順(3-2)では，問題文の言語解析を行う．システムは，言語解析によって，問題文毎に名詞と動詞を取得する．本システムでは，言語解析に既存の形態素解析エンジンを使用する．

手順(3-3)では，問題文毎に特徴ベクトルの抽出を行う．特徴量の抽出には，TF-IDF 法[6]を用いる．TF-IDF 法は，単語の出現頻度および出現範囲から文書内の単語の重要性を測る方法である．TF-IDF 法は以下のように定義される．

$$w(wrd,d) = tf(wrd,d) \cdot idf(wrd) \quad (8)$$

$$idf(wrd) = log\frac{N+1}{df(wrd)} \quad (9)$$

$w(wrd,d)$は特徴量を表す．$tf(wrd,d)$は，文書dの中での単語wrdの出現回数を表す．また，$N+1$は全文書数，$df(wrd)$は，全文書数のうち単語wrdが出現する文書数を表す．こ

	wrd_0	wrd_1	wrd_2	…	wrd_{Nt}
問題文 0	$\boldsymbol{x_0}$ = (0.0 ,	0.2 ,	0.0 ,	… ,	0.0)
問題文 1	$\boldsymbol{x_1}$ = (0.1 ,	0.2 ,	0.0 ,	… ,	0.0)
⋮		⋮			
問題文 N	$\boldsymbol{x_N}$ = (0.1 ,	0.0 ,	0.0 ,	… ,	0.0)

図 1　特徴ベクトルの作成

れらの式を本システムに適応する際，文書dを問題文と設定してしまうと，文章が短すぎて 2 回以上同一単語が出現することがなくなってしまい，重要単語およびそうでない単語との間で$tf(wrd,d)$値に差が生じない．そのため，本システムでは，文書dを題問と設定し，$tf(wrd,d)$を同一の題問中での単語wrdの出現回数を表すものとする．$N+1$は全問題数，$df(wrd)$は，全問題数のうち単語wrdが出現する問題数と設定する．得られた特徴量から図 1 のような特徴ベクトルを作成する．図中の$\boldsymbol{x_i}$は特徴ベクトルを表し，wrd_iは全問題文中に出現する単語を表している．N_t+1は全問題文中に出現する単語数である．

手順(3-4)では，手順(3-3)で得られた特徴量をもとに機械学習を行う．LVQ および AP を複合させたものをここで用いる機械学習として提案する．本機械学習は AP の処理を基とし，(3-4-3-iii)で LVQ の代表ベクトルの更新と類似した処理を加えている．手順(3-3)で得られた問題文毎の特徴ベクトルおよびラベルの組を$\{(\boldsymbol{x_0},y_0),(\boldsymbol{x_1},y_1),\cdots,(\boldsymbol{x_N},y_N)\}$とし，これを学習データとして以下の処理を行う．ここで言うラベルとは，問題文の答えのことであり，$N+1$は全問題数を表す．

(3-4-1)　similarityを計算
(3-4-2)　$a(i,k)=0$に初期化
(3-4-3)　条件を満たすまで(i)~(iii)を反復
　(3-4-3-i)　responsibility を算出
　(3-4-3-ii)　availability を算出
　(3-4-3-iii)　任意のデータ点および最も近い exemplar のラベルに応じて availability の値を修正
　(指定した回数 (iii)を反復)
　(3-4-3-iv)　$r(i,k)+a(i,k)$の最大値から exemplar候補およびそれに属するデータ点を更新
(3-4-4)　exemplar 候補を exemplar として格納

手順(3-4-1)では similarity の計算を行う．計算方法は，負のユークリッドノルムを使用する．ただし，データ点同士のラベルが異なる場合は，負の最大値を格納する．ただし，$i=k$の場合，$s(i,i)$には，負の最大値を持つ similarity 以外の全データ点間の similarity の値の中央値を設定する．計算式は次のとおりである．

$$s(i,k)=\begin{cases}-\|x_i-x_k\|^2 & (y_i=y_k)\\ -\infty & (y_i\neq y_k)\end{cases} \quad (10)$$

手順(3-4-2)の availability の初期化は 2.2 節の手順(2-2)と同様である．また，手順(3-4-3) の条件，手順(3-4-3-i) responsibility 算出，手順(3-4-3-ii)の availability を算出も 2.2 節の手順(2-3)および手順(2-3-i)，手順(2-3-ii)と同様である．手順(2-3-iii)では，任意のデータ点iを取得し，データ点iと最も近くにある exemplar 候補kのラベルが同一であれば，$a(,k)=a(,k)+\beta$とし，同一でなければ，$a(,k)=a(,k)-\beta$とする．βは任意の定数とする．これを指定した回数反復させる．手順(3-4-3-iv)の exemplar 候補およびそれに属するデータ点の更新は，2.2 節の手順(2-3-iii)と同様である．手順(3-4-4)では，反復終了後の exemplar 候補を exemplar としてその情報を格納しておく．格納された exemplar は，3.2 節の手順(4-4)で用いる．

3.2 テストデータから答えの推測方法

本節では，テストデータから機械学習の結果を利用して問題の答えを推測する手順について説明を行う．以下に手順を示す．

(4-1)　テストデータとなる問題文の取得
(4-2)　言語解析
(4-3)　特徴ベクトルの抽出
(4-4)　答えの推測

手順(4-1)では，ファイルから問題文の取得を行う．システムは，それらを題問毎にグループで分けて取得する．

手順(4-2)では，言語解析を行う．処理内容は，3.1 節と同様である．そのため，手順(3-2)で用いる形態素解析エンジンを利用し，問題文毎に名詞と動詞を取得する．

手順(4-3)では問題文毎に特徴ベクトルの抽出を行う．まず初めに特徴量の抽出を行う．3.1 節の手順(3-3)と同様に TF-IDF 法を用いる．$tf(t,d)$は，手順(4-1)で取得した問題文から導出する．ここで，$idf(t)$も手順(4-1)で取得した問題文から導出してしまうと，IDF 法では局所的に出現する単語を重要単語とみなすため，出題分野に偏りがあった場合に学習データと同じ尺度で特徴ベクトルが作成されない可能性がある．例えば，学習データを理科の全範囲の問題から作成したとする．これに対して，テストデータを「植物の光合成」に関する分野の問題のみから作成すると，全問題数のうち「植物」という単語が出現する問題数の割合は確実に高くなる．その結果，式(9)の$idf(t)$の値が低くなるため「植物」という単語のテストデータでの重要度が学習データの重要度より低くなる．特徴が学習データとテス

トデータで異なるため，正しい答えを推測できない可能性が高い．そこで，$idf(t)$は，3.1 節の結果を取得し，これを利用する．このことを考慮して，式(8)および式(9)を用いて特徴量を抽出する．特徴量から特徴ベクトルへの変換方法は 3.1 節と同様である．ただし，wrd_iは互換性を保つため 3.1 節のwrd_iをもとに特徴ベクトルの作成を行う．

手順(4-4)では，答えの推測を行う．答えの推測は，テストデータのデータ点iに最も近いユークリッドノルムを持つ exemplar のラベルを問題文iの答えとする．

4. 評価

本章では，本システムの性能を確かめるための評価実験の内容について述べる．本評価実験では，まず初めに，本稿で提案している機械学習で用いるβおよび(3-4-3-iii)で述べた反復回数をitsを変化させ，その時の正答率を考察する．次に，βおよび反復回数をitsに実験結果から得られた最適値をあてはめ，本提案手法，AP，LVQ の比較を行う．ただし, LVQ は, 予めクラス数を指定しなければならないため，LVQ を行う前にギャップ統計量を行いクラス数の推定を行うものとする．ギャップ統計量とは，k-means を用いてクラスタ数を推定するための手法である．ただし，教師なし学習のためのクラス数の推定方法である．そのため，同じ答えを持つ問題文毎にギャップ統計量を用いてクラス数を推定し，各々のクラス数の合計値を全体のクラス数とする．学習データおよびテストデータの作成には，小学校の理科の教科書に準拠した問題集の問題文とその答えを用いる．この際，問題形式が多様であるため，学習データとして採用する問題には以下の制限を与える．

- 答えの形式が名詞かつ 1 単語の問題に限定
- 穴埋め問題は，空欄に「何」の文字を挿入
  (ただし，穴埋め箇所が複数あるものは除去)
- 前後の問題の知識を必要とする問題は除去
- イラストを見て答える問題および計算問題，2 択問題は除去

これらの制限を満たす問題は，文理[7]が出版している教科書ワーク 6 年理科の教育出版版から 57 問, 同じく東京書籍版から 70 問，学校図書版から 80 問，大日本図書版から 81 問, 啓林館版から 83 問の合計 371 問ある．このうち教育出版版の 57 問および東京書籍版の 70 問，学校図書版の 80 問，大日本図書版の 81 問，合計 288 問を学習データとして使用する．また，啓林館版の 83 問をテストデータとする．本システムの言語解析には既存の形態素解析エンジンである MeCab[8]を使用する．

提案手法におけるβおよびitsに関する実験結果を，図 2 および図 3 に示す．図 2 は，本稿で提案している機械学習

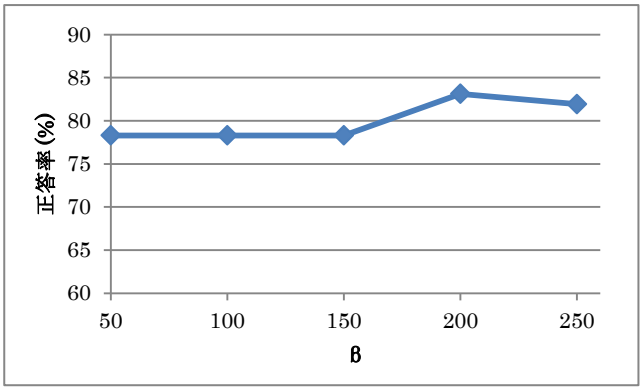

図 2　βと正答率の関係

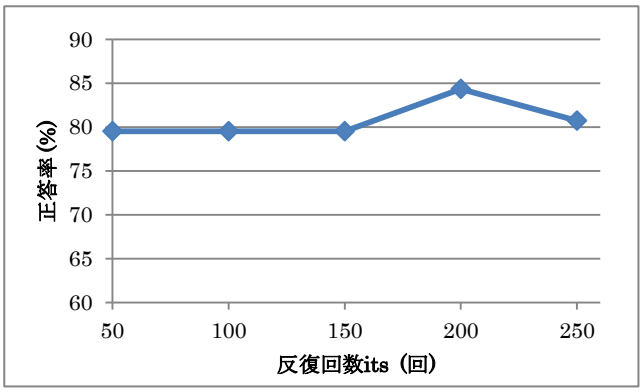

図 3　反復回数itsと正答率の関係

で用いられている変数であるβと正答率の関係をグラフで表したものである．横軸がβ，縦軸が正答率を表し，βを 50 から 50 ずつ増加させている．このとき，手順(3-4-3-iii)で述べた反復回数をitsとし，$its = 200$とする．また，手順(3-4-3)で述べた最大反復回数を$maxits$とし，$maxits = 100$としている．図 3 は，同じく反復回数itsと正答率の関係をそれぞれグラフで表したものである．横軸がits，縦軸が正答率を表し，itsを 50 から 50 ずつ増加させている．このとき，$\beta = 200$，$maxits = 100$としている．図 2 および図 3 より最適値は，β =200，$its = 200$であることがわかる．図からβ，itsともに増減の傾きは，激しく上下を繰り返すこともなく，緩やかで安定しているといえる．また，正答率の振れ幅は共に約 5%であり，両者の正答率に与える影響力は同程度だと推察される．

次に，提案手法，AP，LVQ で実行した結果を図 4 に表す．横軸が最大反復回数maxits，縦軸が正答率を表し，maxitsを 100 から 25 ずつ増加させている．ここで，提案手法の変数の設定は, 図 2 および図 3 の結果より, $\beta = 200$, $its = 200$とする．**エラー! 参照元が見つかりません。**は，最大認識率とその時のmaxitsを表したものである．

本評価実験の結果より，本稿で提案している機械学習は，最大反復回数maxitsが 200 回のときに最大認識率 84.3%を

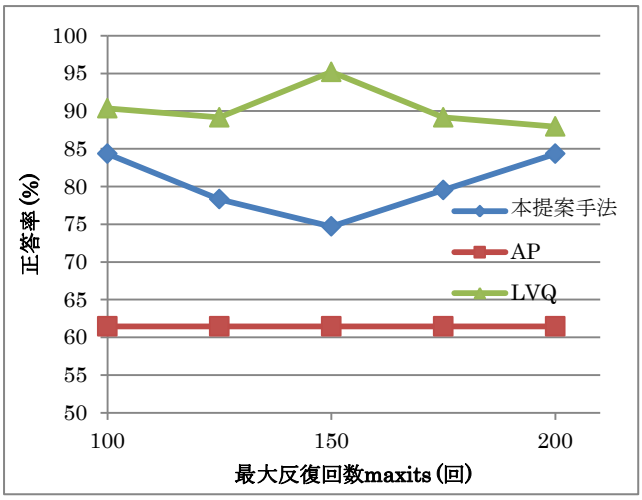

図 4　提案手法，AP，LVQ の比較

表 1　最大反復回数と最大認識率の関係

	最大認識率 (%)	最大反復回数 (回)
提案手法	84.3	100
AP	61.4	100
LVQ	95.2	150

導出することが分かる. また, AP は, 最大反復回数$maxits$を変化させても 100 回から 200 回の間では変化が見られず，最大認識率 61.4%である．LVQ は，最大反復回数$maxits$が 150 回のときに最大認識率 95.2%となることがわかる．AP の認識率が他と比べて低いのは，教師なし学習である AP を教師あり学習に用いたためであると考えられる．これらの結果，本稿で提案している機械学習は，84%という高い確率で問題文に対して正しい答えを推測している．LVQ を用いる手法では，ギャップ統計量によるクラス推定が上手くいくとは限らない．一方，クラス数を学習の途中で取得することができる提案手法では，常に安定した正答率を得ることが可能であるものと思われる．

以上より，提案手法を用いた推論システムにおいて，適切なパラメータは，β =200，its = 200であり，問題文に対する正答率は 84%であるため，有用であると考えられる．

5. まとめ

本稿では，問題文の解答を推測するシステムの開発をした．本システムは，学習データから言語解析および特徴ベクトルの抽出を行った後，機械学習を行い問題文のパターン分類を行う．さらに，テストデータから機械学習の結果を利用して問題文の答えの推測を行う．多クラス判別を行う教師あり学習の代表的な手法に LVQ がある． LVQ では事前にクラス数を求めておく必要がある．クラス数の推定方法であるカーネル主成分分析およびギャップ統計量は，本システムには適さない．一方，AP はクラス数を事前に求める必要はないが, 教師なし学習の手法である. そこで，本稿では，AP 及び LVQ を複合させ，クラス数の推定を行う教師あり学習の手法を提案し，本システムで用いた．

評価実験では，提案手法のパラメータチョイスに関する実験を行い，その後，提案手法，LVQ，AP の比較によって提案手法の評価を行った．パラメータチョイスに関する実験結果から，適切なパラメータは，β =200，its = 200であることがわかった．評価実験の結果，本稿で提案している機械学習の最大認識率は 84%であり，問題文に対して高い確率で正しい答えを推測していることが分かった．LVQ を用いる手法では，ギャップ統計量によるクラス推定が上手くいくとは限らない．一方，クラス数を学習の途中で取得することができる提案手法では，常に安定した正答率を得ることが可能であるものと思われる．このことを考慮に入れると，本稿で提案している機械学習は有用であるといえることが確認された．

参考文献

1) 小林一郎：人工知能の基礎，サイエンス社(2008).
2) A WEB Page,Helsinki University of Technology Laboratory of Computer and Information Science Neural Networks Research Center(online),
available from<http://www.cis.hut.fi/research/lvq_pak/>(accessed 2014-11-06).
3) Schölkopf,B.,Smola,A. and Muller, K.: Kernel Principal Component Analysis,Advances in Kernen Methods, pp.327−353 (1998).
4) R.Tibshirani,G.Walther,and T.Hastie:Estimating the number of clusters in a dataset via the gap statistic, Journal of the Royal Statistical Society,Series B,pp.411-423(2001).
5) Brendan J.Frey,Dellbert Dueck:Clustering by Passing Messages Between Data Points, Science, Vol.315,pp972-976(2007).
6) Salton,G.and Buckley, C.:Termweighting approaches in automatic text retrieval, Information Processing and Management,Vol.24,No.5,pp.513-523(1988).
7) A WEB Page,文理(online), available from< http://www.bunri.co.jp/ >(accessed 2014-11-06)
8) A WEB Page,Kyoto University(online), available from< https://code.google.com/p/mecab/ >(accessed 2014-11-06)