Vol.2017-CVIM-205 No.20
2017/1/19

# 多分決定木を用いた識別と類似画像検索に関する研究

松尾大典† 前田啓† 和田俊和†

**概要**：決定木は，識別や回帰に用いることのできるデータ構造であり，ランダムフォレストの構成要素としても有用である．実数ベクトル集合を学習サンプルとして，決定木を構築する際には，1）射影計算と2）射影成分の閾値処理によるデータ分割法，の二つを定める必要がある．また，3）データの分割と同時に次元圧縮を行う方法も考えられる．本報告では，上記の観点から，数種類の多分決定木の構築法を提案し，1）射影軸を主成分分析と判別分析で決定する場合，2）閾値を等確率分割とエントロピー・ゲイン最大化で決定する場合，3）子ノードの射影軸計算の際に親の射影軸の補空間に射影したデータ集合を対象とする場合とそうでない場合，の組み合わせ合計8通りについて性能の比較を行う．また，Vocabulary Tree の代わりに多分決定木を用いた類似画像検索の実験も行い，計算量と精度についても比較を行う．

**キーワード**：決定木，次元圧縮，データ分割，類似画像検索

# Multiple decision tree for classification and image retrieval

DAISUKEMATSUO† KEI MAEDA†
TOSHIKAZU WADA†

***Abstract***: Decision tree is a data structure that can be applied to classification and regression problems. It is also highlighted as a component of decision forest. On the construction of decision tree for real vectors, we at least have to determine the 1) projection axis and 2) data segmentation rule on the projection axis. Additionally, 3) dimensionality reduction can be applied with the data segmentation. In this report, we examine the performance of multiple decision trees with different combinations of above factors 1-3. That is, 1) projection axis is determined by principal component analysis or discriminant analysis, 2) data segmentation is based on equiprobability or entropy gain maximization, 3) on the data segmentation, segmented data are projected on the complementary space of the projection axis or not. Comparative experiments are performed on several datasets for eight decision trees with different combinations of conditions mentioned above. Also, we performed experiments on similar image retrieval by exchanging the vocabulary tree with a projection based decision tree and compared their accuracy and computational complexity.

***Keywords***: decision tree, dimension reduction, data partition, image retrieval

## 1. はじめに

本報告では，射影に基づく多分決定木の構築法について1）射影軸の決定法，2）射影軸上でのデータ分割法，3）データ分割を行う際に次元圧縮を行うか否か，という3つの観点から検討した結果について述べる．

決定木は，識別や回帰に用いることのできるデータ構造であり，ランダムフォレストの構成要素としても用いられる．ただし，ラベル付きの実数ベクトル集合に対して決定木を構築する際の自由度は高く，最適な決定木を求めることは一般に困難な問題である．このため，フォレストを作成する場合などには,ランダムに分岐パラメータを決定し，分割されるデータに付けられたクラスラベルのエントロピーが最も小さいものを採用するといった方式が採用されている[1].

決定木のノードでは，学習データの分割が行われるが，これは各学習データ（ベクトル値）から何らかの方法でスカラ値を算出し，これらに対する閾値処理によって実現される．ベクトル値から，スカラ値を求める計算としては，固定ベクトルへの射影計算，または固定ベクトルとの距離計算が考えられる．最近傍探索におけるバイナリーコーディングなどでは，ランダムベクトルへの射影を閾値処理するよりも，固定ベクトルとの距離を閾値処理した方がより良い最近傍探索が行えるといった報告[5]もあるが，決定木では，現在の所射影計算が主流である．

射影計算を行う場合，射影軸をどのように決定するかは自明な問題ではない．決定木の場合はクラスラベルが利用できるため，判別分析により射影軸を決定することができる．また，クラスラベルを調べることなく，主成分分析によって射影軸を決定することもできる．これらいずれの方法も，識別率を向上させるという最終目的と完全に合致した計算ではないため，いずれが良いかは自明ではない．

ベクトルへの射影を行った後，閾値処理を行うが，そのノードでデータ集合を分割する前と後でクラスラベルの純度を評価し，純度が最大の分割を採用するという方式がよく用いられる．純度の評価方法としては，エントロピーが最も低下する（エントロピー・ゲインが最大化される）という基準や，ジニ係数が最も大きくなるといった基準が用

† 和歌山大学院システム工学研究科
Wakayama University, Faculty of Systems Engineering

いられるが，エントロピー・ゲインを用いる方法が主流である．その一方で，ラベルの純度を評価せず，射影軸上のデータを等確率に分割する方法も考えられる．

これら，射影軸の決定方法と，射影されたデータを分割する方法の組み合わせで，ほぼ決定木の特性が決まる．但し，これだけでは与えられた学習データに過剰に適合しすぎて，未知データに対する誤りが生じる可能性もある．このような場合への対処法として，データの分割を行った際に，射影軸の成分を除去してから再度射影軸を計算するといった次元圧縮を行う方法がある[6]．この結果，根から葉に至る射影軸は Gram-Shmidt の直交化をしたことになり，木の高さは次元数以上にはならない．これによって，データの過剰な分割が行われなくなる．

このような次元圧縮を行う場合に，データを十分細かく分割するためには，2 分木の構造では分割が十分進まない．このため，多分木の構造をした決定木についても検討する必要が生じる．

以上の通り，射影計算による決定木の構築方法は主に，１）射影軸の計算法，２）射影されたデータの分割基準，３）データ分割後に次元圧縮を行うか否か，の３つに依存しており，さらに木の分岐数も変化させて評価を行う必要がある．

本報告では，このような考えに基づき，１）射影軸の計算法として判別分析と主成分分析の２種類，２）データの分割としてエントロピー・ゲインを最大化する分割とデータが等確率に各ノードに流れる分割の２種類，３）次元圧縮を行う場合と行わない場合の２種類，の合計８通りの条件で構築した決定木を用いて，様々なデータベースおよび，分岐数で識別性能を評価した結果について述べる．

また，SIFT や SURF などの局所特徴を用いた画像の類似検索で用いられる David Nister が提案した Vocabulary Tree[8]は，各画像をクラスと捉えれば，局所特徴を個々の画像に識別する決定木と見なすこともできる．Vocabulary Tree は k-means クラスタリングを再帰的に適用して作成された木構造であるため，木を辿る際には，分岐の数×木の高さ（回）の距離計算が必要になる．これに対して，射影計算の場合には，各階層で 1 回の射影計算しか行わないため，木の高さ（回）の内積計算だけで済む．内積計算と距離計算は同等のコストであるため，分岐の数だけの高速化が果たせるはずである．この観点から，Vocabulary Tree を射影に基づく決定木に置き換えた場合の識別精度と速度がどのようになるかについても検討を行う．

以下，第 2 章では関連研究，第 3 章では提案手法について，第 4 章で実験，第 5 章で実験の結果を示し，第 6 章でまとめを述べる．

## 2. 関連研究

本報告では，決定木の構築に際して射影軸と分割面の決定法，次元圧縮の有無に着目して性能の比較を行う．

画像から抽出された高次元の局所特徴などのデータに対して識別や最近傍探索を行う課題に対して，ランダムフォレスト[3]がしばしば適用される．ランダムフォレストは，ランダムにサンプリングされた学習サンプルに対して構築した決定木を複数用いて識別や回帰の計算を行う手法であり，複数の決定木を用いることでバイアスや分散を低減させ，汎化性能の向上を図るという手法である．

この手法では，

- 決定木の分岐ノードの候補を複数作成し，それらのノードでトレーニングデータが分割された後のクラスラベルのエントロピーが最も低下する候補を選ぶ．

という手法が一般的に用いられる．分岐ノードのパラメータはランダムに選ばれることが多く，候補数の大小によって，ランダムさを調整することができる．

これらの手法では，エントロピー・ゲインさえ大きければ，ランダムに生成した決定木でも，ある程度の性能は担保され，木の間の相関が低ければ過適合は起きにくくなると言われている．

しかし，実ベクトルに対する決定木の分岐ノードには数多くのパラメータが存在し，構築方法も複数の可能性がある．このため，ランダムにパラメータを決定するだけでは，決定木として十分な性能を出すことは困難である．特に学習サンプルの個数に著しい偏りがある場合には，汎化と分化のバランスがとりにくい．

このような問題意識に基づき，過去の研究を俯瞰すると，射影軸の決定や，射影後のデータの分割，さらに次元圧縮などで用いることのできる手法が幾つか存在する．

射影軸としては，以下のものが考えられる．

- クラスラベルが利用できない場合には，PCA tree[3]や主成分木[6]などで用いられる主軸（共分散行列の第一固有ベクトル）
- クラスラベルが利用できる場合には，判別分析で得られる判別軸（$\Sigma_W^{-1}\Sigma_B$の第一固有ベクトル）

決定木構築の際には，判別軸が有利であると考えられるが，クラスラベル数が多い問題では，唯一の判別軸が決定できないため，主軸と判別軸の優劣は付けがたい．

空間分割の手法としては PCH[7]で用いられた射影後のデータの等確率分割や，Grafting Trees[2]やランダムフォレストのようにエントロピー・ゲインで分割位置を決定する方法もある．

さらに，主成分木[6]で採用されているように，空間分割と次元圧縮を反復しながら木を作成する方法もある．

これら１）射影軸の計算法，２）射影されたデータの分割基準，３）データ分割後に次元圧縮を行うか否か，についてはこれまでにも個別に検討されてきたが，全体としてどのような組み合わせがよりよい結果をもたらすのかにつ

いては検討がなされてこなかった．さらに，決定木は二分木として構築されるケースが多いが，射影に基づく決定木の構築では，多分木にすることも可能であり，前述の等確率分割もエントロピー・ゲイン最大化分割を使って多分決定木を構築することもできる．

また，多分木の決定木は，類似画像検索に用いられるVocabulary tree[8]の代替としても利用可能であり，単に識別のタスクだけでなく，画像検索のタスクに於いても精度と速度両方の評価が可能である．

以上，関連研究について述べたが，その詳細について以下に説明する．

## 3. 射影に基づく分岐を行う決定木の構成法

本報告では，射影計算による決定木の構築における複数の手法の組み合わせについて考える．射影計算による決定木の構築方法では主に，１）射影軸の計算法，２）射影されたデータの分割法，３）次元圧縮を行うか否か，の３つの手順で複数の手法を適用することが考えられる，それぞれの手順と用いる手法について以下で詳しく説明する．

### 3.1 射影軸の決定法

射影軸の決定法としては，前章で述べた主成分木やPCHと同様に主成分分析により求める方法と，データに付与されたクラスラベルからクラスの集合を最もよく分割する軸を求める判別分析の２つを用いる．各手法について以下で詳しく説明する．

#### 3.1.1 主成分木

主成分木とは，射影軸の決定と空間分割，次元圧縮を反復し2分木を作成し，最近傍探索などに応用される手法である．主成分木では図1に示すように，まずデータのバラつきを表す軸として主成分分析により求めた第一固有ベクトルを求め，これを射影軸とし，次にデータと射影軸との内積からデータを1次元に射影し，射影結果の重心位置を閾値としてデータを二等分する．最後に分割後のデータを主軸の補空間へ射影する．これらの処理を反復することでGram-Schmidt の直交化と同じ効果から親子間のノードで射影軸が直交する2分木を作成する．直交化を行うので射影軸を基底とすると次元数と同じ木の高さでデータを十分に表現することが可能となっている．しかし，主成分木では2分木を構築するため木が高くなり，木をたどる時にコストがかかるという問題がある．この問題は2分木から多分木に切り替え，木の高さを抑える事で解決できると考えられる．

#### 3.1.2 判別分析

主成分木ではデータのバラつきを最もよく表す主軸として主成分分析により求めた第一固有ベクトルを射影軸として用いたが，データの分布ではなく，データに付与されたクラスラベルを考慮し，クラスのグループを最もよく分割する軸を求め，それを射影軸として用いることが考えら

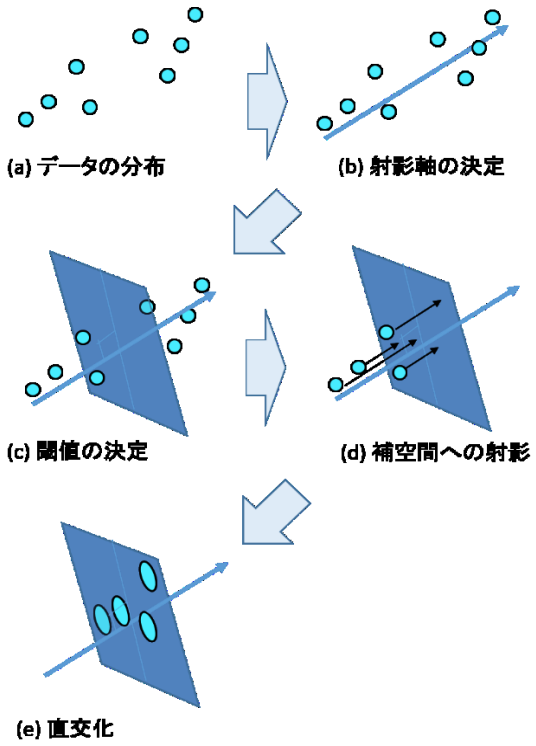



図 1 主成分木の作成手順

れる．クラスラベルを考慮した主軸を求める方法として判別分析が挙げられる．

クラス数を$C$，クラス平均を$\bar{\boldsymbol{x}}_{\boldsymbol{i}}(\mathbf{i}=\mathbf{1}\ldots C)$，クラスの事前確率を$P(\varpi_i)(\mathrm{i}=1\ldots C)$とするとクラス間分散$\Sigma_{\mathrm{b}}$とクラス内分散散$\Sigma_{\mathrm{w}}$はそれぞれ

$$\Sigma_{\mathrm{b}}=\sum_{i=1}^{C}P(\varpi_i)\Sigma_i$$

$$\Sigma_{\mathrm{w}}=\sum_{i=1}^{C}\sum_{j<i}P(\varpi_i)P(\varpi_j)\left(\bar{\boldsymbol{x}}_{\boldsymbol{i}}-\bar{\boldsymbol{x}}_{\boldsymbol{j}}\right)\left(\bar{\boldsymbol{x}}_{\boldsymbol{i}}-\bar{\boldsymbol{x}}_{\boldsymbol{j}}\right)^{\mathrm{T}}$$

により求められる．判別分析はこれを$\Sigma_{\mathrm{w}}^{-1}\Sigma_{\mathrm{b}}$としたときの固有ベクトルを求め，それを主軸としてクラスのグループを最もよく分割する分割面を考える手法である．

#### 3.1.3 Power Iteration

本報告では射影軸の決定データのバラつきを最もよく表す軸のみを用いるため，第一固有値に対応する固有ベクトルのみを求めれば良い，そこで Power Iteration という手法を用いる．この手法は，例えば主成分分析の場合では，まずデータの共分散行列Σを求め，これを線形変換とみなして任意のベクトル$\boldsymbol{b}$と掛け合わした結果を正規化する計算を繰り返すだけで最大固有値に対応する固有ベクトルを導くことができる．

$$\boldsymbol{b}_{\mathrm{k+1}}=\frac{\Sigma\boldsymbol{b}_k}{\|\Sigma\boldsymbol{b}_k\|}$$

この手法を用いることでメモリ使用や計算量の無駄を回避し，かつ高速に第一固有ベクトルのみを計算することがで

きる． また，判別分析においても，$\Sigma_w^{-1}\Sigma_b$を$\Sigma$とおくことで主成分分析と同様に Power Iteration を用いることで第一固有値に対応する固有ベクトルを高速に求める事で射影軸が決定できる．

### 3.2 射影軸上でのデータ分割法

3.1 節で述べた射影軸の決定法により求めた射影軸を用いて，データと射影軸との内積により射影した結果からデータを分割する方法について述べる．決定木の探索においては，木の分割数をnとすると閾値は$\theta_i(i = 1, \dots, n-1)$，求めた主軸を$\boldsymbol{\varphi}$，クエリデータを$\boldsymbol{x}$とすると木の探索時には

$$\theta_{i-1} \leq \boldsymbol{\varphi}^{\mathrm{T}}\boldsymbol{x} < \theta_i$$

により閾値との比較を行い，たどる子ノードを決定する．分割面の決定では閾値$\theta$を決定する処理を行う．手法としては等確率分割と，エントロピー・ゲインを最大化する分割の２つの手法を用いる．各手法について以下で詳しく説明する．

#### 3.2.1 等確率分割

等確率分割は PCH で用いられた空間分割の手法である．等確率分割は図に示すように，データ全体の分布が正規分布に従うと仮定し，データを射影軸上に射影して累積確率を調べ，これを等間隔に分割する事で得られる．PCH では各区間のデータの個数が一定なのでクエリに対する区間が必ず存在する事，射影結果から容易に距離計算の打ち切りを行えることから効果的な最近傍探索手法であることが示されていた．

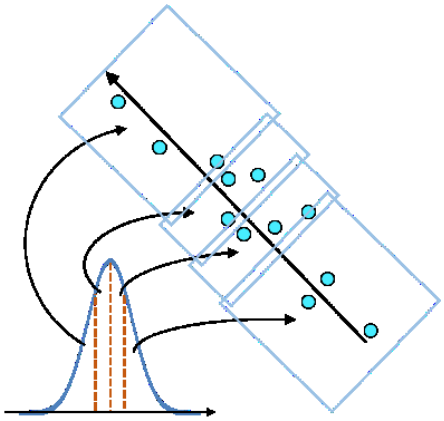


図 2 等確率分割の模式図

また実際のデータでは，全体の分布が正規分布に従わない場合も考えられるが，軸に射影したデータの累積度数から，これを等分割することで分割個数が一定に定まらない問題を回避できる．

また，決定木の作成では，ノードの親子間でのエントロピー・ゲインを最大にするようなデータの分割をする手法が多く取られる．ここで，等確率分割をエントロピー・ゲインを最大化する分割に用いる拡張が考えられる．次にこのエントロピー・ゲインを最大化する分割について説明する．

#### 3.2.2 エントロピー・ゲインを最大化する分割

エントロピー・ゲインを最大化する分割は木の親子間でエントロピー･ゲインが最大となる分割をする手法である．エントロピー・ゲインを最大化する分割では，目的の分割数の定数倍に等確率分割を行い，過分割した結果の各区間で，隣接する区間同士でマージした場合の局所的なエントロピーを計算する．エントロピーの計算結果から最もエントロピーの低い組み合わせを選択しマージを行う．以上の操作を，過分割した分割数が目的の分割数に落ち着くまで反復する．反復した結果として，分割後のエントロピーを最小化することで，親子間でのエントロピー・ゲインを最大化する．以上の拡張を本研究ではエントロピー・ゲインを最大化する分割と呼称する．

本報告における射影軸の決定法としては，等確率分割とエントロピー･ゲインを最大化する分割の 2 種類を用いる．エントロピー・ゲインを最大化する分割は等確率分割よりも区間の大きさと保持するデータ数に偏りが出るが，早期にクラスの純度の高いノードを切り分けられ，より少ない属性で決定木が作成されることが期待できる．かつ，本報告では述べていないが過分割数を決定する定数を調整することで訓練データとの適合を調整できることが考えられる．

### 3.3 軸の補空間への射影の有無

軸の補空間への射影の有無について説明する．データの分割時には，主成分木と同様に射影軸の補空間へデータを射影する処理を行う場合が考えられる．主軸として求めたベクトルを$\boldsymbol{\varphi}$，射影するベクトルを$\boldsymbol{x}$とすると，補空間への射影は

$$\boldsymbol{x}' = \boldsymbol{x} - \boldsymbol{\varphi}^{\mathrm{T}}\boldsymbol{x}\boldsymbol{\varphi}$$

の計算を行うことで対象のベクトル$\boldsymbol{x}$からベクトル$\boldsymbol{\varphi}$の成分を引くことができる．この処理により Gram-Shmidt の直交化と同様の効果があり，子ノードで求めた射影軸は親ノードの射影軸と直交することになる．ただし兄弟ノード間では射影軸の直交は保証されない．評価する軸同士が直交する事で決定木の根から葉にかけて異なる分布の特徴を評価することが可能になり，木の高さが次元数分だけあればデータを十分に表現できると考えられる．また，残差$\boldsymbol{\varphi}^{\mathrm{T}}\boldsymbol{x}\boldsymbol{\varphi}$を保存しておくことで元のデータを再現する事も可能であり，木をたどったクエリの元データとの比較をすることもできる．提案手法では補空間への射影を行う場合と行わない場合の両方の結果を求め，クラス識別間違い数の差を確認する．

また，本報告では決定木の類似画像検索への適用を考え，適用した場合の精度と速度について評価する．類似画像検索手法の一つとして Vocabulary Tree を挙げる．次に Vocabulary Tree についての説明をする．

### 3.4 Vocabulary Tree

Vocabulary Tree は K-means により反復してクラスタリングを行うことで階層的なコードブックを作成する手法であ

る. まず全局所特徴の集合を少数のクラスタ数に分割する．クラスタリングによって出来た集合に対して，再度クラスタリングを行う．これを繰り返し行うことで階層的なコードブックが構築される．クラスタリングの回数は多いが，各階層において少数のクラスタ中心に対し最近傍探索を行うため計算量を抑えることができる．また Vocabulary Tree では，探索時に通過した各ノードに投票していく．これにより探索結果が本来目的とする場所でなくとも，投票の結果から結果が反省でき，探索の誤りによる検索結果への影響を減らすことができる．

本報告では，木の作成時における複数の処理の組み合わせを考慮した多分木型の決定木の構築法を考える．

決定木構築手法の主な手順は

1). 射影軸の決定
2). 軸とデータの内積結果を基に閾値を決定し，データを分割する．
3). 軸の補空間へデータを射影する

の 3 つの手順から構成されている．ただし手順 3 に関しては木の親子間で射影軸を直交させる場合のみ実行する．本報告では，１）射影軸を主成分分析と判別分析で決定する場合，２）閾値を等確率分割とエントロピー・ゲイン最大化で決定する場合，３）子ノードの射影軸計算の際に親の射影軸の補空間に射影したデータ集合を対象とする場合とそうでない場合, の組み合わせ合計 8 通りについて考える．

本報告で考える手法の組み合わせ 8 通りでは，射影軸にデータを射影した 1 次元の結果に対する閾値の決定を行うことにより，主成分木と同様に木の探索時にクエリと射影軸との内積計算の結果と閾値との比較のみの計算でたどる子ノードが決定できるので，計算コストを抑える事ができる. また, 多分木にすることで, 主成分木で問題であった，作成した木が高くなることによる探索時間の問題を回避できることが考えられる．

以上で述べた計 8 通りの組み合わせについて，複数のデータセットに対して適用した実験を行う．

## 4. 実験

| データセット名 | データ数 | 訓練数 | テスト数 | 次元数 | クラス数 |
|---|---|---|---|---|---|
| breast cancer | 569 | | | 30 | 2 |
| shuttle | 58,000 | 43,500 | 14,500 | 9 | 7 |
| mnist | 70,000 | 60,000 | 10,000 | 784 | 10 |

表 1 実験に使用したデータセット一覧

3 章で述べた射影軸の決定法，射影軸上でのデータ分割法，補空間への射影の有無からなる合計８通りの条件で構築した多分決定木の識別性能や特徴を明らかにすることを目的に，様々なデータベースを対象とした場合の手法の，組み合わせと識別性能との関係を実験により明らかにする．また，木の分割数は 2 から 10 までの場合を，それぞれを実験する．

データベースは識別問題において一般に用いられるものを UCI Machine Learning Repository [4]より引用した．用いたデータセットについて表にまとめる．

用いたデータベースの内，訓練データとテストデータに分かれていないものはデータベースから一つ抜き出したものをクエリとし，その他のデータで木を構築した．識別性能の評価では，クラス識別の間違い数を集計しこれをデータ数だけ繰り返した．また，提案手法では求める基底の数で弁別性が左右されることが考えられる．よって各組み合わせで最大の深さを制限し木を構築した場合の結果も合わせて示す．木の高さを制限する場合，決定木の葉ノードでデータを複数保持することが考えられるが，木の探索時には，部分空間へ落とされたクエリと葉ノードが保持するデータとの最近傍探索を行い，最近傍データに付与されたクラスを識別結果とする．これは，通常決定木では多数決によるクラス判別が用いられる場合が多いが，極端にデータベースのクラス間で個数に差がある場合に常に判別されないクラスが発生することが考えられるので，これを回避するためである．

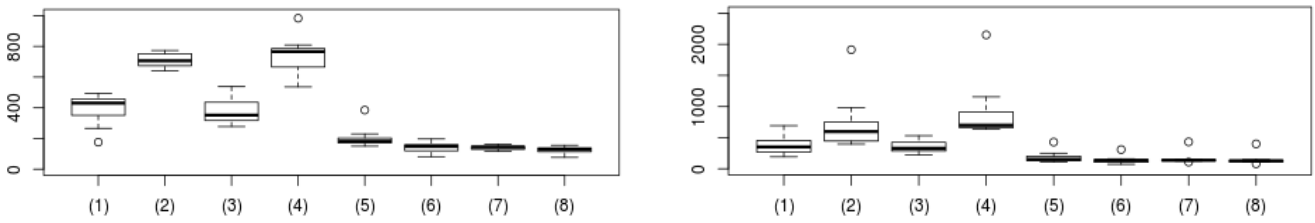
以上の条件により作成した多分決定木を各データベースに適用した場合の識別性能を評価した結果を次章で示す.

## 5. 実験結果

まず，Shuttle データベースを対象とし，多分決定木の各手法を組み合わせた場合のクラス識別間違い数を集計した．識別間違い数を縦軸に，各手法の組み合わせを横軸とした Box Plot にまとめた結果を図 3 に示す．木の高さの制限は制限のない場合で最も低くなった高さ 6 とした．

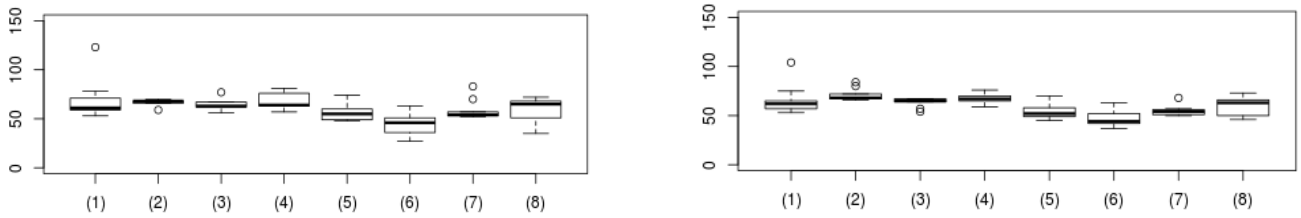
結果から，決定木の高さの制限の有無に関係なく，誤識別数に似た傾向が見て取れた．射影軸の決定においては主成分分析と判別分析では，判別分析を用いる場合の手法の組み合わせで識別間違いが少ないことが見て取れる．主成分分析を射影軸の決定に用いる場合に注目すると，補空間への射影を行う方が，識別間違い数が少ない結果が見られたが，射影軸上でのデータの分割方法の間での識別率に大きな差は見られなかった．また，判別分析を射影軸の決定に用いる場合に注目すると，データ分割法の選択と補空間への射影の有無をそれぞれ変更しても識別率に大きな差は見られなかった．しかし判別分析を用いる場合のみに注目すると，射影軸の補空間への射影を行う場合に，若干誤識別数が下がっていることが見て取れる．

このことから，決定木の作成において，データ分割時の計算方法よりも射影軸の決定方法の方が識別率に影響を及ぼすことが考えられる. しかし Shuttle を対象とした時に見




(a) (b)

図 3 shuttle データベースを対象とした場合の提案手法の誤識別結果．縦軸は誤識別数，横軸はそれぞれ 8 通りの手法の組み合わせとなっている，それぞれ（1）主成分分析かつ等確率分割，補空間への射影を行う場合，（2）主成分分析かつ等確率分割，補空間への射影を行わない場合，（3）主成分分析かつエントロピー・ゲイン最大化分割，補空間への射影を行う場合，（4）主成分分析かつエントロピー・ゲイン最大化分割，補空間への射影を行わない場合，（5）判別分析かつ等確率分割，補空間への射影を行う場合，（6）判別分析かつ等確率分割，補空間への射影を行わない場合，（7）判別分析かつエントロピー・ゲイン最大化分割，補空間への射影を行う場合，（8）判別分析かつエントロピー・ゲイン最大化分割，補空間への射影を行わない場合となっている．また，（a）は木の最大の高さを制限しない場合（b）は木の最大の高さを制限する場合




(c) (d)

図 4 breast cancer データベースを対象とした場合の提案手法の誤識別結果．縦軸は誤識別数，横軸はそれぞれ 8 通りの手法の組み合わせとなっている，それぞれ（1）主成分分析かつ等確率分割，補空間への射影を行う場合，（2）主成分分析かつ等確率分割，補空間への射影を行わない場合，（3）主成分分析かつエントロピー・ゲイン最大化分割，補空間への射影を行う場合，（4）主成分分析かつエントロピー・ゲイン最大化分割，補空間への射影を行わない場合，（5）判別分析かつ等確率分割，補空間への射影を行う場合，（6）判別分析かつ等確率分割，補空間への射影を行わない場合，（7）判別分析かつエントロピー・ゲイン最大化分割，補空間への射影を行う場合，（8）判別分析かつエントロピー・ゲイン最大化分割，補空間への射影を行わない場合となっている．また，（c）は木の最大の高さを制限しない場合（d）は木の最大の高さを制限する場合

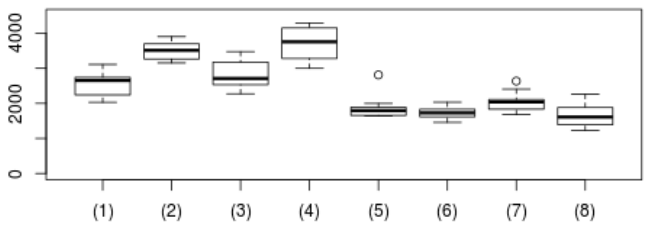
られた傾向が他のデータベースを対象とした場合でも見られるとは限らない．見られた傾向の一般性を確認するために提案手法を他のデータセットに適用した結果も示す．

Breast Cancer データベースを対象として Shuttle と同様に多分決定木を作成した場合のクラス識別間違い数を集計した結果を図 4 に示す． 木の高さの制限は制限のない場合で最も低くなった高さ 3 とした．

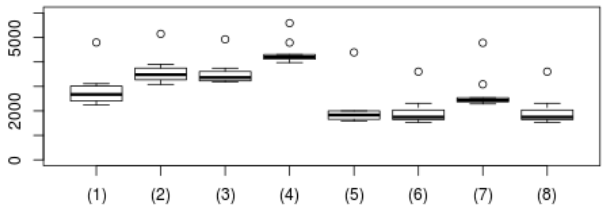
結果から，Breast Cancer を対象とした場合では，Shuttle を対象とした時と同様に高さの制限の有無で結果に大きな差は出なかった．射影軸の決定において主成分分析と判別分析を入れ替えても大きく識別率に大きな差は見られなかった．しかし，最も誤識別数が低い組み合わせは判別分析を用いた場合であった．主成分分析を射影軸の決定に用いる場合に注目すると，射影軸の補空間にデータを射影する場合の誤識別数が若干だが平均的に低いことが見て取れる．これは Shuttle と同様の傾向と言える．Breast Cancer は Shuttle と比べてデータ数が少ない，かつ次元数が高く，クラスが 2 種類のみであるなど，データの特徴が大きく異なっているために Shuttle ほど結果に顕著な差が出なかったと考えられる．特に次元数が高く，データ数が少ない特徴は，データの分割と補空間への射影を繰り返し，木を構築する提案手法では，必然的に木の高さが低くなり，十分な属性数でデータが表現できていない影響が考えられる．

最後に mnist データベースを対象とした場合の多分決定木の誤識別数の集計をする．識別間違い数を縦軸に，各手法の組み合わせを横軸とした Box Plot にまとめた結果を図 5 に示す．木の高さの制限は制限のない場合で最も低くなった高さ 5 とした．

結果からは他のデータセットを対象とした場合と同様の傾向が見て取れた．木の高さを制限する場合では射影軸




(e)




(f)

図 5 mnist データベースを対象とした場合の誤識別結果．縦軸は誤識別数，横軸はそれぞれ 8 通りの手法の組み合わせとなっている，それぞれ（1）主成分分析かつ等確率分割，補空間への射影を行う場合，（2）主成分分析かつ等確率分割，補空間への射影を行わない場合，（3）主成分分析かつエントロピー・ゲイン最大化分割，補空間への射影を行う場合，（4）主成分分析かつエントロピー・ゲイン最大化分割，補空間への射影を行わない場合，（5）判別分析かつ等確率分割，補空間への射影を行う場合，（6）判別分析かつ等確率分割，補空間への射影を行わない場合，（7）判別分析かつエントロピー・ゲイン最大化分割，補空間への射影を行う場合，（8）判別分析かつエントロピー・ゲイン最大化分割，補空間への射影を行わない場合となっている．また，（e）は木の最大の高さを制限しない場合（f）は木の最大の高さを制限する場合

の決定に主成分分析において，データの分割法に等確率分割とエントロピー・ゲインを最大化する分割を用いる場合とで識別間違い数に差が見て取れた．mnist データベースの次元数は 784 であり，breast cancer と同様に次元数と比べて作成される木の高さが十分とは言えないが，同様の傾向が見て取れた．このことから breast cancer を対象とした時に見られたやや異なった傾向は breast cancer 特有のデータの分布やクラス数に依存するものと考えられる．

実験結果として，3 種類のデータの特徴の異なるデータベースを対象とした場合に，多分決定木の各手法の組み合わせを適用した場合の誤識別数をまとめた．結果からは，複数のデータベースを対象とした場合でもよく似た傾向が見られた．データの分割時に用いる手法として等確率分割とエントロピー・ゲインを最大化する分割の 2 手法をそれぞれ適用した場合で誤識別数に大きな差は見られず，射影軸の補空間への射影の有無でも誤識別数に差はあっても全体としての傾向に大きな差がある結果はなかった．全体を俯瞰して，射影軸の決定に主成分分析または判別分析を用いる場合では，明らかに判別分析を軸の決定に用いる場合に誤識別数が低くなっており，決定木の構築においては，データ分割の手法の決定よりも射影軸を決定する手法の選択が重要であると結果からは見て取れた．

また，本報告では類似画像検索へ決定木を適用する場合についても検証を行う．Vocabulary Tree を多分決定木に置き換えた場合の識別の精度と速度についての実験結果を以下に示す．

Recognition Benchmark-Image[2]の画像セットを対象とした場合の類似画像検索のスコアについて調べた．スコアとしては N-S score を用いる．N-S score とは，Recognition Benchmark-Image の画像セットでは，同一の物体を別々の視点から撮影した 4 枚の画像が 2,550 組（計 10,200 枚）含まれているが，その 4 枚の画像の内 1 枚をクエリとして類似画像検索を行った時に，その同じ組の画像が選ばれる枚数のことを指している．よって N-S Score の最小は 0，最大は 4 となっている．

実験の条件として，作成する木の分割数は 10，高さを 6 に設定した．これは標準的な Vocabulary Tree の木の構造と近づけるためである．また，局所特徴を用いた類似画像検索では木の根ノード付近ではクラスラベルを用いた射影軸の決定と分割の決定では十分な性能が期待できないことが考えられる．そこで，クラス数が十分に少なくなるまでクラスタリングを行い，木の作成の途中から，射影軸への射影を用いた木の作成を行う手法を用いることを考える．この手法を本報告では MIX と呼称する．MIX では分割するデータの持つクラスラベルの種類が，分割数よりも少なければクラスタリングから射影計算を用いた分割の決定をするようにした．

多分決定木，Vocablary Tree，MIX の各手法で Recognition Benchmark-Image を対象とした時の N-S score について調べた．スコアをまとめたものを表 2 に示す．

| | N-S score |
|---|---|
| Vocabulary Tree | 2.99 |
| 多分決定木 | 2.39 |
| MIX | 2.98 |

表 2 Recognition Benchmark Images の画像セットを対象にした N-S score の結果．多分決定木は，最も N-S score が高い主成分分析かつエントロピー最大化分割，補空間への射影を行い木構築したものである．

また，木探索における時間コストを計測した．ノードをたどるためにかかる時間を GNU gprof を用いて時間を測定した．各ノードでかかった時間の平均を結果として表 3 に，

また，実験に用いた PC のスペックを表 4 に示す.

| | 探索時間(ns) |
|---|---|
| Vocabulary Tree | 0.10 |
| 多分決定木 | 0.03 |

表 3 Vocabulary Tree と多分決定木における探索時間
探索時間は gprof の Flat profile において，
self seconds を calls で割った値である.

| | |
|---|---|
| CPU | Intel Core i7-2700K 3.50GHz × 8 |
| Memory | 16GB |

表 4 実験に使用した PC の CPU およびメモリ.

結果から，Vocablary Tree，多分決定木，MIX ではスコアに大きな差が見られなかった．また，探索時間においては多分決定木が明らかに高速であることが確認できた.

## 6. まとめ

決定木の構築法として，提案手法の射影軸の決定法，射影軸上でのデータ分割法，射影軸の補空間へのデータの射影の有無からなる 8 通りの組み合わせを考え，複数のデータベースに対して適用した結果を示した.

結論から言えば，射影軸の計算法が性能を最も左右するファクターであり，判別分析が主成分分析よりも有効であった．また，データの分割法はさほど重要ではなく，エントロピー・ゲインの最大化と等確率分割の２つに明確な差は見受けられなかった．また，Gram-Schmidt の直交化を行うか否かは，主成分分析と判別分析で有効性に差がみられた.

このことから，射影軸や分割方法をランダムに決定し，エントロピー・ゲインで選択を行う決定木構築法の妥当性については検証が必要であると考えられる.

また，類似画像検索へ多分決定木を適用した結果からは，Vocablary Tree での各ノードでの距離計算に比べ，多分決定木での射影軸へ射影した 1 次元データでの閾値を用いた処理の方が明らかに高速であることが示された．類似画像検索のスコアは上回ることが出来なかったが，識別性能を向上させることでスコアも向上させられることが考えられる.

## 参考文献

[1] J. R. Quinlan, “Induction of decision trees,” Machine learning 1.1, pp. 81-106, 1986
[2] 大谷，大池，和田, “距離計算を行わない近似最近傍探索アルゴリズム: Grafting Trees,” 電子情報通信学会技術研究報告 PRMU，Vol.112, No.441，pp.137-142，2013
[3] A. Criminisi, J. Shotton, E. Konukoglu, “Decision Forests: A Unified Framework for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning,” Foundations and Trends® in Computer Graphics and Vision: Vol. 7: No 2-3, pp 81-227, 2012
[4] R.F. Sproull,“Refinements to nearest-neighbor searching in k-dimensional trees, ”Algorithmica, vol. 6, pp.579-589, 1991.
[5] J.-P. Heo, Z. Lin, and S.-E. Yoon, “Distance Encoded Product Quantization,” 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), DOI: 10.1109/CVPR.2014.274, 2014
[6] 荒井，中村，和田, “主成分木を用いた写像学習法の提案と性能比較“ 画像の認識・理解シンポジウム (MIRU), pp516—521 (2007)
[7] Yusuke Matsushita, Toshikazu Wada, “Principal component hashing: An accelerated approximate nearest neighbor search“, Proceedings of the 3rd Pacific Rim Sumposium on Advances in Image and Video Technology, pp374—385 (2009)
[8] D. Nister and H. Stewenius, “Scalable Recognition with a Vocabulary Tree“ 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), 2006, pp. 2161-2168.
[9] “UCI Machine Learning Repository“, (http://archive.ics.uci.edu/ml/)
[10] “The University of Kentucky Center for Visualization & Virtual Environments “, (http://vis.uky.edu/~stewe/ukbench/)