DEIM Forum 2011 C9-6

不均一データストリームのための オンライン決定木のノード構築アルゴリズムの提案と評価

峰岸　達也† 　新美　礼彦†† 　小西　修†††

† 公立はこだて未来大学大学院システム情報科学研究科　〒 041–8655 北海道函館市亀田中野町 116 番地 2
†† 公立はこだて未来大学システム情報科学部情報アーキテクチャ学科
〒 041–8655 北海道函館市亀田中野町 116 番地 2
††† 公立はこだて未来大学システム情報科学部複雑系知能学科　〒 041–8655 北海道函館市亀田中野町 116 番地 2
E-mail: †{g2109043,niimi,okonishi}@fun.ac.jp

あらまし　不均一データストリームからの決定木構築において従来の学習アルゴリズムを用いることは充分な性能を発揮することができない．不均一データストリームでは情報利得や Gini Index といった決定木分割基準値は，重要なクラスよりもより大きいクラスに有利に働き，分類精度を最大にする傾向にある．オンライン型決定木である VFDT は，データストリームに対応した決定木学習手法であるが，VFDT もそれらと同じ影響をうけることがわかっている．本稿では VFDT の構築アルゴリズムにおいて VFDT のノード構築アルゴリズムにおける新しい情報量基準値として情報量にクラスごとに重み付けを行った．データのクラスごとの重み付けを行い，データ分布を正規化させることで不均一データストリームが VFDT に及ぼす影響を解決する．我々は提案手法の有効性を検証するために約 2,500 万件の実データであるクレジットカード取引データを用いた実験と評価を行い，提案手法の有効性を示すことができた．
キーワード　データストリームマイニング，不均一データストリーム，オンライン型決定木，VFDT，情報量基準値

Proposal and Evaluation of Nodes Construction Algorithm of Online Decision Tree for Imbalanced Data Stream

Tatsuya MINEGISHI†, Ayahiko NIIMI††, and Osamu KONISHI†††

† Graduate School of Systems Information Science, Future University Hakodate
116-2 kamedanakano, Hakodate, Hokkaido, 041–8655 Japan
†† Faculty of Systems Information Science, Future University Hakodate
116-2 kamedanakano, Hakodate, Hokkaido, 041–8655 Japan
††† Faculty of Systems Information Science, Future University Hakodate
116-2 kamedanakano, Hakodate, Hokkaido 041–8655 Japan
E-mail: †{g2109043,niimi,okonishi}@fun.ac.jp

1. 序　　論

近年のネットワーク社会では，情報処理技術の発達により大規模なデータを収集・蓄積することが容易になり，そのようなデータを有効活用できないかという要望が多く挙げられる．このような流れを受けて，データから有益な情報を発見・活用する技術であるストリームマイニングが注目され，様々な分野において利用されている．ストリームマイニングには様々な分析手法があるが，データストリームを分類し，分類されたクラスから有益な情報を取り出すという分類学習が注目されている．しかし，近年のデータの変化からあるひとつのクラスが他のクラスに比べて極めて数が少ない，かつそのクラスの持つ情報が重要であるようなデータストリームが存在していて，それらは通信から金融，医学，web のカテゴリ化，生物学にまで及んでいる [1].

そのようなデータストリームに対して分類学習を適用する際に従来のアルゴリズムではいくつかの問題が発生することが知られている．従来の学習アルゴリズムでは最大限の分類精度を目指すようデータのどのクラスも等しく重要と扱い分類するので，このような条件下では常に最適な分類学習を行うことがで

きない [2] [3]．さらに，このような状況では，分類精度は分類学習における分類器の性能を適切に示す評価基準ではなくなってしまう [4]．分類学習の代表的なものである決定木，特に C4.5 や CART は決定木の導出のための有名なアルゴリズムであるが，そこで用いられている情報利得や Gini Index のノード分割基準値はクラスの分布に大きな偏りがあるようなデータに対して影響を受けやすいとされている [1]．

本稿では，そのようなあるひとつのクラスの数が極めて少ないデータストリームを”不均一データストリーム”と定義する．そして不均一データストリームに対応したオンライン型決定木である VFDT [5] を構築することを提案する．VFDT とはデータストリームに対応した決定木構築アルゴリズムであり，学習が高速であり，導出される分類器の表現が人間にとって理解しやすい．したがって，膨大な量のデータストリームを高速に処理し時間を短縮できるという点と，可読性の高さから分類ルールが理解しやすいという点で優れている．また，不均一データストリームとしては実際に取引されたクレジットカード取引データを使用する．

しかし，VFDT で用いられている Hoeffding bounds [3] [5] [6] と呼ばれる分割基準値も不均一データストリームに影響されてしまい，特にデータ到着の順序に依存されることがわかっている．本稿では，VFDT の新しいノード構築アルゴリズムを提案することでこれを解決する．

2. 関 連 研 究

2.1 VFDT

C4.5 や CART のように初めにすべての事例を入力として受け取り，決定木を構築するものをオフライン型決定木と呼ぶ．しかし，これはすべての事例がそろわないと決定木を構築することができないということと，事例にランダムアクセスをしなければならないということからデータストリームに適用することができない．これに対して，データストリームの短い間隔で次々と新しい事例が到着し，かつ累積する事例数が大量になるという特徴に対応した決定木をオンライン型決定木といい，代表的なものに VFDT(Very Fast Decision Tree learner) が挙げられる．VFDT はすべての事例の到着を待たずに決定木を徐々に成長させていくことができるので，メインメモリに事例を蓄積しない．VFDT のアルゴリズムはメモリ消費量と処理時間を減らすために，事例そのものを決定木中に蓄積するのではなく，事例のクラスと属性値の同時出現頻度のみを各ノードで蓄積する．VFDT では事例を受け取るごとにルートノードのみの決定木から枝を成長させ葉ノードを作成していくことで決定木を順に成長させていく．新規にノードを作成する際には，それまでのノードに頻度情報が蓄積し，そこから C4.5 で用いられているノード構築アルゴリズムと同様にデータの属性の情報利得を算出する．そしてその情報利得と VFDT で用いられる分割基準値を比較し，条件を満たすかどうかの判定を行い，ノードを分割し決定木を成長させる．VFDT ではノードを分割する際の分割基準値として Hoeffding bounds [3], [5], [6] と呼ばれる情報量基準値が用いられる．

2.2 Hoeffding bounds

Hoeffding bounds とは VFDT でノードを分割する際に用いられる情報量基準値である．値域が R の数値変数 r を n 回独立に観測し，その平均が $\overline{r}$ のとき，Hoeffding bounds は $1-\delta$ の確率で変数 r の真の平均が $\overline{r}-\epsilon$ より大きくなることを保証する．ここで，ϵ は以下のように定義される．

$$\epsilon = \sqrt{\frac{R^2 \ln(1/\delta)}{2n}} \tag{1}$$

ノードを分割する際に，ある葉ノードにおける最良の情報利得と，次の情報利得との差が ϵ より大きくなればその葉ノードからさらに分岐を作成する．Hoeffding bouds を用いると，$\overline{\Delta G()} = \overline{G(X_a)} - \overline{G(X_b)} > \epsilon$ のとき，属性 X_a でノードを分割することが $1-\delta$ の確率で正しいことがわかる．ここで $G()$ は情報利得関数，X_a は情報利得を最も大きくする属性，X_b は情報利得を 2 番目に大きくする属性である．

3. 提 案 手 法

3.1 不均一データストリームの問題点

分類学習で不均一データストリームを取り扱う際にいくつかの問題点があげられる．C4.5 や CART といった決定木では，データのクラスの不均一さに大きく影響されてしまうことが知られている．それらで用いられている情報利得や Gini Index はクラスの歪みに敏感な分割基準値であり，分布の歪度が増加するほど効率よく機能することが難しくなってくる．特に Gini Index では顕著に現れる．これらはデータのサンプリングによりデータ分布を調節することで性能が大きく改善されることがわかっている [1]．

データの不均一さは VFDT にも影響を及ぼす．2.2 で挙げた Hoeffding bounds が対象としているデータストリームは不均一データストリームのようにクラスの分布が偏ったものではなく，データ分布はガウス分布に従うと仮定している [6]．したがって，不均一データストリームに Hoeffding bounds を用いてしまうと，図 1 のクラス 1 のデータ分布のように他のクラスを持つデータ数に比べ極端に数が少ないために，他のクラスのデータ分布に埋もれてしまうことがある．その結果，クラス 1

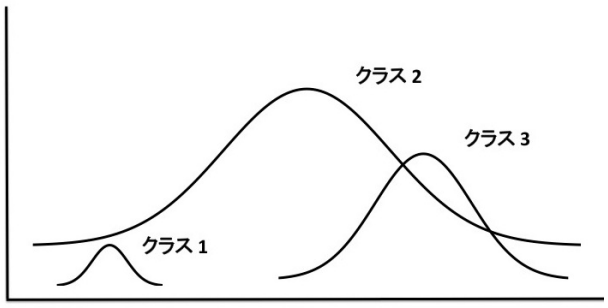

図 1　データ分布の違い

の分類を無視し，クラス 2 とクラス 3 のみを分類し最大の精度の分類を行おうとしてしまう．それにより VFDT の分類精度のみ高くなってしまうという問題があげられる．また，VFDT

ではデータの到着する順序から影響を受けやすい [3]．もしデータストリームで初期に観測されるデータ分布がデータストリーム全体のデータ分布と大きく異なっていたら，最終的なデータ分布を無視した VFDT を構築してしまう．

このような状況で VFDT を構築することは重要なクラスを無視することになり，結果として VFDT の全体の精度のみ高くなってしまう．それにより重要だが少数データで構成されたクラスの分類精度を無視した VFDT を構築してしまうことになる．

3.2 不均一データストリームのための VFDT のノード構築アルゴリズム

本稿では，3.1 であげた問題点にも対応できるよう不均一データストリームのための VFDT のノード構築アルゴリズムを提案する．VFDT を構築する際に入力される学習データのクラスごとに重み付けをすることで図 2 のようにデータ分布を正規化し，データ数の少ないクラスを学習に反映させることにより，それまでよりも効率的な分類ができるようにする．

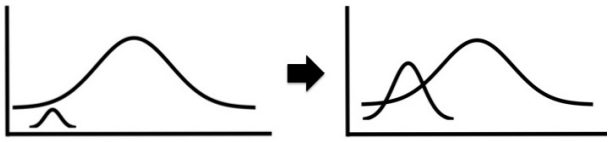

図 2　データ分布の正規化

本稿で提案するアルゴリズムとしては，VFDT を成長させていく上でのプロセスであるノード構築アルゴリズムにおいて，2.2 であげた情報利得の差と Hoeffding bounds の比較

$$\overline{\Delta G()} = \overline{G(X_a)} - \overline{G(X_b)} > \epsilon \qquad (2)$$

に用いられる情報利得 $G(X_a)$ の情報量と 情報利得 $G(X_b)$ の情報量を求める部分にクラスごとに重み付けを行う．本稿ではクラスは 2 つとし，数の少ないクラスに w として重み付けを行う．重みは $0 \leqq w \leqq 1$ の範囲とする．したがって数の少ないクラスをクラス C_1，もう一方のクラスを C_2 とすると，事例の集合 S に対して $freq(C_i, S)$ を S の中でクラス C_i に属する事例の数，集合 S に含まれる事例数を $|S|$ とすると，S からランダムに 1 つの事例を選び出し，それがクラス C_i に属しているとすると平均情報量 $\overline{info(S)}$ は

$$\begin{aligned}\overline{info(S)} = &-w \times \frac{freq(C_1,S)}{|S|} \times \log_2\Big(\frac{freq(C_1,S)}{|S|}\Big) \\ &-(1-w) \times \frac{freq(C_2,S)}{|S|} \times \log_2\Big(\frac{freq(C_2,S)}{|S|}\Big)\end{aligned} \qquad (3)$$

となり，式 (2) は，

$$\overline{\Delta G'()} = \overline{G'(X_a)} - \overline{G'(X_b)} > \epsilon \qquad (4)$$

となる．

また，VFDT では離散データストリームを対象としたアルゴリズムになっており，数値データストリームへの対応はしていない．しかし，本稿で VFDT を構築する際に用いたツールである VFML(Very Fast Machine Learning) [7] で公開されている VFDT のプログラムでは数値属性を取り扱うための改良が加えられている．具体的には Entropy-Based Discretization [6] という離散化法が導入されているが，このとき各数値属性の情報利得を最大とする属性値で 2 つの区間に離散化しているが，この情報利得の算出でも，上で挙げた同様の重み付けを行っている．

しかし，重み付けを行ったあとの情報利得と，Hoeffding bounds の比較である $\overline{\Delta G'()} > \epsilon$ (式 (4)) では，重み付けされた値と重み付けされていない値を比較してしまっている．そこで，右辺の Hoeffding bounds に 2 つのクラスに対しての重み w と $(1-w)$ の平均値である 0.5 をかけて，$\overline{\Delta G'()} > 0.5 \times \epsilon$ とすることで両辺の釣り合いを取ることとする．

4. 評 価 実 験

4.1 クレジットカード取引データ

本稿では実データとして実際に取引されたクレジットカード取引データをデータストリームに見立てて，不正利用の検出を目的とした評価実験を行う．クレジットカード取引データは，24 時間 365 日発生するデータであることからまさにストリームデータであると言える．さらに不正利用率は 0.02〜0.05%ほどと極めて低い割合であり，膨大な量のデータから極めて少ない不正利用を検出しなければならないということが課題とされていることから不均一データストリームであると言える．先行研究 [8] では，低い不正利用率のためにほぼすべての不正利用を正常利用と分類してしまう結果となっている．

今回の評価実験に用いたデータは，予備実験 [9] によりオフライン型決定木である C4.5 で性能が良かったデータを用いた．これは予備実験 [9] では不正利用率 0.02〜0.05%のデータで C4.5 を構築したが，低すぎる不正利用率のためにクラスが正常クラスであるルートノードのみの決定木を構築しただけであったためである．したがって，学習用データとしては約 50,000 件，不正利用率を 10%としてリサンプリングしたものを，またテスト用データとしては約 25,000,000 件，不正利用率が 0.048%の実際の 1ヶ月間のデータとした．

4.2 評価実験の設定

4.1 であげた学習用データを用いて 3.2 で提案した VFDT と，アルゴリズムに変更を加えていない VFDT をそれぞれ構築し，結果を比較した．さらにそれぞれで構築した VFDT に対して 4.1 であげたテスト用データを分類した結果を比較した．

結果の比較は，不正クラスに対する再現率，および VFDT の全体の精度，VFDT のサイズ，不正ルール数，VFDT の構築時間を比較する．リサンプリングしたデータの不正利用率は 10%であるので，3.2 の提案手法であるアルゴリズムの重み w はデータ分布の逆数として $w = 0.9$ とした．また，3.2 で示した情報利得と Hoeffding bounds の比較時の操作により，実際にはアルゴリズムに変更を加えていない VFDT の結果と，$w = 0.5$ は同じものである．

4.3 実験結果

VFDT の学習　4.1 であげた学習用データを用いて 3.2 で提案した VFDT と，アルゴリズムに変更を加えていない VFDT をそれぞれ構築した結果を比較した．ここに示す結果は 10-folds cross validation を行ったものである．表 1 は VFDT の全体の精度，VFDT のサイズ，不正ルール数，VFDT の構築時間を示している．精度を見ると従来手法よりも提案手法では分類精度が向上している．これは木のサイズが増加し，それに伴い不正ルール数が増えたことにより不正クラスに対する分類がより詳しく行われたためだと考えられる．その結果，分類精度も同様に向上したと考えられる．また，VFDT の構築時間はそれほど変化はしていない． しかし，先行研究 [8] から，分類精度

表 1　全体の精度，サイズ，不正ルール数，構築時間 (秒)

	精度 (%)	木のサイズ	不正ルール数	構築時間 (秒)
提案手法	92.325	106.600	5	6.722
従来手法	90.851	91.000	3	5.907

はほぼすべてが正常クラスを正常クラスに分類したものであるとわかっている．したがって，どれだけ不正クラスを正しく分類できたかの比較ができない．そこで表 2 と表 3 の Confusion Matrix から比較する．

表 2　Confusion Matrix(提案手法)

		実際のクラス	
		0(正常)	1(不正)
葉のクラス	0(正常)	40,174	1,982
	1(不正)	2,145	2,790

表 3　Confusion Matrix(従来手法)

		実際のクラス	
		0(正常)	1(不正)
葉のクラス	0(正常)	40,825	2,174
	1(不正)	1,494	2,598

Confusion Matrix を比較すると表 2 の提案手法では表 3 の従来手法よりも不正クラスを不正クラスと正しく分類できている数が 2,598 件から 2,790 件と 192 件増加している．不正クラスに対する再現率 (本稿では，不正クラスを持つデータ中の実際に VFDT が不正と分類したデータが占める割合と定義) でも 4%上昇している．ここで不正クラスに対する再現率とした理由は，不均一データストリームでは従来の再現率，適合率のような評価基準では判断することが難しいためである．今後，本研究では新しい評価基準を検討することも必要であると考えている．

また，提案手法と従来手法の不正ルールを比較するためにそれぞれの場合の不正ルールに分類された不正クラス数と正常クラス数を表 4 と表 5 に示す．

表 4　不正ルールに分類されたクラス数 (提案手法)

	不正クラス数	正常クラス数
不正ルール 1	292	236
不正ルール 2	98	72
不正ルール 3	1,343	961
不正ルール 4	70	58
不正ルール 5	987	818
合計	2,790	2,145

表 5　不正ルールに分類されたクラス数 (従来手法)

	不正クラス数	正常クラス数
不正ルール 1	38	15
不正ルール 2	1,628	1,247
不正ルール 3	932	232
合計	2,598	1,494

表 4 と表 5 からそれぞれの重みの値の場合における不正ルールの全通りの組み合わせ (提案手法：31 通り，従来手法：7 通り) において，どれだけのデータがそれぞれの不正ルールの組み合わせに分類され，そのなかに不正クラスがどれだけ分類されたのかを図 3 に示す．図 3 から，提案手法では不正ルールに

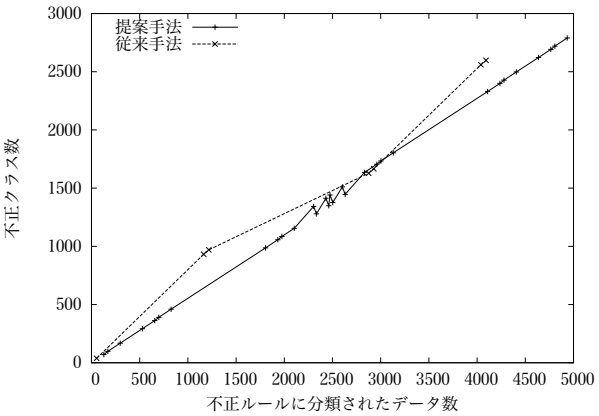

図 3　不正ルールの組み合わせにおける不正クラスの分類数

分類された不正クラスの合計数は従来手法よりも高くなっているものの，従来手法での不正クラスの合計数まででは不正クラスを分類した数が下回っている．しかし，提案手法では従来手法よりも不正ルール数が多いためより細かい不正ルールに分類されたデータ数で不正クラスの分類が可能である．

また，提案手法では不正クラスに対する重みを $w = 0.9$ と大きくしているため正常クラスを誤って不正クラスと分類してしまう数が増加している．しかし，4.1 であげたクレジットカード取引データの特徴から，少数データの分類精度 (不正クラスの再現率) を最も重要視していることから提案手法の重み付けが有効的であると考えられる．

VFDT のテスト　4.1 であげた実際の不正利用率であるテスト用データを前述した「VFDT の学習」で構築した VFDT のそれぞれで分類した結果を比較した．表 6 と表 7 はそれぞれのテスト後の Confusion Matrix を示している．このテストは，

不正利用率 10%の学習データで構築した提案手法，および従来手法のそれぞれの決定木を用いて，不正利用率 0.048%の実際の不正検出の分野における不正利用率のデータを分類したものである．表 6 と表 7 から，従来手法ではすべてのデータを正常

表 6　Confusion Matrix(提案手法)

		実際のクラス	
		0(正常)	1(不正)
葉のクラス	0(正常)	24,968,527	11,934
	1(不正)	145,457	346

表 7　Confusion Matrix(従来手法)

		実際のクラス	
		0(正常)	1(不正)
葉のクラス	0(正常)	25,113,984	12,280
	1(不正)	0	0

クラスに分類してしまっているが，提案手法では 346 件の不正クラスを持つデータを正しく不正クラスに分類できている．クレジットカード取引データでは膨大な量のデータから極めて少ない不正利用を検出しなければならないということが課題とされているため，VFDT による分類でも不正クラスを不正クラスと正しく分類することが難しい．しかしながら，実際の不正検出の分野における不正利用率のデータを従来手法ではすべて誤分類してしまったにもかかわらず，提案手法では少なからず分類することができたために不正クラスに対する再現率を向上させることができた．また，表 6 において VFDT が不正クラスに分類したデータが提案手法に現れた 5 つの不正クラスにどれだけ分類されたかを表 8 に示す．表 8 から不正クラスの 346 件

表 8　不正ルールに分類されたクラス数

	不正データ数	正常データ数
不正ルール 1	346	145,457
不正ルール 2	0	0
不正ルール 3	0	0
不正ルール 4	0	0
不正ルール 5	0	0
合計	346	145,457

と，VFDT が不正クラスと誤分類した正常クラス 145,457 件はすべて不正ルール 1 に分類されていることから不正ルール 1 は不正クラスを良く分類することができるルールであると考えられる．

以上から，VFDT のノード構築基準値として，ノード構築時の情報利得と Hoeffding bounds の比較の際の情報利得にクラスごとに重み付けをして学習させることにより，通常ではデータ数が少ないクラスが分類時にうまく分類されないような場合でも，精度を向上させることができることがわかった．今後は，予測モデルの性能を評価するために用いられる CAP (Cumulative Accuracy Profiles) 曲線を用いた結果の比較や静的なデータに対する検証，また外れ値検出などとの比較も予定している．

5. まとめと今後の展開

本稿では，あるひとつのクラスの数が極めて少ないデータストリームを不均一データストリームと定義し，オンライン型決定木である VFDT を不均一データストリームにも対応できるよう拡張した．VFDT のノード構築アルゴリズムにおける新しい情報量基準値として情報量にクラスごとに重み付けを行った．評価実験ではクレジットカード取引データのを用いて重み付けの有効性を検証した結果，提案手法の重み付けにより不正クラスの分類精度をより多く分類することができた．提案方法はストリームマイニングに特化した方法ではなく，一般的なデータマイニングでも利用可能である．しかし，ストリームマイニングでは，流れてくるデータをその場で処理しなければいけないので，通常のデータマイニングよりデータの分布が重要な意味を持ってくる．

今後の展開としては，Hoeffding bounds の決め方を含めてパラメータの設定の検討，提案手法の一般性を検証するための評価実験として他のデータを用いた実験を予定している．

文　献

[1] David A. Cieslak and Nitesh V. Chawla: Learning Decision Trees for Unbalanced Data, Proceedings of the 2008 European Conference on Machine Learning and Knowledge Discovery in Databases, pp.241-256,2008

[2] Yan Li; Yuhong Zhang; Hu Xuegang; Li Peipei; , "A classification algorithm for noisy data streams," Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on , vol.5, no., pp.2239-2244, 10-12 Aug. 2010

[3] Bernhard Pfahringer, Georey Holmes, Richard Kirkby: Handling Numeric Attributes in Hoeding Trees, Proceedings of the 12th Pacific-Asia conference on Advances in knowledge discovery and data mining, pp.296-307,2008

[4] Chris Drummond, Robert C. Holte: Exploiting the Cost (In)sensitivity of Decision Tree Splitting Criteria, Proceedings of Seventeenth International Conference on Machine Learning, pp.239-246,2000

[5] P. Domingos. G. Hulten: Mining High-Speed Data Streams, Proceedings of the ACM Sixth International Conference on Knowledge Discovery and Data Mining, ACM Press, pp.71-80,2000

[6] 西村 聖，寺邊 正大，橋本 和夫: 数値データストリームからの決定木導出, FIT2009, 第 8 回情報科学技術フォーラム, 2009

[7] P. Domingos. G. Hulten: VFML - a toolkit for mining high-speed time-changing data streams, http://www.cs.washington.edu/dm/vfml/, 2003

[8] Tatsuya Minegishi, Masayuki Ise, Ayahiko Niimi, Osamu Konishi: Extension of Decision Tree Algorithm for Stream Data Mining Using Real Data, IEEE, 5th International Workshop on COMPUTATIONAL INTELLIGENCE & APPLICATIONS 2009, 2009

[9] 峰岸 達也，伊勢 昌幸，新美 礼彦，小西 修: ロジスティック分析でのステップワイズ法と決定木による属性選択法の実データをもちいた比較, 日本知能情報ファジィ学会, 第 25 回ファジィシステムシンポジウム, 1A2-02 (6 pages in CD-ROM),2009