

決定木の差分分析によるコンセプトドリフトの抽出と分析手法

松澤 裕史[†] 久保 晴信^{††} 鹿島 久嗣[†]
井手 剛[†] 比戸 将平[†]

ビジネスの世界においては、販売促進や新規顧客の獲得のためにデータマイニング技術を用いて分析を行い、顧客行動のモデル化を行っている。しかしながら、前年と同じ特徴を持つ顧客層であっても、社会の環境変化などにより、その行動パターンが前年と異なってしまうため、再度、顧客分析のためモデル化を行っており、新しいモデルを用いて顧客の分析を行っている。このようなモデルの変化はコンセプトドリフトと呼ばれているが、コンセプトドリフトに対する従来研究の対象となっているのは、モデルの変化の有無や変化の大きさであるが、本稿では、モデルの変化内容を分析する方法について提案する。また、実在のデータを用いて実験を行い、その有効性について検証を行った。

Concept Drift Extraction and Analysis by Using a Difference Model of Decision Tree

HIROFUMI MATSUZAWA,[†] HARUNOBU KUBO,^{††}
HISASHI KASHIMA,[†] TSUYOSHI IDE[†] and SHOHEI HIDO[†]

This paper proposes a new approach to concept drift analysis, which is a recent important research topic in stream mining. While most of the conventional methods focus only on either how to detect concept drift or how to measure the degree of change, our problem is how to analyze the content of the change. Specifically, our method enables us to visualize the difference between two decision trees built at two different time points. Using a real-world data set, we demonstrate the utility of our approach.

1. ま え が き

1.1 コンセプトドリフト

ビジネスの世界において、売れ筋商品の分析や販売促進戦略の立案などを目的として、蓄積された顧客の購買行動データを用いて分析が行われている。データマイニング手法などを用いたデータ分析によって、顧客の購買行動モデルが構築され、構築された購買行動モデルから顧客がどのような商品に興味があるのか、あるいは、ある商品に興味があるのがどのような顧客層かという知識が抽出され、購買行動の分析や予測などに用いられている。このような顧客の購買行動モデルは、必ずしも恒常的なものではない。環境（景気動向、流行、法規制など）の変化により、購買行動モデルには変化が起きる。従って、ある一定期間毎に、新しいデータセットを用いて分析を行い、購買行動モデ

ルを構築し直す必要がある。各期間毎に顧客の購買行動を分析することも重要であるが、同時に、顧客の購買行動モデルに変化が発生しているかどうか、どのくらい大きな変化なのか、そして、どのような変化なのか、について分析することも将来の顧客の購買動向を分析する上で非常に重要なことである。データマイニングの研究分野において、モデルによって学習される知識はコンセプトと呼ばれ、コンセプトの時間的な変化はコンセプトドリフトと呼ばれている^{1)~3)}。

顧客の購買パターンを分析して、購買行動モデルを作成する方法の一つとして、決定木^{4)~6)}があり、これを購買行動モデルとして捉えることができる。決定木とは、複数の説明属性と1つの目的属性からなるデータセットに対して、目的属性を説明する知識を木構造により表現したものである。決定木では、各ノードにはテストが付与されており、各ノードにおいてテストを満たしているかどうかでデータが下位ノードに分割される。これをルートノードから終端ノードまで繰り返すことで、全ての入力データはいずれかの終端ノードに分類される。そして、終端ノードには、分類

[†] 日本アイ・ピー・エム (株) 東京基礎研究所
Tokyo Research Laboratory, IBM Japan, Ltd.

^{††} 日本アイ・ピー・エム (株) 大和システム開発研究所
Systems Development Laboratory, IBM Japan, Ltd.

ラベルが付与されており、辿り着いた終端ノードのラベルにより、そのデータがどの分類に属するものかを予測するのに用いることができる学習器である。

1.2 関連研究

これまで、コンセプトドリフトに関する研究として、二つのデータセットから生成されるモデルを比較し、両者に差異があるかどうか、あるいは差異の大きさを測る指標を与える方法について提案がなされてきている^{1)~3)}。例えば、Fan¹⁾は、2つのコンセプトの違いを特徴付けるパラメータを与えており、この方法は、どの程度コンセプトが異なるのかを一つの実数値で与えている。

しかしながら、従来手法は、コンセプトドリフトの有無やその変化量という点にのみ着目しており、具体的にどのような変化が起こったのかを分析する手法については、ほとんど研究がなされていない。

本稿では、コンセプトドリフトの内容を知るために、2つの決定木の違いを表す「差分モデル」という概念を導入する。それに基づいて、発生したコンセプトドリフトの内容の詳細分析を行う方法を提案する。

我々は⁸⁾において、教師付き学習器を用いて、2つのデータセット間の変化を学習する一般的な方法について提案を行っており、本稿は、その方針に従い、差分を分析するためのモデルとして決定木を用いてコンセプトドリフトを抽出する手法の提案と詳細な実験結果を示したものである。

2. コンセプトドリフト抽出アルゴリズム

我々は、与えられたデータセットに対して、コンセプトドリフトの抽出を次のように行う。全データセットを時系列順にソートし、これを時間軸に沿って分割する。分割されたデータセットからそれぞれ決定木を用いてモデルを構築(コンセプトを抽出)し、時間軸上で隣り合うデータセット間のコンセプトを比較することで、コンセプトドリフトを抽出する。

隣り合う2つのデータセットからそれぞれ構築した決定木を比較することができれば、両者のモデルの違い、すなわちコンセプトドリフトについて説明することが可能である。しかしながら、一般に、2つの決定木を比較することは、例えば、それが同じ属性を持つデータセットから構築された決定木であっても、直接、人手によって差分を観察して、コンセプトドリフトが発生したか、どのような変化であったかを分析することは困難である。そこで、我々は、2つの決定木を比較するため、その違いを説明するための差分モデルの導入を行い、コンセプトドリフトの内容分析を試みた。

我々が採用したコンセプトドリフトの分析に対するアプローチは、以下に示すように非常に単純なものである。比較すべきデータとして、2つのデータセット D_A と D_B があるとする。 D_A と D_B を学習データとして、それぞれ決定木 DT_A と DT_B の構築を行う。ここで、2つの決定木 DT_A と DT_B を比較するために、決定木 DT_A にデータ D_B を与え、また、決定木 DT_B にデータ D_A を与える。決定木に各データを与えることにより、目的属性に対する予測値が得られる。元データが持つ目的属性と(他方の学習データから構築された)決定木において得られる予測値が一致した場合には、そのデータを X_{agr} に分類し、予測が不一致だった場合のデータを X_{dis} に分類する。注意すべき点は、学習に用いたデータを構築された決定木に与えているのでは無く、決定木に対して別のデータセットを与えることで、差異の有無に関する検証を行おうとしているのである。

2つのデータセット X_{agr} と X_{dis} は、元の学習データ D_A 及び D_B の中で、変化が発生したものと発生しなかったものに分類したものである。2つのデータセット X_{agr} と X_{dis} に対して、それぞれ、一致、不一致のラベル (agr 及び dis) を付与し、これを目的属性とする学習データとして、再び、決定木 DT_{diff} を構築する。決定木 DT_{diff} は、 DT_A と DT_B の差分を表す差分モデルであり、決定木上の各ノード上のテストを観察することにより、どのようなデータが一致、あるいは、不一致に分類されたかを知ることができる。特に、不一致については、同じ内容のデータであっても異なる予測値を生成していることになるので、不一致と分類されるためのルールを決定木から読み取ることにより、コンセプトドリフトの内容について分析することができる。

図1にコンセプトドリフト抽出アルゴリズムを示す。図1において、*disagreement score* ρ を以下のよう

$$\rho \equiv \frac{|X_{dis}|}{|X_{agr}| + |X_{dis}|} \quad (1)$$

ここで、 $|\cdot|$ はデータセットの数を表す。 ρ は、 DT_A と DT_B で発生した変化度合いを表しており、 ρ の大きさによりコンセプトドリフトの発生があったかどうかを判断することが出来る。

2.1 不一致データに対する詳細な分析

元のデータセット D_A , D_B について、目的変数の取りうる値が少ない時には、コンセプトドリフトの詳細

agr, dis は、agree(一致),disagree(不一致)を表す

アルゴリズム： コンセプトドリフトの抽出

入力データ

- 2 つのデータセット D_A および D_B
- 2 つの決定木構築アルゴリズム L_{base} 及び L_{diff}
- Decision threshold $\nu > 0$

1. データセット D_A と D_B を決定木構築アルゴリズム L_{base} に適用し、それぞれ、決定木 DT_A と DT_B を得る
2. 構築した決定木 DT_A に対してデータセット D_B を適用し、決定木 DT_B に対してデータセット D_A を適用して、一致データセット X_{agr} と不一致データセット X_{dis} を得る。
決定木とデータセットの組み合わせが 1. と互い違いであることに注意
3. 式 (1) で与えられる *disagreement score* ρ を計算する。
4. もし、 $\rho \leq \nu$ ならば、コンセプトドリフトの発生が無いと返し、終了する。
5. もし、 $\rho > \nu$ ならば、コンセプトドリフトが発生しているので、データセット X_{agr} の各データに一致ラベルを付与し、データセット X_{dis} の各データに不一致ラベルを付与し、2 つのデータセット $X_{agr} + X_{dis}$ をマージし、決定木構築アルゴリズム L_{diff} に適用し、差分モデル (決定木) DT_{diff} を構築する。
6. 差分モデル (決定木) DT_{diff} を観察し、 X_{agr} と X_{dis} の差異を調べる。

図 1 コンセプトドリフト抽出アルゴリズム

Fig. 1 Algorithm for concept drift analysis

しい性質を解析することが可能である。例えば、不一致を詳細に分析することが可能である。分類すべきラベルが $Y = \{p, q\} (|Y| = 2)$ と与えられていたとすると、この場合には不一致データ X_{dis} を 2 つに分割して、 $X_{p \rightarrow q}$ と $X_{q \rightarrow p}$ にすることが可能である。 $p \rightarrow q$ は DT_A では p に分類されたデータが DT_B では q に分類されたことを示している。

3. コンセプトドリフトの分析実験

本章では、我々のアプローチに従い、実データを用いた実験結果を示し、提案手法について検証を行う。

実験対象とするデータセットは、ある組織の研究活動の履歴を収集したものである。最優秀論文賞の受賞や招待講演などの学術的な活動実績は、ワークショップのプログラム委員や研究会の運営委員などの活動なども含めて、その当事者によって入力されている。データベースの管理担当者は、活動実績について、四半期毎に集計を行い報告を行う必要がある。例えば、ノーベル賞の受賞など非常に顕著な活動実績は当然報告する必要があるが、査読の無い国内ワークショップなどは顕著では無いとして報告が求められていない。そのため、データベース管理担当者は、活動内容が顕著なものであったかどうかを表す“重要”ラベルを各活動実績に付与しておき、報告時には“重要”というラベルが付与されたものだけをピックアップすれば良いように運用されている。データベース管理担当者は、ガイドラインを設けて、できるだけ安定した判断基準で重要性の判断を行っているが、学会活動の種類などは広範であり、管理者が全ての研究分野に精通しているわけではない。また、その組織の研究分野も年々変化しており、新しい学会での研究活動が増えたりするこ

とがあるため、全ての活動実績内容に関して、機械的に判断可能な客観的な判断基準を設けることは困難である。例えば、ガイドラインでは、重要な活動実績の定義として、採録難易度の高い著名な国際的な会議における論文の採録や議長としての活動、世界的に名誉ある賞の受賞や活動など、となっていた場合、容易に判断できる活動内容もあるが、日本国内でしか開催されない国際会議などについては、その活動実績を重要とするかどうかは必ずしも自明でないため、管理者の主観的な判断によって、“重要”というラベルが付与されている。

しかしながら、データベース管理者の交代や、報告を求められる基準やガイドラインの改訂が、時々発生していることから、重要かどうかの判断基準が時間と共に変化していることが想像されるデータベースである。このような“重要”と付与される判断基準が変化していくデータセットに対して、これをコンセプトドリフトとみなし、いつ、どのように“重要”ラベル付与の判断基準が変化したのか解析を行った。

このデータベースの内容は、1 つのテーブルから構成されており、各行には、1 つの活動実績が入力されている。各活動の当事者によって、簡易入力システムから活動実績が入力され、登録申請されるとデータベース管理担当者は、その活動実績の内容に対して、内容を確認し、重要かどうかラベルを付与してデータベースに登録を行っている。

実験に用いたデータベースには、過去 6 年間の活動実績が登録されている。ここで、表 1 にはデータの例を表している。各データは、16 の属性値を持つ。例えば、(*JrnalSer*, *PanelNum*, *Position*, *OrgGov*) はそれぞれ、雑誌の種類、代表組織、職位、団体種別を表してい

表 1 学会活動実績データベース

Table 1 Examples of entries in the academic activity database

JrnalSer	PanelNum	Position	OrgGov	...	importance label
Transactions	Committee	Editor	WW	...	important
Journal	others	Program Comittee Member	WW	...	important
Magazine	Standard	Member	Japan	...	unimportant
Issue	Committee	Editorial Board Member	others	...	important

る。そして“重要”というラベルを表すため、*important label* カラムには、(‘important’ か ‘unimportant’) の 2 値のどちらかが付与される。例えば、*JrnalSer* の種類は ‘Transactions’, ‘Journal’, ‘Magazine’, ‘Issue’ などからなる。

データの入力、管理者ではなく、各活動の当事者による入力であるため、同じ学会名であっても異なる表記が入力されていたり、同じ役職であるのに学会によりその英語表記が異なっているなど、入力データにはバラツキが発生しており、名寄せを行って、同じ内容ではできるだけ統一表記となるように、データクレンジングを施した。また、データの入力内容に不備があって決定木構築に利用できないレコードを除去して、残った 1,214 件のデータを用いて実験を行った。

これらのデータを時系列に並べ、年度毎にデータを分割して、6 つのデータセットを作成した。図 1 のアルゴリズムに従い、これらのデータセットに対して、コンセプトドリフトの分析を行った。1 回の分析で用いるデータは、連続する 2 年分のデータであり、6 年分のデータであるので、全部で 5 回の分析を行っている。

なお、 L_{base} と L_{diff} として用いた決定木構築ツールとして、 R^7 の *mvpart* パッケージを用いた。このパッケージでは、Gini 係数を用いてデータの分割を行っている。

アルゴリズム (図 1) ステップ 1. によって生成された決定木の一例として、2003 年度データから構築された決定木を図 2 に示す。図 2 について、簡単に説明する。この決定木は、147 件のデータから構築されており、各データには、“imp” または “uni” の何れかのラベルが目的属性に付与されている。ルートノードには *JournalSeries* が “Magazine”, “Transactions”, “Journal” かどうかというテストがあり、該当する 17 件のデータが左のリーフに割当てられている。左のリーフにおいて、17/0 とあるのは、“imp” というラベルが付与されたデータが 17 件、“uni” というラ

ベルが 0 件であることを示している。棒グラフは、その数字を表すものとなっている。また、その数字の上にある “imp” というのは、このリーフで大勢を占めるデータのラベルである。ルートノードで上記ルールに該当しなかったデータは、ルートノードから右のノードへ割当てられ、ノード上のルール (*PanelNm* が “Standard”, “Committee” かどうか) に該当するデータは左へ、そうでないものは右へ分割されている。最右のリーフには、18 件の “imp” ラベルを持つデータ、及び 97 件の “uni” データが割当てられている。このリーフは “uni” が多数を占めているので、リーフのラベルとして “uni” が付与されている。

図 3 は、5 回の計算においてステップ 3. における *Disagreement score* ρ を計算した結果である。2003–2004 と 2006–2007 に *Disagreement score* ρ のピークがあることがわかる。*Decision threshold* ν はユーザによって決められるため、 ν の取り方によっては、毎年発生する僅かな変化であるという解釈も可能であるが、ピーク時である 2003–2004 と 2006–2007 は、実際にデータベース管理者が交代した時期と一致しており、少なくともこの二つの *Disagreement score* ρ のピーク時にはコンセプトドリフトが発生していたと考えられる。そこで、2003–2004 のピークについて着目して差分モデルの解析を行う。

図 4 に 2004 年度データからそれぞれ構築された決定木を示す。図 2 と図 4 を比較することが、本研究の主題であるが、見てわかるように、人手による観察では、これらの比較が容易に行えないことがわかるであろう。

2003 年のデータを 2004 年のデータから構築した決定木に与え、また、2004 年のデータを 2003 年データから構築した決定木に与えて、それぞれ、一致、不一致の集計を行い、その結果を元にして差分モデル (決定木) を作成した。図 5 に作成した決定木を示す。2.1 節に示した方法で、我々は、 $X_{imp \rightarrow uni}$, $X_{uni \rightarrow imp}$ 及び X_{agr} の 3 つのクラスをもつ決定木を作成した。

図 5 の決定木の見方は図 2 と同様であるが、各リーフは、3 つのラベルからなるため、各リーフには $X/Y/Z$ という形式で各ラベルの分布が表示されている。 X, Y, Z

imp, uni は、それぞれ important, unimportant を省略したものである

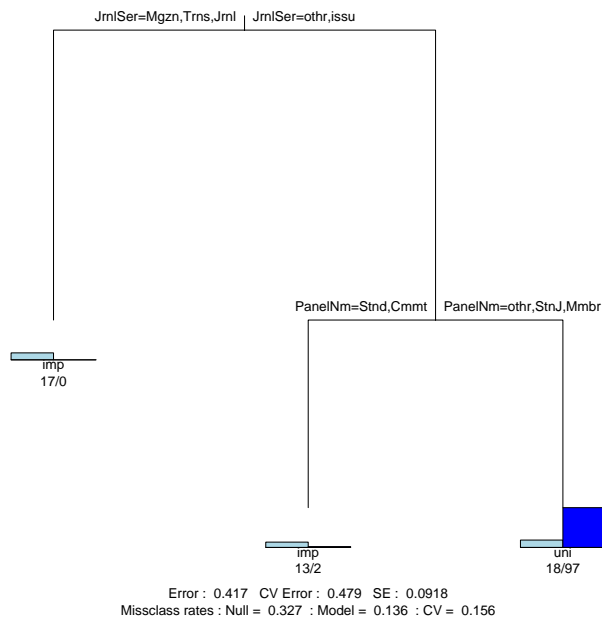


図 2 2003 年データから構築される決定木
Fig. 2 Decision Tree for 2003

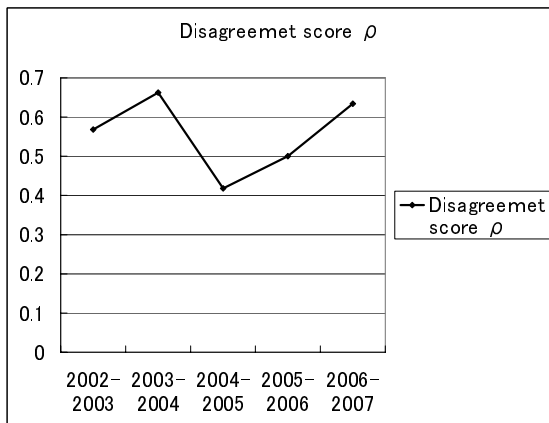


図 3 Disagreement score ρ
Fig. 3 不一致度 ρ

はそれぞれ, “agr”, “uni→imp”, “imp→uni” というラベルを持つデータの数を表している. 同様に, 棒グラフも X,Y,Z に該当する高さの棒 (bar) から構成されている.

図 5 は, 全部で 2003 年度 (147 件) 2004 年度 (133 件) の合計 280 件のデータから構築された決定木であり, 大半は agr に分類されていることがわかる. いま, コンセプトドリフトの観点から観察したいのは, どのようなデータが “uni → imp” 及び “imp → uni”

に変化したかであるので, そのようなリーフを見つけ, そのリーフに分類されるためのルールを調べることにより確認することが出来る.

まず, “uni → imp” というラベルを持つデータ, 即ち, 2003 年には “重要” というラベルは付与されていなかったが, 2004 年には “重要” ラベルが付与されるようになったデータに着目する. 決定木上からそのようなリーフを探すと, 最右のリーフが該当することがわかる. このリーフにたどり着くためのルールを調べることができる. 即ち, (X.University=others, TIoT) かつ (JrnIser=issue, others) かつ (PanelNm=Member, others, Standard Japan, Standard WW) かつ (SocAcad=others) かつ (Descrip=Award, others, Panelist, speaker) であるデータが, 2003 年から 2004 年にかけて “重要” ラベルを付与すべきデータになったことがわかる. なお, 利用した決定木作成ツールには, 決定木を木構造で表示するだけでなく, 各ノードのルールをテキストで出力する機能があるため, 上記のルール生成も人手により決定木を眺めるのではなく, 自動的に抽出することが可能である. 図 6 に, 出力されたルールを示す.

次に, “imp → uni” というラベルを持つデータ, 即ち, 2003 年には “重要” というラベルを付与してい

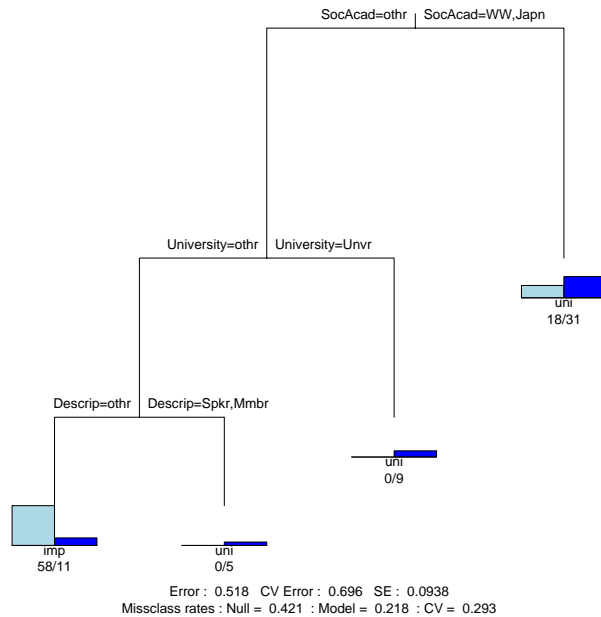


図 4 2004 年データから構築される決定木
 Fig. 4 Decision Tree for 2004

たが、2004 年には“重要”ラベルを付与しなくなったデータに着目する。左から 4 番目のリーフに含まれている 9 件のデータが存在していたが、このリーフは枝狩りの結果、これ以上、分類することができなかったということを表しており、この決定木からは、これ以上の分析を行うことができなかった。

以上の分析により、2003 年と 2004 年の間において、重要な活動実績の判断基準の変化について、特に、2003 年には“重要”というラベルは付与されていなかったが、2004 年には“重要”ラベルが付与されるようになったデータがどのようなものであったかを分析することができた。

4. おわりに

我々は、2 つのデータセットから構築される 2 つの決定木の間に発生するコンセプトドリフトに注目した。相互に元データを与えることにより、その一致・不一致を判断し、これを学習データに用いて再度、決定木を構築することで、2 つの決定木間の差分を差分モデルとして分析を行った。

コンセプトドリフトに対する従来研究は、主にモデルの変化点の検出やモデル変化の大きさに対するものであるのに対し、モデルの変化の内容を抽出し分析を行う一つの手法を提案した。実在のデータセットを用

いて、実験を行い、提案手法が有益であることが確認することができた。

図 1 において、我々は L_{base} と L_{diff} が異なっても良いという意味で、 L_{base} 及び L_{diff} と 2 つの学習器を区別した。本論文の実験では、どちらも同じ学習器を利用したが、例えば、 L_{base} がニューラルネットワークを用いた学習器で、そのコンセプトドリフトを抽出するために、 L_{diff} として決定木を用いるという利用方法もあると考えているからである。決定木は、不安定な学習器であるため、過学習 (overfitting) しやすいなどの問題はあつたものの、出力結果について説明能力が高いため、本稿の目的には有利であると思われる。

実験に用いたデータセットは小規模なものであり、今後、より大規模なデータセットを用いて手法の検証を行いたいと考えている。

参考文献

- 1) W. Fan, "Streamminer: A classifier ensemble-based engine to mine concept-drifting data streams.", In Proc. the 30th Intl. Conf. Very Large Data Base, 1257-1260, 2004.

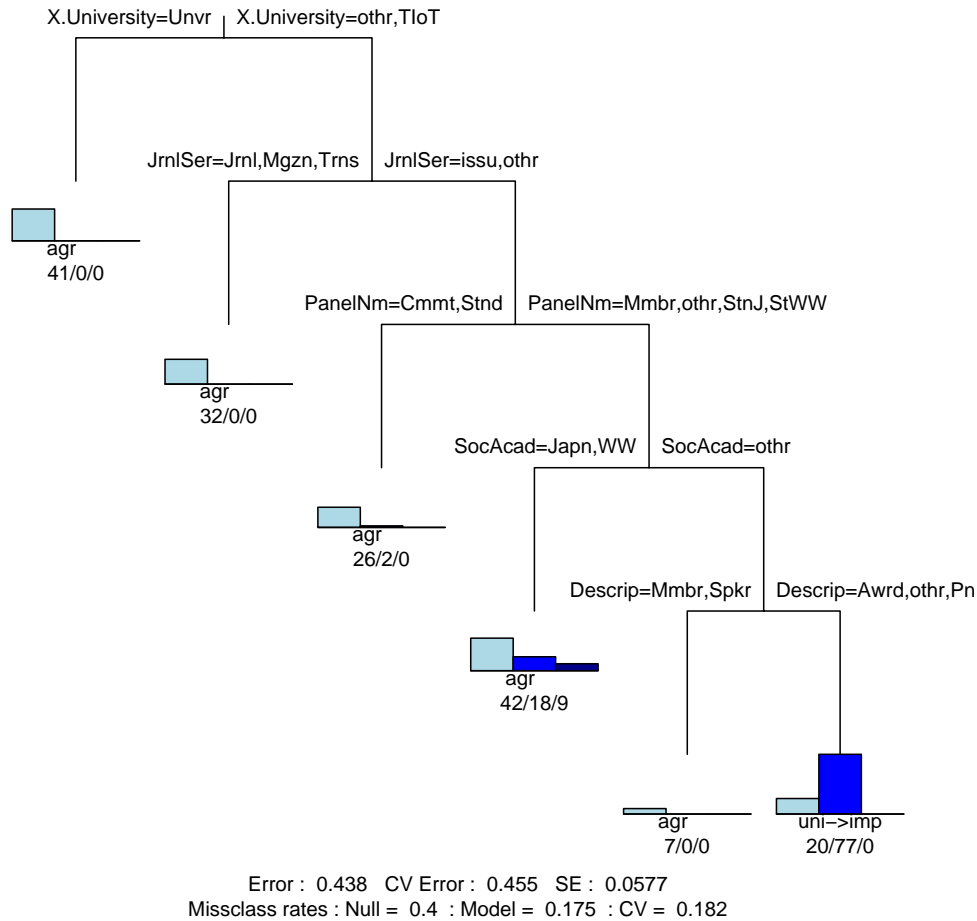


図 5 2003 年と 2004 年間の差分モデル
 Fig. 5 Difference Model between 2003 and 2004

- 2) H. Wang, J. Yin, J. Pei, P.S. Yu, J.X. Yu, "Suppressing model overfitting in mining concept-drifting data streams", In Proc. the 12th ACM SIGMOD Inter. Conf. Knowledge Discovery and Data Mining, 20-23, 2006.
- 3) Y. Yang, X. Wu, X. Zhu, "Combining proactive and reactive predictions for data streams", In proc. the 11th ACM SIGKDD Intl. Conf. Knowledge Discovery and Data Mining, 710-715, 2005.
- 4) L. Brieman, J. H. Friedman, R. A. Olshen, C. J. Stone, "Classification and Regression Trees", Wadsworth, 1984.
- 5) J. R. Quilnan, "Induction of decision trees.", Machine Learning, 1: 81-106, 1986.
- 6) J. R. Quinlan, "C4.5 : Programs for Machine Learning.", Morgan Kaufmann, 1993.
- 7) R Development Core Team, "R: A language and environment for statistical computing", R Foundation for Statistical Computing, 2005, <http://www.r-project.org/>
- 8) S. Hido, T. Ide, H. Kashima, H. Kubo, H. Matsuzawa, "Unsupervised Change Analysis using Supervised Learning", In Proc. the 12th PAKDD Pacific-Asia Conf. Knowledge Discovery and Data Mining, pp.148-159, 2008.

1) root 280 112 agr (0.60000000 0.36785714 0.03214286)
2) X.University=University 41 0 agr (1.00000000 0.00000000 0.00000000) *
3) X.University=others,TIoT 239 112 agr (0.53138075 0.43096234 0.03765690)
6) JrnlSer=Journal,Magazine,Transactions 32 0 agr (1.00000000 0.00000000 0.00000000) *
7) JrnlSer=issue,others 202 104 uni->imp (0.47029703 0.48514851 0.04455446)
14) PanelNm=Committee,Standard 28 2 agr (0.92857143 0.07142857 0.00000000) *
15) PanelNm=Member,others,Standard Japan,Standard WW 174 78 uni->imp (0.39655172 0.55172414 0.05172414)
30) SocAcad=Japan,WW 69 27 agr (0.60869565 0.26086957 0.13043478) *
31) SocAcad=others 105 27 uni->imp (0.25714286 0.74285714 0.00000000)
62) Descrip=Member,Speaker 7 0 agr (1.00000000 0.00000000 0.00000000) *
63) Descrip=Award,others,Panelist,speaker 97 20 uni->imp (0.20618557 0.79381443 0.00000000) *

図 6 生成された決定木のルール

Fig. 6 Rule of generated decision tree