２００６年度　卒業論文

データマイニングアルゴリズム
「アプリオリ」と「ＩＤ３」の比較

研究指導　菊池浩明　教授

東海大学　電子情報学部　情報メディア学科

3ADM1127　倉野　奈央子

3ADM3119　阿久津　忍

目次

付録２

第１章

はじめに

現在、情報技術の発達に伴い大量のデータの中から有益な情報、知識を抽出する技術、データマイニングが社会的に必要とされている。データマイニングでは、アプリオリ[1]と ID3[2] という２つの主要なアルゴリズムがしばしば用いられることが多い。しかし、与えられた目的属性についての論理決定木を作る ID3 に対して、アプリオリでは全ての属性間の組み合わせから相関ルールを抽出するという違いがある。そこで、実際のアンケートデータに適用した出力結果を比較し、出現属性や論理関係には相関があるのか確かめることにした。

本研究では、100 名に対してのアンケートを、代表的な手法である決定木学習アルゴリズム ID3 とアプリオリ分析との比較を行い、その出力結果から出現属性や論理関係から相関があることを確かめる。

第２章
データマイニングアルゴリズム

2.1　データマイニングとは
売店の販売データや電話の通話履歴、クレジットカードの利用履歴など、企業に大量に蓄積されるデータを解析し、その中に潜む項目間の相関関係やパターンなどを探し出す技術のことで、以下のような手法がある。

≪代表的なデータマイニングアルゴリズム≫

・決定木
分析の結果を木構造のモデルで出力する方法。
結果が木構造で出力するため理解しやすく高速に処理出来るが、誤差が発生する可能性がありモデル構築でコンピュータに負荷が掛かる。
代表的なアルゴリズムは ID3。

・クラスタ分類
複雑な集合を類似したカテゴリーに分けて、それを理解しようとする方法。
探索的手法で容易にデータ入力が出来るが、正しい距離尺度やウェイトを設定することが困難なケースが有り結果の解釈が難しい。
代表的なアルゴリズムは K-means 法。

・マーケットバスケット分析
要素の組み合わせから、それらの関連性を探る方法。
強力な探索的手法で明確に結果を理解できる。データ属性が限定的に扱われ問題の規模に対して指数的に計算量が増大する。
代表的なアルゴリズムはアプリオリアルゴリズム。

2.2 アプリオリ

アプリオリアルゴリズムとは Rakesh Agrawal らによって提案されたアルゴリズムである。本アルゴリズムは、ユーザが確信度のしきい値(最小確信度：minconf) とサポートのしきい値(最小サポート：minsupp) を与えて、最小確信度以上の確信度と最小サポート以上のサポートをもつ全ての相関ルールを効率よく出力する。データベース中の全トランザクションのうち、アイテム集合 X を含むトランザクションの割合を X のサポートといい、supp(X) と表記する。また、ユーザが与えた最小サポート以上のサポートをもつアイテム集合を頻出アイテム集合(frequent itemset) とよぶ。このとき、相関ルール“X⇒Y”のサポートと確信度は、それぞれ、

$$supp(X \Rightarrow Y) = supp(X \cap Y)$$

$$conf(X \Rightarrow Y) = supp(Y \cap Y) / supp(X)$$

と定義されている。したがって、条件を満たす相関ルールを求めるためには、$supp(X \cap Y) \geqq minsupp$, かつ、$supp(X \cap Y)/supp(X) \geqq minconf$ であるようなアイテム集合 Y 、X を生成し、そのサポートを求めればよい。

2.3 決定木

データから知識を獲得する方法として, 決定木学習アルゴリズム ID3, C4.5 がある. 決定木を構成するために用いるデータの事を学習データと呼ぶ。ここで学習データ D は、m 個の事例 d1,...,dm からなる。事例はそれぞれに属性値を持つ n 個の属性 a1,...,an とターゲット属性であるクラスからなる。k 個クラスがあったとするとクラスは C1,...,Ck と表され、同じクラスを持つ事例によって k 個の部分集合が作られる。
決定木のノードには属性、枝には属性値、葉にはクラスが対応する。

2.3.1 ID3

決定木学習アルゴリズム ID3 は、最も簡潔な木を作ることを目的としている。そのために ID3 は、各属性に対して情報量利得を計算し、その値が最も大きな属性から順にノードとして選択し、決定木を構成していく。次に、これらのアルゴリズムについて本研究で重要な点について述べる。

2.3.2　枝刈り

大規模な学習データから決定木を生成するとき、意味的な内容とは無関係な、非常に複雑な木となることがある。そこで、次のパラメータに基づいて木を簡単化する。

1. M（重み）

分類を進めるために必要な事例数の最小値を意味する。この値を大きくすると、枝刈りの効果が大きくなり、木が小さくなる。

2. CF（信頼度）

決定木の葉の誤り率をどこまで許すか設定できる。小さな値ほど大きな枝刈りが行われるが、誤差が大きくなる。

第 3 章
実験評価

3.1　使用ツール紹介
今回の実験では、先輩方の作ったツールを使わせていただいたので、ここで紹介する。

3.1.1　Gather
2005 年度卒業生、本多淳司「アプリオリアルゴリズムに基づいたデータマイニングツール"Gather"」

アプリオリアルゴリズムに基づいたデータマイニングツール。実験により”Gather”は人間の計算時間の１％も満たない時間で計算することが分かったと、述べられている。

3.1.2　ID3E
2004 年度卒業生、並木翼「ユーザビィリティィの高い GUI ベースの決定木学習ツール"ID3E"」

ユーザビリティィの高い GUI ベースの決定木学習ツールを開発。ユーザビリティ評価実験により、開発ツールの総合実験時間は従来ツールの 85%で済むことが分かったと、述べられている。

3.2　アンケートデータ
本研究で用いたアンケートは、東海大学生 100 人を対象に、2006 年 5 月下旬から 7 月上旬にかけて実施した。アンケートの質問を表 1 に示す。これらの２択の質問項目を属性として用いる。

表１：アンケートデータ

Q1	あなたの性別は？	男	女
		63	37
Q2	異性の友達が･･･	多い	少ない
		39	61
Q3	異性間の友情は･･･	成立する	成立しない
		89	11
Q4	海外に行ったことが･･･	ある	ない
		51	49
Q5	さびしがり屋ですか？	はい	いいえ
		67	33
Q6	犬と猫ならどっちが好き？	犬	猫
		68	32
Q7	自分が好き？	好き	嫌い
		48	52
Q8	今ダイエットを･･･	している	していない
		22	78
Q9	感情を表に出すタイプ？	はい	いいえ
		49	51
Q10	絶叫マシーンが･･･	好き	嫌い
		70	30
Q11	付き合った経験が･･･	ある	ない
		84	16
Q12	自分は遠距離恋愛が出来ると･･･	思う	思わない
		41	59
Q13	やきもち焼きですか？	はい	いいえ
		64	36
Q14	相手色に･･･	染まる	染まらない
		44	56
Q15	いつでも自分を1番に考えてほしい？	はい	いいえ
		42	58
Q16	人のケータイを黙ってみたことが･･･	ある	ない
		25	75
Q17	一人の時間は大切ですか？	はい	いいえ
		94	6
Q18	ペア商品を持っている（今いない人は持ちたい？）	持ってる	持ってない
		42	58
Q19	結婚と恋愛は別だと･･･	思う	思わない
		66	34
Q20	あなたは束縛をする人ですか？	はい	いいえ
		25	75

3.3　アプリオリシステムの性能

アプリオリシステムの性能を示すため、Java を用いて相関ルールを抽出するシステム”Gather” を用いて、3.2 節のアンケートデータをアプリオリシステムで分析し、最小サポートや最小確信度についての得られるルール数を調査した。この結果を図 1、図 2 に示す。

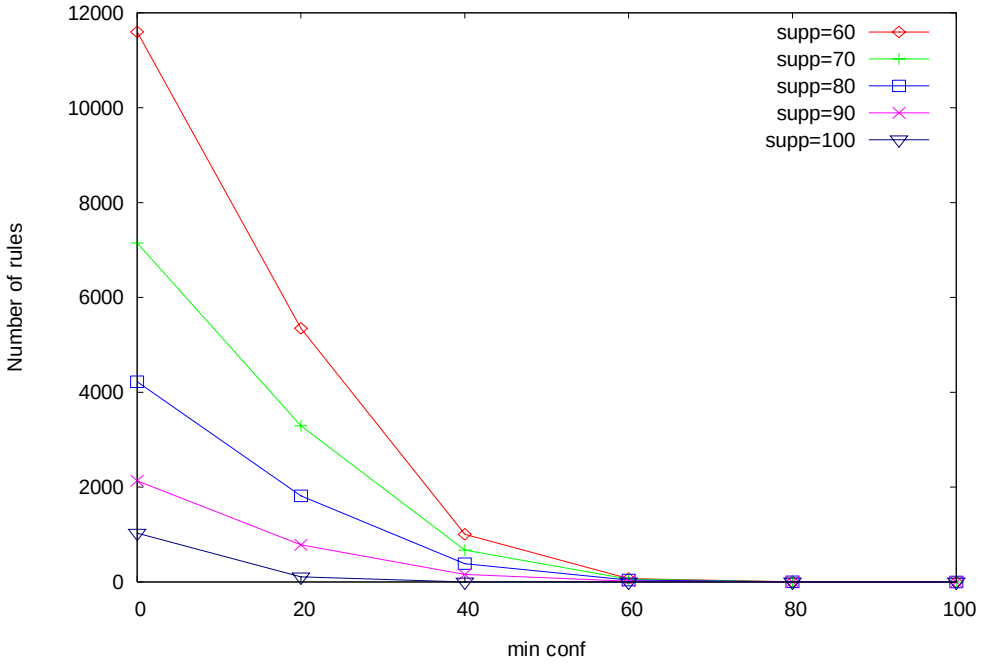


図 1: min supp とルール数の関係

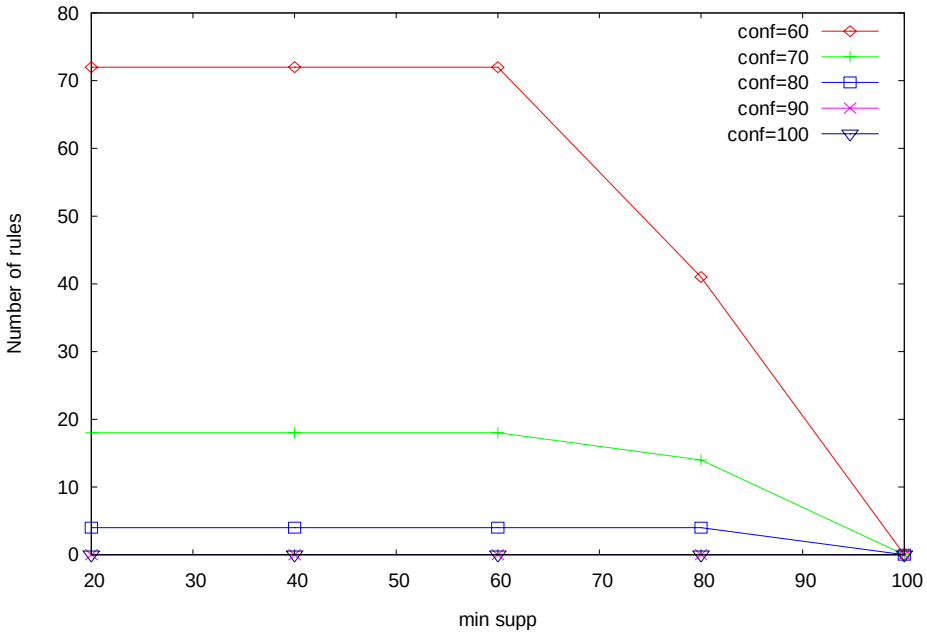


図 2: min conf とルール数

図 1 の結果より、最小確信度の値と得られるルール数には反比例の関係があることが明らかになった。また Conf の値によって、ルール数は大きく変わる。図 2 の結果より、最小サポートの値が一定に達すると、急激にルール数が少なくなることが示された。

3.4 出現属性の比較

ID3 とアプリオリの出力結果による出現属性の信頼性を確かめるため、アンケートデータを ID3 とアプリオリの 2 つのアルゴリズムでデータマイニングし、ID3 とアプリオリそれぞれ、あるターゲット属性について相関のあった属性に関して、2 種類の比較を行う。

1 つ目は、ID3 の出力した決定木で高い位置に出現した属性、アプリオリの出力したルールに何度も出現した属性を重要属性とし、その比較をする。ID3 については枝狩りを行い、出現ノードを減らすことでそれらを重要属性としている。

海外に行った経験があるかどうかというターゲット属性に関してのアプリオリの出力を表 1 に、ID3 の出力を図 3 に示す。また、これらの結果をまとめて表に表したものを表 2 に示す。

表 2：海外旅行経験の相関ルール

	出力されたルール
1	感情を表に出す⇒海外経験ある
2	自分は遠距離恋愛できない⇒海外経験ある
3	相手色に染まらない⇒海外経験ある
4	異性間の友情は成立する∧自分は遠距離恋愛できない⇒海外経験ある
5	ダイエット中でない∧相手色に染まらない⇒海外経験ある
6	一人の時間は大切∧相手色に染まらない⇒海外経験ある
7	束縛しない∧相手色に染まらない⇒海外経験ある
8	感情を表に出さない⇒海外経験ない
9	感情を表に出さない∧一人の時間は大切⇒海外経験ない

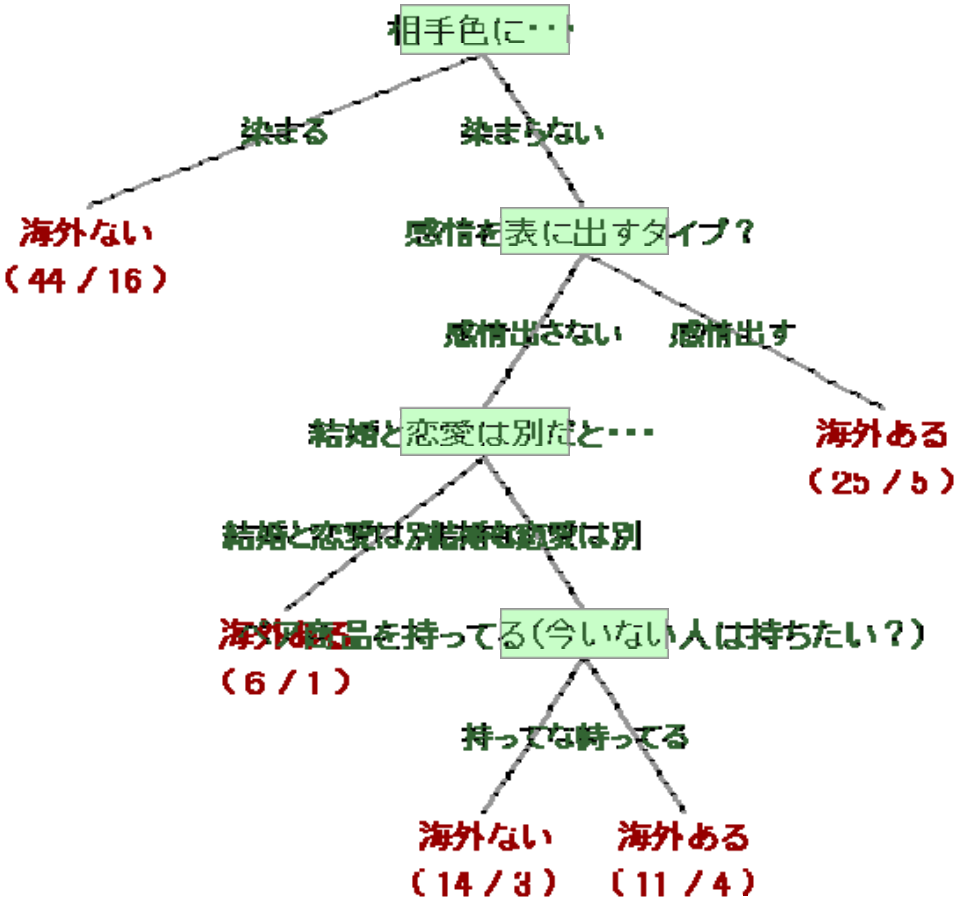


図３：海外旅行経験の決定木
（クラスに分類される事例の数/分類誤りの事例数）

表 3：海外旅行経験のアプリオリ、ID3 の出現属性

	属性	ターゲット	ID3 m=5,C=9 高さ 4,ノード 4	アプリオリ Supp=0.3,Conf=0.6 ルール 9 個 m=2	 m=3
Q1	あなたの性別は？				
Q2	異性の友達が・・・				
Q3	異性間の友情は・・・				1
Q4	海外に行ったことが・・・	○			
Q5	さびしがり屋ですか？				
Q6	犬と猫ならどっちが好き？				
Q7	自分が好き？				
Q8	今ダイエットを・・・				1
Q9	感情を表に出すタイプ？		2	2	1
Q10	絶叫マシーンが・・・				
Q11	付き合った経験が・・・				
Q12	自分は遠距離恋愛が出来ると・・・			1	1
Q13	やきもち焼きですか？				
Q14	相手色に・・・		1	1	3
Q15	いつでも自分をはい番に考えてほしい？				
Q16	人のケータイを黙ってみたことが・・・				
Q17	一人の時間は大切ですか？				2
Q18	ペア商品を持ってる（今いない人は持ちたい？）		4		
Q19	結婚と恋愛は別だと・・・		3		
Q20	あなたは束縛をする人ですか？				1

表 3 は、Q4 をターゲット属性としたときの ID3 とアプリオリの出現属性を表している。ID3 の列は、属性が出現した属性のノードの高さを表す。アプリオリは、m=2 の列は表 1 のルール 1 のように X⇒Y というように 2 つの属性からなるルール、m=3 の列は表 1 のルール 4 のように X∧Y⇒Z というように 3 つの属性からなるルールの各属性の出現回数を表している。どちらも出現しなかった属性については空欄となっている。

表 3 を見ると、アプリオリでは Q14 が 4 回、Q9 が 3 回出現している。アプリオリの出力した全ルール数が 9 であるなかで、他の属性が 1,2 回しか出現していないのに対しこれらの属性は数多く出現したことから重要属性であるといえる。また、これらの属性は図 1 の決定木の中でも高い位置に出現している（表 3 では大きな数字となっている）。よって、海外に行った経験があるかどうかという属性に対して、アプリオリと ID3 の重要属性は十分一致したといえる。

また同様に Q9、Q17、Q20 をターゲット属性としたときの ID3 とアプリオリの出現属性を表 4、表 5、表 6 に示す。

表 4：感情を表に…のアプリオリ、ID3 の出現属性

		ターゲット	ID3 m=9,C=24 高さ 2,ノード 3	アプリオリ Supp=0.3,Conf=0.6 ルール 21 個	
				m=2	m=3
Q1	あなたの性別は？				1
Q2	異性の友達が・・・		1	1	5
Q3	異性間の友情は・・・				1
Q4	海外に行ったことが・・・			2	1
Q5	さびしがり屋ですか？		2	1	5
Q6	犬と猫ならどっちが好き？		1		2
Q7	自分が好き？				
Q8	今ダイエットを・・・				2
Q9	感情を表に出すタイプ？	○			
Q10	絶叫マシーンが・・・				
Q11	付き合った経験が・・・				2
Q12	自分は遠距離恋愛が出来ると・・・				1
Q13	やきもち焼きですか？				1
Q14	相手色に・・・				
Q15	いつでも自分をはい番に考えてほしい？				
Q16	人のケータイを黙ってみたことが・・・				3
Q17	一人の時間は大切ですか？				3
Q18	ペア商品を持ってる（今いない人は持ちたい？）				
Q19	結婚と恋愛は別だと・・・				4
Q20	あなたは束縛をする人ですか？				3

表 4 の場合においても、Q2 や Q5 はアプリオリで出現回数が多く、ID3 でも出現していることから、重要属性が一致したといえる。アプリオリでは Q19 も出現回数が多いように思うが、m=2 の部分に出現していないことから、minSupport と minConfidence を変えてみれば、アプリオリと ID3 の重要属性の相関関係がもう少し顕著な結果となったのではないかと予想される。

表 5：一人の時間…のアプリオリ、ID3 の出現属性

		ターゲット	ID3 m=9,C=24 高さ 4,ノード 4	アプリオリ Supp=0.3,Conf=0.6 ルール 6 個	
				m=2	m=3
Q1	あなたの性別は？				
Q2	異性の友達が・・・				
Q3	異性間の友情は・・・		1	1	1
Q4	海外に行ったことが・・・				
Q5	さびしがり屋ですか？				
Q6	犬と猫ならどっちが好き？				
Q7	自分が好き？				
Q8	今ダイエットを・・・			1	
Q9	感情を表に出すタイプ？				
Q10	絶叫マシーンが・・・				
Q11	付き合った経験が・・・		3	1	1
Q12	自分は遠距離恋愛が出来ると・・・				
Q13	やきもち焼きですか？		2		
Q14	相手色に・・・				
Q15	いつでも自分をはい番に考えてほしい？				
Q16	人のケータイを黙ってみたことが・・・		4	1	
Q17	一人の時間は大切ですか？	○			
Q18	ペア商品を持ってる（今いない人は持ちたい？）				
Q19	結婚と恋愛は別だと・・・				
Q20	あなたは束縛をする人ですか？			1	

表 5 の結果では、アプリオリと ID3 両者で出現属性が少なかったが、それらがほぼ一致している。つまり、この場合においても重要属性が一致したといえるだろう。

表 6：束縛のアプリオリ、ID3 の出現属性

		ターゲット	ID3 m=9,C=24 高さ 5,ノード 5	アプリオリ Supp=0.3,Conf=0.6 ルール 25 個 m=2 m=3
Q1	あなたの性別は？			1
Q2	異性の友達が・・・			1
Q3	異性間の友情は・・・			1
Q4	海外に行ったことが・・・			2
Q5	さびしがり屋ですか？		4	2
Q6	犬と猫ならどっちが好き？		2	1
Q7	自分が好き？			2
Q8	今ダイエットを・・・			1
Q9	感情を表に出すタイプ？		1	2
Q10	絶叫マシーンが・・・			1
Q11	付き合った経験が・・・		3	1
Q12	自分は遠距離恋愛が出来ると・・・			2
Q13	やきもち焼きですか？			2
Q14	相手色に・・・			1
Q15	いつでも自分をはい番に考えてほしい？		5	1
Q16	人のケータイを黙ってみたことが・・・			1
Q17	一人の時間は大切ですか？			1
Q18	ペア商品を持ってる（今いない人は持ちたい？）			1
Q19	結婚と恋愛は別だと・・・			1
Q20	あなたは束縛をする人ですか？	○		

表 6 では、アプリオリの出力したルール数が多く、重要属性がわかりにくい結果となってしまった。minSupport と minConfidence の設定を変え、ルール数の調整をして m=3 の場合も求めて比較を行えば重要属性がわかり、他のターゲット属性の場合と同様な結果が得られたのではないかと考えられる。

このように、いくつかのターゲット属性に関しても比較を行ったところ、アプリオリと ID3 の出力するルールの重要な属性は十分一致するといえることが分かった。

2 つ目は、データに対しデータマイニングし、両者の出力に出現した属性を比較、適合率を求める。適合率を求める式を以下に示す。

- アプリオリの適合率=両者に出現した属性数/アプリオリが出力した属性数
- ID3 の適合率=両者に出現した属性数/ID3 が出力した属性数

これらを３つのターゲット属性について求めた結果を表７に示す。

表 7：出現属性の適合率

ターゲット属性	アプリオリ	ID3
Q4	2/7	1/2
Q7	3/5	3/4
Q9	3/14	1
平均	0.37	0.75

アプリオリの平均適合率は 0.37、ID3 の平均適合率は 0.75 となった。全属性数が 20 であることを考慮すると、ターゲット属性 Q7 の例のように出力した 4,5 個の属性のうち半分以上が一致しているというのは高い数字であるといえる。他のターゲット属性ではアプリオリの適合率が低く出てしまったが、これは一般的にアプリオリの方が出力されるルールが多く出るためである。

3.5　論理関係の比較

属性の種類だけではなく、抽出された規則の論理的な矛盾がないかどうかを検討するため次の実験を行った。100 件の集めたデータの中から、Y と N の比率が 50:50 に近い 4 つの属性（Q4. 海外に行ったことがあるか、Q7. 自分が好きか、Q9. 感情を表に出すタイプか、Q7. 相手色に染まるか）を選び、それらの属性だけからなるデータ D′ に対して、アプリオリ、ID3 を適用して一致を調べる。アプリオリの条件は minsupp =0.2、minconf =0.5 で、ルールは 30 個生じた。そこで出てきたルールを conf において高い順に並び替え、その中から答えが海外（決定木のクラス）に関するものだけを抜き出したのが、表 8 である。決定木に対応する葉を最後の列に示しており、括弧は部分的に一致している葉である。

表 8: アプリオリから抽出された相関ルール

ルール	supp	conf	順位	決定木の葉
感情出す ∧ 染まらない ⇒ 海外ある	0.2	0.8	1位	f
自分嫌い ∧ 感情出さない ⇒ 海外ない	0.2	0.6897	2位	(e)
染まる ⇒ 海外ない	0.28	0.6364	3位	(a,c)
感情出す ⇒ 海外ある	0.31	0.6327	4位	(f)
染まらない ⇒ 海外ある	0.35	0.625	5位	(d,f)
感情出さない ⇒ 海外ない	0.31	0.6078	6位	(e,c)
自分好き ⇒ 海外ある	0.29	0.6042	7位	(b)
自分嫌い ⇒ 海外ない	0.3	0.5769	8位	(a,e)

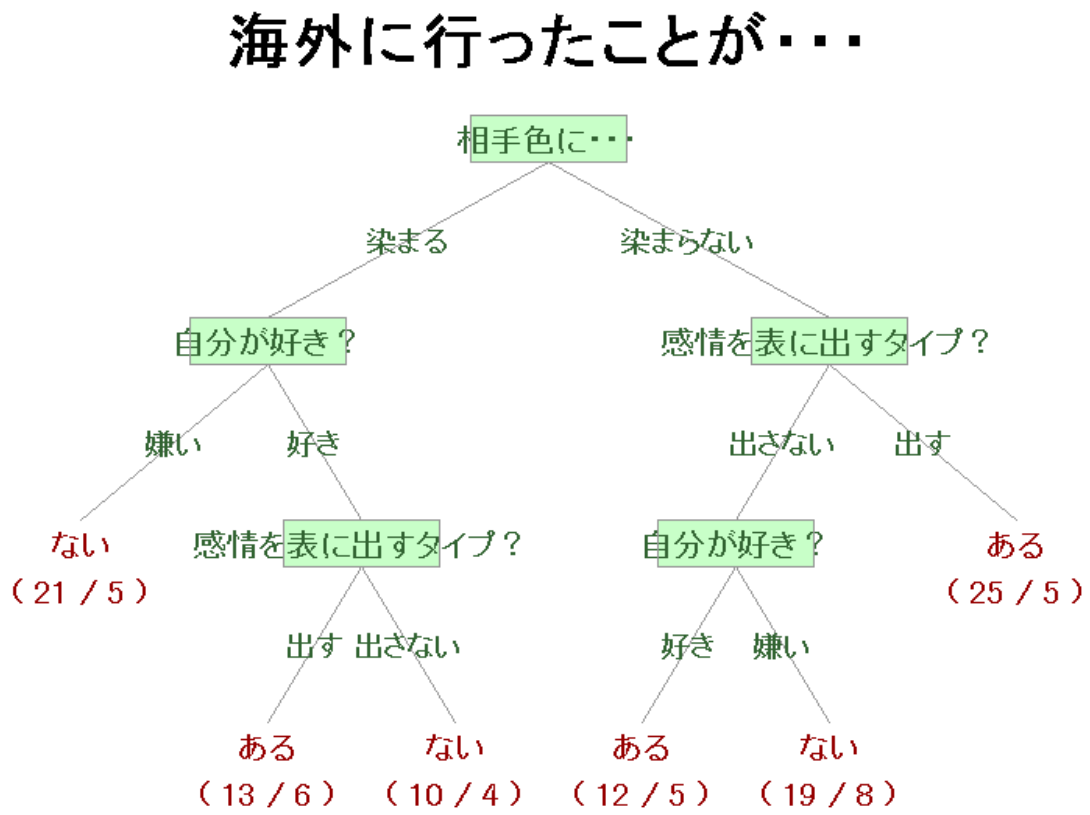


図 4: 「海外旅行経験」の決定木(2)

表 9： 決定木から抽出された相関ルール

	supp	conf
a	0.16	0.76
b	0.07	0.53
c	0.06	0.6
d	0.07	0.58
e	0.11	0.57
f	0.2	0.8

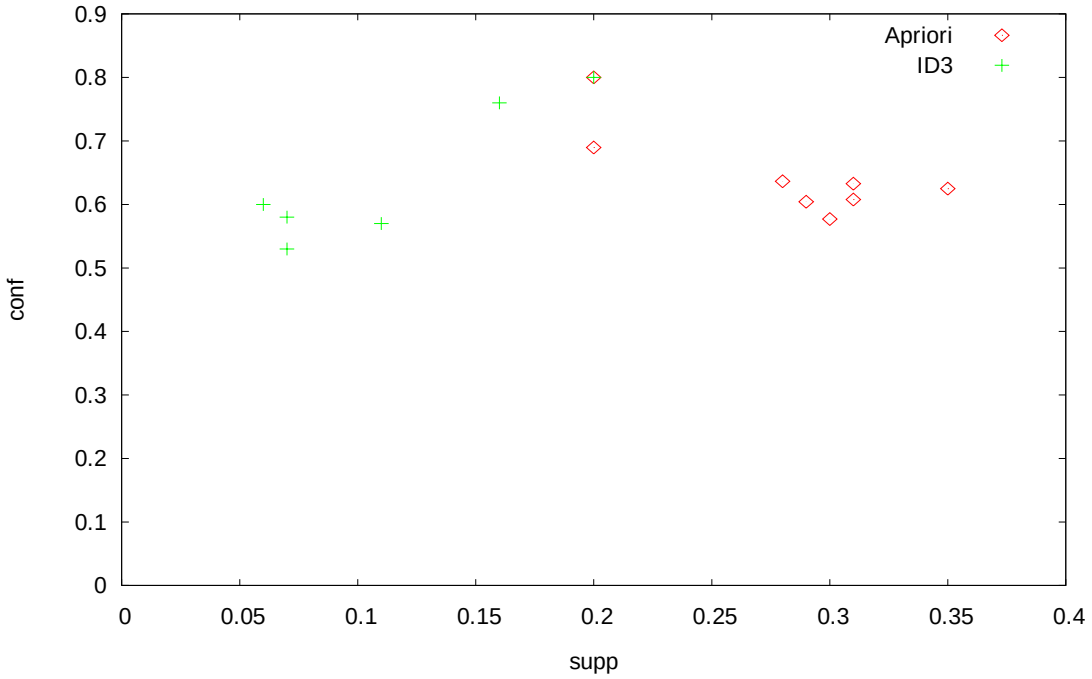


図 5: サポートと確信度についての散布図

次に、図 4 の ID3 の決定木の葉に左から順に番号を振り、supp と conf を表 9 に示した。図 5 は、表 8 から得られた結果から X 軸を supp、Y 軸を conf とする散布図の上に、二種類のルールをプロットした。それらを比較する。図 5 を見て分かるように、アプリオリと ID3 とが一致するもの(表 8 ルール 1 と表 9 の葉 f) があることが分かる。木の枝狩りがうまくいけば、もう少し一致することが予想される。また、ID3 は supp が低く conf が高いものが与えられる傾向が高い。アプリオリの条件で supp を下げれば ID3 と一致するより多く生じると考えられる。

2.6 考察

出現属性の比較では、ID3 で高い位置に出現した属性はアプリオリでも出現回数が多かったこと、適合率が高い数値であったことから、両者の出力する属性にはある程度相関があることが分かった。

論理関係の比較では、アプリオリに出力されたルールと ID3 の出力の葉に矛盾がないことから、両者の出力するルールに矛盾が生まれることはないことが分かった。ID3 の葉に一致するアプリオリのルールは少なかったが、決定木から抽出される規則のサポートはアプリオリよりも低いためである。これは、アプリオリの相関ルールがそれぞれ独立なのに対し、決定木は全体で一つの矛盾のない木を構成しなければならないからだと考えられる。

第４章
評価ツールの開発

4.1　概要
両アルゴリズムでの属性の比較にあたり、その一部を自動で行うツールを作成した。

4.2　機能

Java ソース名	機能	入力ファイル（拡張子）	出力ファイルの拡張子
tocsv.java	Gather の出力ファイルを csv ファイルに直す。	Gather の出力ファイル(txt)	csv
appear.java	出現属性の比較を行う。	属性テーブル(csv) tocsv の出力ファイル(csv)、 ID3 の出現属性(txt)	txt
Logic.java	アンケートの回答が半分程度に割れた属性部分を抜き出す。	アンケートデータ(csv)	csv

4.3　使用方法
［出現属性の比較］

・データマイニングする

本ツールは研究室の卒業生のツールを使用し、それらの結果を比較する仕様になっている。まずデータを用意し、Gather、ID3E を使用し、結果（例として図 6、図 3 参照）を保存する。Gather では minSupport と minConfidence の設定、ID3E では枝狩りができるが、両者の出現属性の数を本論文を参考に調整してほしい。

図 6：Gather の出力ファイル例

・アプリオリでの出現属性をまとめる

tocsv.java を実行し、Gather の結果を引数で指定すると、テキストファイルに文章で表示されていた相関ルールが csv ファイルの表にまとめられる。Excel で表示し、相関ルールを「仮定⇒結論」としたとき、結論部の列を昇順（降順でも良い）に並び替え、ターゲット属性が結論となっている相関ルールの仮定部をテキストファイルにコピーアンドペーストする（図 7 参照）。

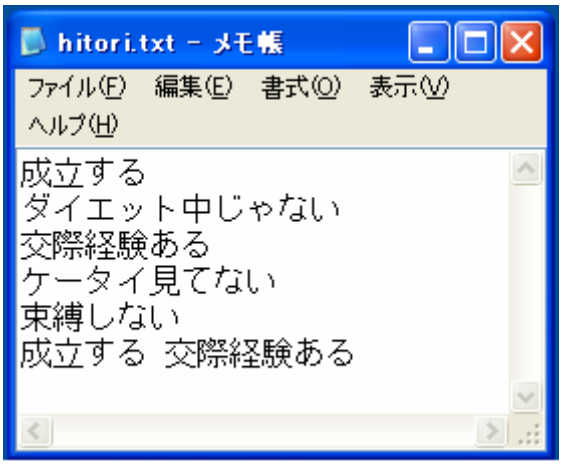


図 7：仮定部をコピーしたファイル例

・属性と属性値の表を用意する

属性（質問）と属性値（回答の種類）を表にして保存する（図 8 参照）。

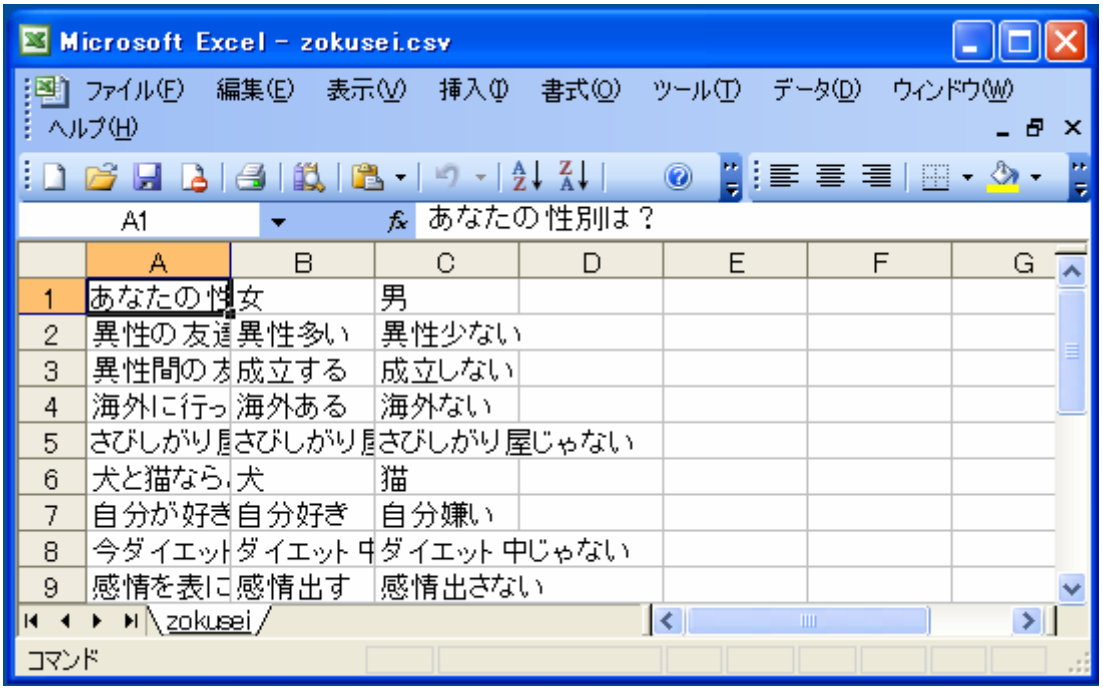


図 8：属性テーブル例

・ID3E での出現属性をまとめる

ID3E での出現属性をテキストファイルに手動でまとめる。このとき、高い位置のノードから書いていくと、属性の重要さがわかりやすく、結果がわかりやすい。

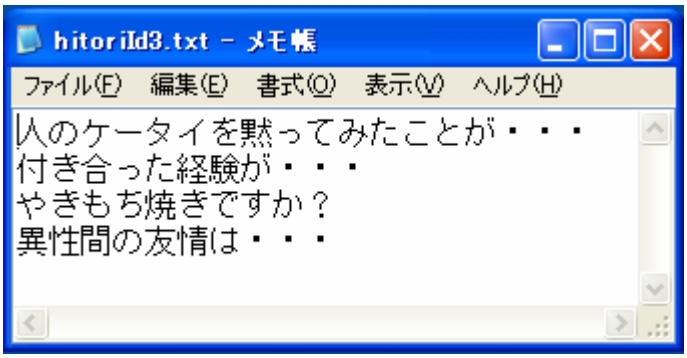


図 9：ID3E 出現属性ファイル例

・比較をする

appear.java を実行し、属性と属性値の表、アプリオリでの出現属性、ID3E での出現属性を引数としてあたえる（図 10 参照）と、各属性のアプリオリでの出現回数、ID3E での出現属性がアプリオリで何回出現しているかが表示され、また同様の内容をアプリオリでの出現属性に「出力結果.txt」が追加されたファイル名で出力する（図 11 参照）。

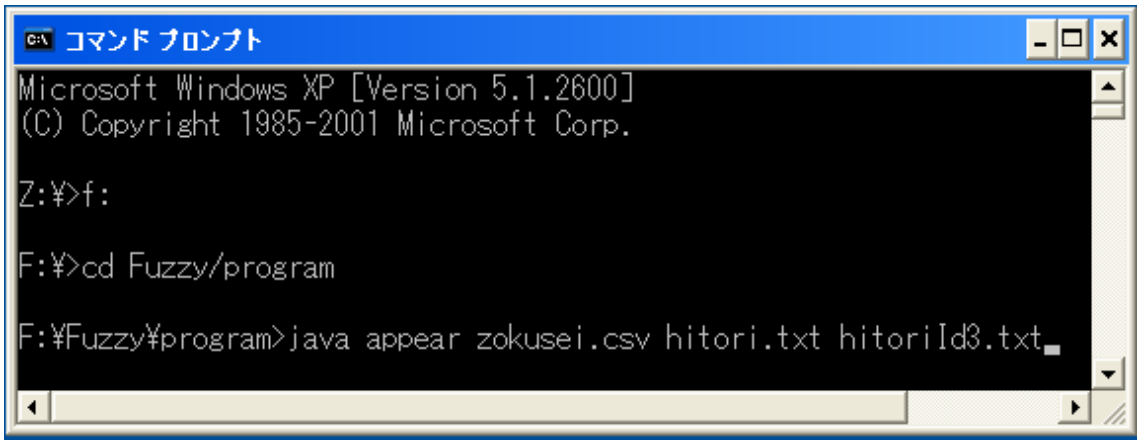


図 10：appear.java 実行画面

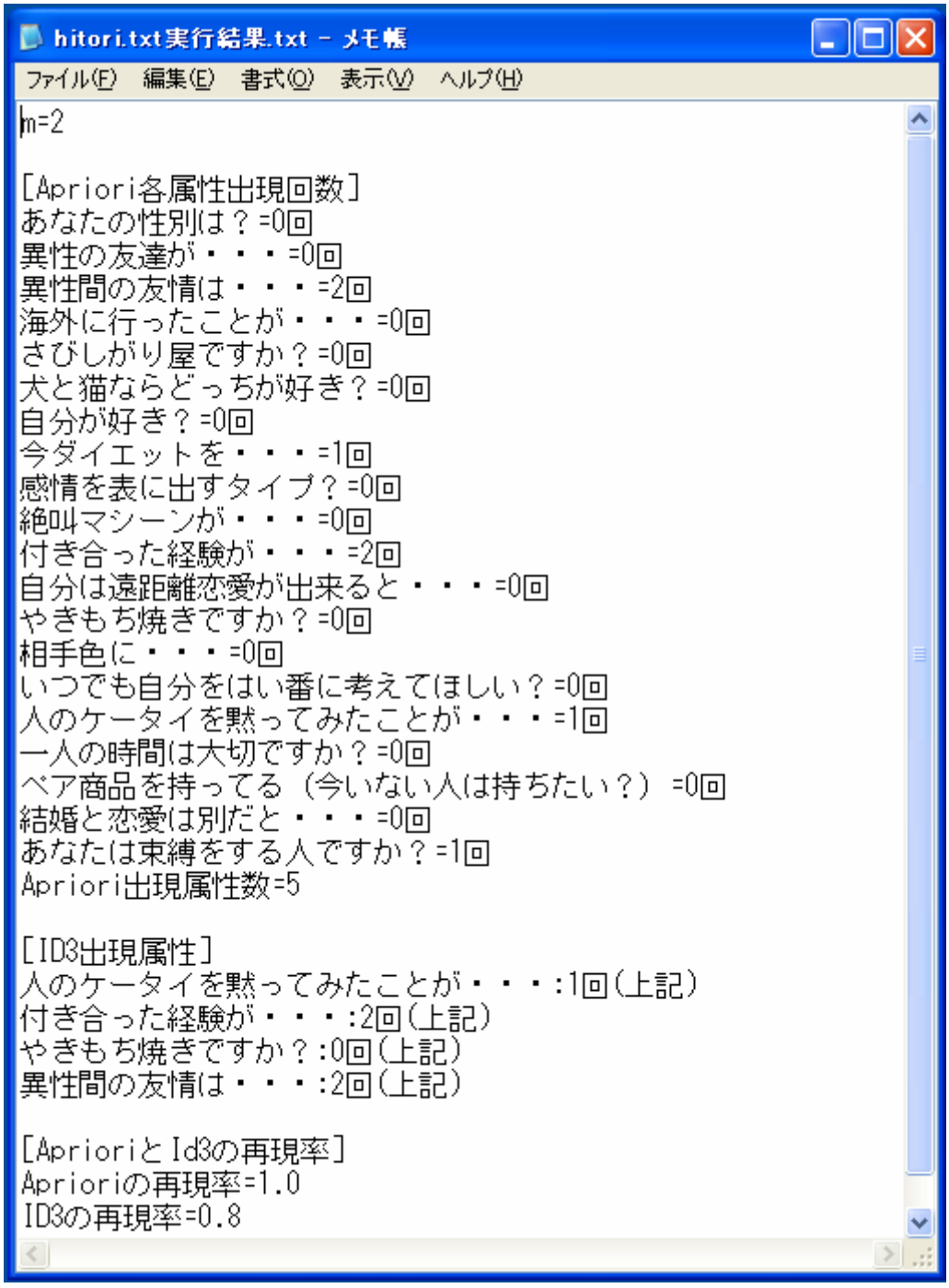

hitori.txt実行結果.txt - メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
m=2
[Apriori各属性出現回数]
あなたの性別は？=0回
異性の友達が・・・=0回
異性間の友情は・・・=2回
海外に行ったことが・・・=0回
さびしがり屋ですか？=0回
犬と猫ならどっちが好き？=0回
自分が好き？=0回
今ダイエットを・・・=1回
感情を表に出すタイプ？=0回
絶叫マシーンが・・・=0回
付き合った経験が・・・=2回
自分は遠距離恋愛が出来ると・・・=0回
やきもち焼きですか？=0回
相手色に・・・=0回
いつでも自分をはい番に考えてほしい？=0回
人のケータイを黙ってみたことが・・・=1回
一人の時間は大切ですか？=0回
ペア商品を持ってる（今いない人は持ちたい？）=0回
結婚と恋愛は別だと・・・=0回
あなたは束縛をする人ですか？=1回
Apriori出現属性数=5
[ID3出現属性]
人のケータイを黙ってみたことが・・・:1回(上記)
付き合った経験が・・・:2回(上記)
やきもち焼きですか？:0回(上記)
異性間の友情は・・・:2回(上記)
[AprioriとId3の再現率]
Aprioriの再現率=1.0
ID3の再現率=0.8

図 11：出力ファイル例

[論理関係の比較]

論理関係の比較の実験では回答が半分程度（45～55）に分かれている属性を使用している。そこで、Logic.java を実行し入力ファイルを引数で指定する（図 12 参照）と、それらの属性とターゲット属性のみの csv 形式（カンマ区切り）のデータが出力される（図 13 参照）ようになっている。このデータを Gather と ID3E の２つのツールにかけ、結果を手動で比較する。ただし、このプログラムは２択のアンケートを想定している。

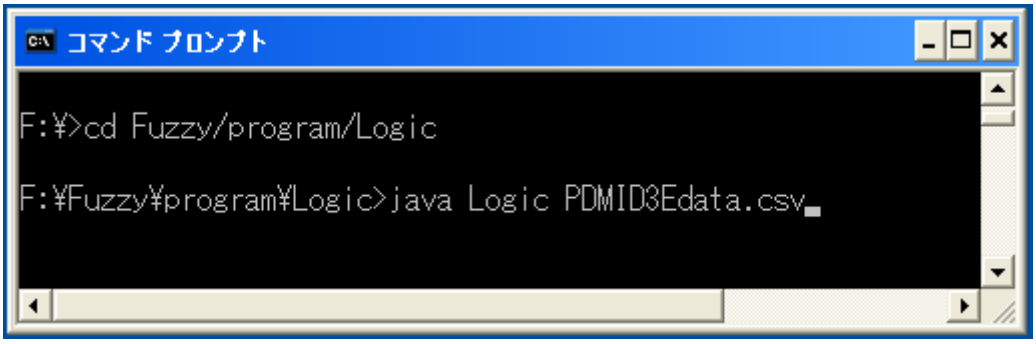
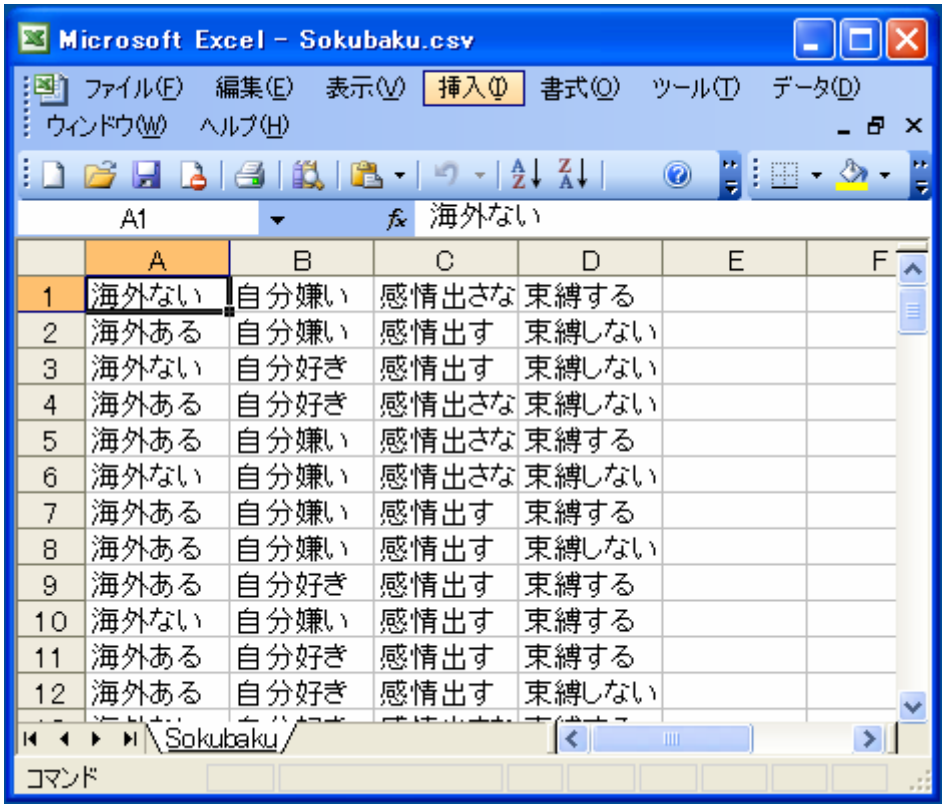


図 12：Logic.java 実行画面



	A	B	C	D	E	F
1	海外ない	自分嫌い	感情出さな	束縛する		
2	海外ある	自分嫌い	感情出す	束縛しない		
3	海外ない	自分好き	感情出す	束縛しない		
4	海外ある	自分好き	感情出さな	束縛しない		
5	海外ある	自分嫌い	感情出さな	束縛する		
6	海外ない	自分嫌い	感情出さな	束縛しない		
7	海外ある	自分嫌い	感情出す	束縛する		
8	海外ある	自分嫌い	感情出す	束縛しない		
9	海外ある	自分好き	感情出す	束縛する		
10	海外ない	自分嫌い	感情出す	束縛する		
11	海外ある	自分好き	感情出す	束縛する		
12	海外ある	自分好き	感情出す	束縛しない		

図 13：Logic.java 出力ファイル例

第 5 章
おわりに

アプリオリと ID3 の出力結果を比較し、出現属性や論理関係には相関があるのか確かめた。出現する属性に関しては、決定木で上位に出現する属性はアプリオリでも目的属性と相関が高く、平均値でそれぞれ 0.75 と 0.35 の適合率であった。また、抽出された知識の論理的な関係には矛盾はないが、アプリオリで出力された相関ルール 8 つのうち決定木に完全に対応する葉は 1 つしかなかった。これは、アプリオリの相関ルールがそれぞれ独立なのに対して、決定木は全体で一つの矛盾のない木を構成しなければならないところに起因していると考えられる。決定木から抽出される規則のサポートはアプリオリよりもかなり低いことも原因の一つである。これらのことから、アプリオリと ID3 の出力結果に、出現属性や論理関係の相関があると確かめられた。

参考文献

[1]　本多，アプリオリアルゴリズムに基づいたデータマイニングツール”Gather”の開発，東海大学情報メディア学科 2005 年度卒業研究論文.

[2]　松岡，データマイニングによる授業評価アンケートの解析，日本知能情報ファジィ学会，気持ちのワークショップ 2003.

[3]　福田，森本，徳山，データサイエンス・シリーズ(3)データマイニング，共立出版，(2001)

[４]阿久津，倉野，「データマイニングアルゴリズム“アプリオリ”と“ID3”の比較」，日本知能情報ファジィ学会，曖昧な気持ちに挑むワークショップ，pp. 1-5，2006

謝辞

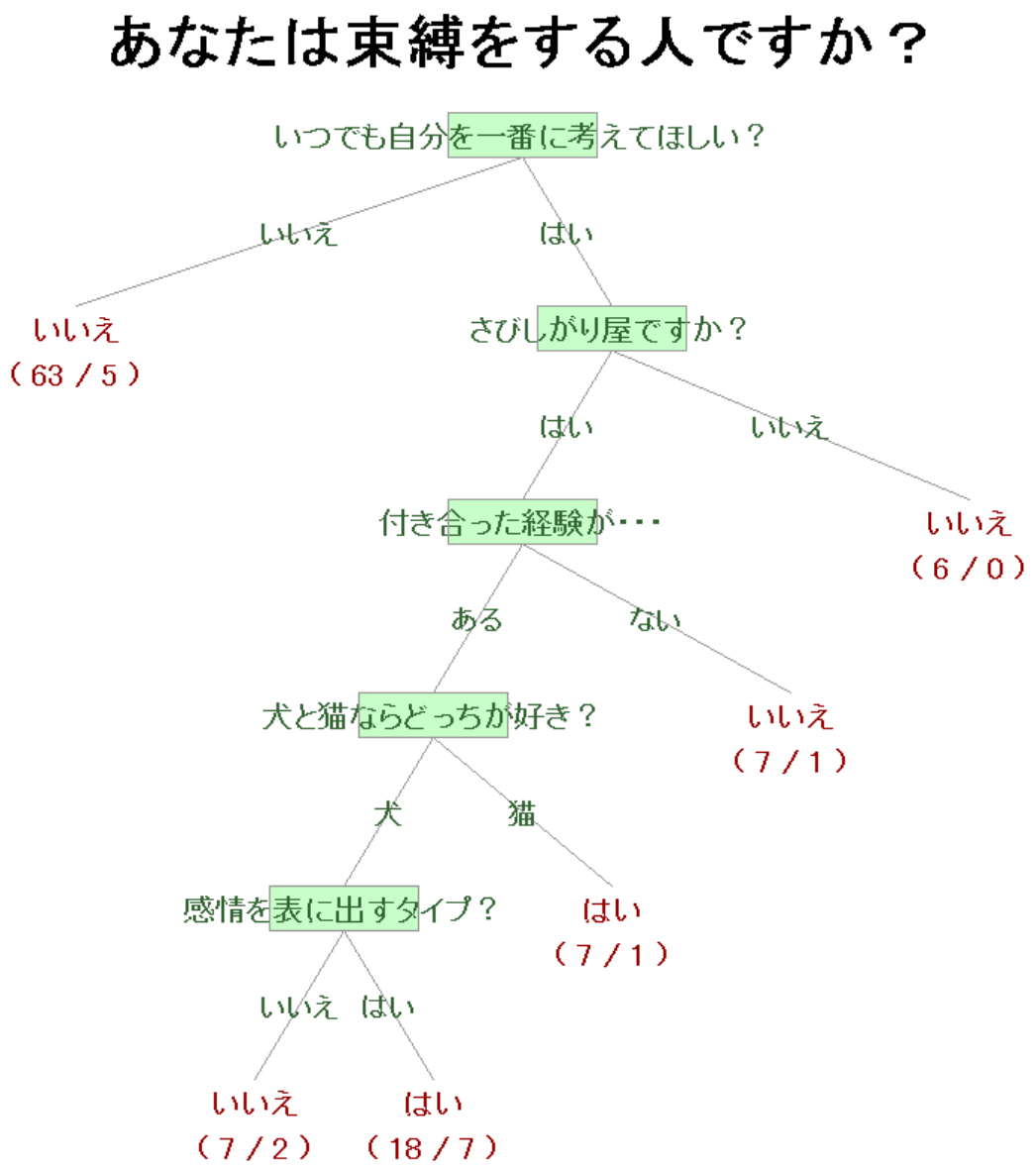
本研究を完遂するにあたり、多大なるご指導を受け賜りました東海大学電子情報学部情報メディア学科菊池浩明助教授に心より感謝申し上げます。

また、多大なるご指導を賜り、本研究を導いて頂きました三橋将氏、ツールを使用させていただいた本多淳司氏、並木翼氏に深くお礼申し上げます。

そして、菊池研究室の同期の皆さんに感謝の意を述べると共に、謝辞とさせて頂きます。

付録１

今回アンケートを取るにあたって、束縛するかどうかをターゲット属性とした。大学生の恋愛感について興味があったためだ。その ID3E の結果をここに示す。



付録２

今回のデータマイニングから発見された相関ルールを用いて、Psychology Data Mining という心理テストを夏の課題として作成した。興味のある方は以下の URL よりどうぞ。

http://www.cs.dm.u-tokai.ac.jp/~kitakita/OfficialPage/kadai/index.html