岡山理科大学紀要 第39号 A pp. 77-86(2003)

機械学習を中心としたデータマイニング

津田 倫彰・成久 洋之*

岡山理科大学大学院工学研究科修士課程情報工学専攻

*岡山理科大学工学部情報工学科

(2003年11月7日 受理)

1. まえがき

データマイニング(データ発掘)[1],[2]とはデータベース(Data Base)に保管されている属性値(Attribute)のような生のデータ群から有益な情報や知識を抽出することである。従来の情報探索(Information retrival)とは少しニュアンスが違い、Data mininng,Knowledge discovery(知識発見)と呼ばれる非常に注目されている研究分野である。このデータマイニングは1989年に American Association for Artificial Intelligence(AAAI)のWorkshop on Knowledge Discovery in Databases以降にこの用語が定着するようになったものとされている。したがって、人工知能分野から派生したものであるが、データベースや機械学習さらには統計学にも関連した学際的(Interdisciplinary)研究領域とも考えられる。最近では理論的研究の基礎段階を超えて、ビジネスの実践段階に入っているものもかなり見受けられている。これらの中での代表的なものとしてはマーケティングへの活用であり、顧客情報の分析結果を将来の企業戦略に生かそうとするものである。

本論文は機械学習(Machine Learning)で提案された決定木(Decidion tree)などの学習アルゴリズムを中心としたデータマイニングにつき、その概要を記述し分割統治法(Divide and conquer)を用いたアプローチで属性22個からなる約8200個のMushroomデータから知識を抽出し、その有効性につき検討したものである。

2. 機械学習

2－1 機械学習

機械学習の分野はコンピュータの出現以来諸種の学習を実現できる(特にデータや問題例から知識を抽出するメカニズムを持った)数学的手法であると考えられてきた。しかしながら今日の先端コンピュータ技術においてはこのような知識導出の問題は、そのソフトウェア開発でのボトルネックと考えられてきた。そこに台頭したのが人工知能分野の研究であり、従来考えられてきたソフトウェアとしてのプログラムを

program = algorithm + data

↓

program = alogrithm + data + domain knowledge

とすることで対象とする問題領域固有知識(domain knowledge)の導入により問題解決を計ろうとするものである。すなわち人工知能においてプロダクションルール(Production rule)やフレーム(Flame)、セマンティックネットワーク(Semantic network)で表される知識に基づいた情報処理はかなり有効なものとして知られている。

しかしながら、このことは知識の導入がプログラマのボトルネックからknowledge engineer にシフトしただけに過ぎないのではないかという見方もある。その理由は現実世界の応用において知識獲得とその符号化のプロセスは非常に困難であるからである。

2－2 機械学習システム

機械学習に対する一般的な枠組みは次図のとおりである。

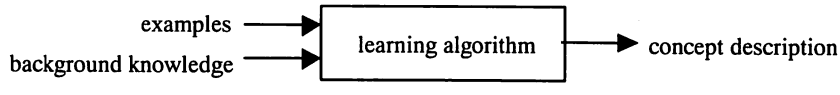

図1. 機械学習のフレームワーク

学習システムは教師やその分野の背景となる知識からなる概念例(concept examples)の集合から与えられた概念記述を決定するものである。background knowledgeは問題例や概念を記述するための言語についての情報を含んでいる。例えば、属性の可能な値やその階層、述語、補助的文法規則、主観的な好みなどがある。

学習アルゴリズムは大別して2つの手法に分類される。一つはニューラルネットワークや統計学のようなブラックボックス法、もう一つは知識依存法(knowledge-oriented method)である。ブラックボックス法は主に概念認識に使用され、知識依存法は理解可能性の原理を満たす記号認識構造(symbolic knowledge structure)を使用している。

2－3 知識表現法

機械学習で問題例や概念を表現するための言語として以下の各種論理言語が使用されている。

(1) 0階論理(Zero Order Logic)あるいは命題論理(Propositional Logic)
(2)属性論理(Attributional Logic)
(3) 1階述語論理(First Order Predicate Logic)またはホーン節(Horn Classes)
(4) 2階述語論理(Second Order Predicate Logic)

命題論理は論理記号と命題定数のみで表される。

$$C \leftarrow X \wedge Y \wedge Z$$

(概念 CはX,Y,Zの条件が成立するときのみ真である)

属性論理は基本的には命題論理と同等であるが柔軟性にとんだ豊富な表現を期待できる。これは命題変数、命題定数を使用するもので、属性を命題変数と見なしていることである。この属性論理は機械学習での記述言語としてはかなり実用的なものとされており、TDIDT(Quilan)やAQアルゴリズム(Michalski)などが良く知られている。

1階述語論理は対象物やその部分である対象物間の関連性などについての記述や理由付けのための公的な枠組みを具備したもので、関数定数、述語変数、個体変数などに使用するものである。ホーン節はheadとbodyから構成される。以下に例をしめす。

$$grandparent(X,Y) \text{:-} \ parent(X,Z), parent(Z,Y)$$

($grandparent(X,Y)$:head, $parent(X,Z), parent(Z,Y)$:body)

これはXがZの*parent*であり、ZがYの*parent*であるような*person* Zが存在すればXはYの*grandparent*であるとする。このときX,Y,Zは限定変数となっている。また、*grandparent*や*parent*は述語であり( )のなかの変数は*argument*(引数)とよばれその数は任意であるけれども与えられた述語については固定される。もし述語が正確に1個の*argument*であれば、属性論理となり、全ての述語が*zero argument*ならばその言語は命題論理となってしまう。

2階述語論理が述語変数や関数変数を持つ一般的な体系である。たとえば、

$$p(X,Y) \text{:-} q(X,XW) \wedge q(Y,YW) \wedge r(WX,WY)$$ (p:brother,q:son ,r:equal)

をp:*brother*,q:*son* ,r:*equal*すると、

$$brother\ (X,Y) \text{:-}\ son\ (X,XW) \wedge son\ (Y,YW) \wedge equal\ (WX,WY)$$

となる問題例と同等である。

2－4 解探索

知識表現言語が決定され学習者がデータ列から概念を学習するものとしても、その記述言語に基づく探索空間は巨大なものとなってしまう。まして、高階かつ複雑な言語記述では想像を絶するものになりうる。このような巨大な探索空間に対して有効な探索戦略としては推論によるか、あるいはヒューリスティックな探索しかないとされている。

概念学習の広義な枠組みは学習者の表現言語で記述されている可能空間での探索である。この探索手法は人工知能の研究分野で広く検討されているものである。これには幅優先と深さ優先の探索戦略がとられている。また、ヒューリスティックス探索では最良解優先アルゴリズムと山登り法のようなビーム探索アルゴリズムが考えられている。

本研究は、属性論理表現に対して代表的な学習手法として考えられている分割統治法によるデータマイニングについて検討する。これは決定木を生成するための最もポピュラーなアルゴリズムであり、1986年に

Quinlanにより提案されたもので、TDIDT(Top-Down Induction of Decision Tree)あるいはID3として知られている。

3. データマイニング

データマイニングはデータベースにある莫大な量のデータから知識を抽出することである。これはデータベースに含まれる構造的なパターンの発見やデータに含まれる構造の発見や記述に関するものであり、論理的でなく現実的学習を含むトッピックスといえる。すなわち、本論は現実的学習法としてのデータマイニングにつき記述する。

3－1 発見された知識の望ましい特性

データマイニングで抽出された知識は次の特性を持たなければならない。

1. 正確であること。(*Accurate*)
2. 理解できるものであること。(*Comprehensible*)
3. 新規性に富み、興味深いものであること。(*Interesting*)

これらの各特質は知識の質的評価尺度とも考えられるが、それらの根本的な重要さは解決すべき問題の種類や適用領域に依存する。

3－2 データマイニングにおける代表的表現方法

(a)決定表(Decidion Tables)
(b)決定木(Decidion Tree)
(c)分類ルール(Classification Rules)
(d)相関ルール(Assosiation Rules)
(e)最近傍表現法(Instance-Based Representation, Nearest Neighbor Method)

(a)の決定表は条件と行動(決定)との関係を表した表であり、表1. のようなものでplayできるか否かを決定するための条件が表されている。

(b)の決定木は木構造で知識を表したものであり、木におけるノード(node)は特定の属性を示し、葉(reaf)は分類上のクラスを表すものである。

(c)の分類ルールは*if*～*then*ルールで表現され、ルールの前件は属性などに関する条件、後件は分類クラスを与える結論を示している。

(d)の相関ルールは分類ルールとほとんど違いはなく、分類ルールのクラスを予測するのではなく属性やその組み合わせを予測できる。このルールにおけるカバーレッジ(coverage)はそのルールが正しく予測できる問題例の数であり、これをサポート(support)と呼んでいる。その信頼度はコンフィーデンス(confidence)とよばれ、全問題例に対して予測できる数の割合をいう。

3－2 データマイニングの代表的な手法

(ⅰ)単純推論法, IR法(Infering rudimentary rules)
これはとも呼ばれ深さ1の決定木を生成し、1属性ごとに全てのルールを表すもので最も単純な方法である。
(ⅱ)統計的モデル法(Statistical Modeling)
これは全体の問題例から各属性の統計量を求めるものである。
(ⅲ)分割統治法(Divide and Conquer)
これは決定木を構成する方法である。
(ⅳ)被覆法(Covering algorithm)
これはルールを構築する方法である。
(ⅴ)相関ルール法(Assosiation rules Method)
これは知識としての相関ルールを生成してデータマイニングを行うものである。
(ⅵ)問題例依存学習法(Instance-based learning)
知識表現として最近傍法を利用したものである。

4. 分割統治法による決定木の作成

4-1決定木

決定木とはあるデータから得られた知識やルールを人にわかりやすく表したものであり、図2のような形をしている。図中の最上部の根と呼ばれるもの以外の先端部分は葉と呼ばれ、クラス（導出したい結果を表す要素）が入る。そして、根と葉をつなげる部分は枝と呼ばれ、属性（クラスを導出するまでの条件を表す要素）が入る。

決定木は単純な形をしたものほど優れた木といわれている。その理由は簡単である。木の高さ（根から葉までの枝の総数）が浅いほどルールは簡素であり、葉の数が少ないほど個々のルールの価値が高まる。例えば、車（クラス）を製造するときに組み込むべき必要最小限の部品（属性）が判れば、車の製造コストは抑えることが可能となり車の製造会社は無駄を省いたという利益が生まれる。

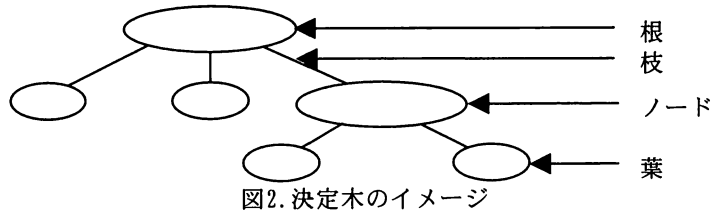

図2. 決定木のイメージ

4－2 平均情報量

平均情報量とは、曖昧さや不確かさを表す尺度で（1）から（3）式のように定義されている。（3）式において $Info(A)=0$ となるとき確率の集合 A は確実に発生し、$\log_2 m$ に近づくほど不確実さは増していく。

$$Info(A) = -\sum_{i=1}^{m} P_i \log_2 P_i \quad \cdots \quad (1)$$

$$\sum_{i=1}^{m} P_i = 1 (i = 1,2,...,m) \quad \cdots \quad (2)$$

$$0 \leqq Info(A) \leqq \log_2 m \quad \cdots \quad (3)$$

4-3 分割統治法

分割統治法とは問題に対してある条件に従って小さな問題と解に分割し、分割した問題に対して再び同じ条件に従い小さな問題と解を得る、という処理を繰り返して行い、処理の際に生じた解を1つの解としたときに最初の問題を満たせば上記の繰り返し処理は終了する。決定木作成において分割する際には平均情報量を使って問題を分割していく。

4-4 アルゴリズム

分割統治法のアルゴリズムを示す。

Step0. 初期化

属性の集合$A=\{A_1,A_2,...,A_i,...,A_m\}$($m$:属性数)として、属性$A_i$の属性値$a_i=\{a_{i1},a_{i2},...,a_{ij},...,a_{in}\}$（$n$:属性値の数）、クラスの値$C$を$C=\{c_1,c_2,...,c_k,...,c_z\}$（$z$:クラス値の数）、データ集合$D=\{d_{xy}\}$（$x$:データ、$y$:属性）とする。

Step1. Dにデータを読み込ませる。

Step2. 各属性に対応するデータの数え上げ

データ集合Dから属性A_iの属性値a_iに対応するものを数え上げ数え上げたものを$sac(i,j,k)$とする。

Step3. 平均情報量と相互情報量の計算

Step2. で得られたsacを用いて各属性値の平均情報量を導出する。$SS(j)$は属性A_jに対応したクラスの総数、$PP(j,k)$はクラスCの確率の総和を表す。(7), (8), (9)式は相互情報量の計算を表し、$gain(A_j)$の値が大きいほど

信頼性が高い。

$$S(i,j)=\sum_{k=1}^{z} sac(i,j,k) \quad \cdots (4)$$

$$p(i,j,k)=\frac{sac(i,j,k)}{S(i,j)} \quad \cdots (5)$$

$$Info(i,j)=\sum_{k=1}^{z} P(i,j,k)\log_2 P(i,j,k) \quad \cdots (6)$$

$$Info_{ave}(i,j)=\sum_{j=1}^{n}\frac{S(i,j)}{SS(j)} Info(i,j) \quad \cdots (7)$$

$$Info_0(A_i)=\sum_{j=1}^{n} PP(i,j)\log_2 PP(i,j) \quad \cdots (8)$$

$$gain(A_i)=Info_0(A_i)-Info_{ave}(A_i) \quad \cdots (9)$$

Step4. 葉の決定。

(10), (11)で最も信頼性の高い属性が選ばれ、(13), (14)で葉が決定される。

$$Gain_{\max}=\max\{gain(A_i)\} \quad \cdots (10)$$

$$i_{\max}=\max\{i \mid gain(A_i)=Gain_{\max}\} \quad \cdots (11)$$

$$Info_{\max}=\max\{Info(i,j)\} \quad \cdots (12)$$

$$j_{\max}=\{j \mid Info(i,j)=Info_{\max}\} \quad \cdots (13)$$

$$j_0=\{j \mid Info(i,j)=0\} \quad \cdots (14)$$

Step5. 全てのノードの接続先がクラスの値になるまでStep2からStep4. を繰り返す。

このアルゴリズムを表1で与えられる天候データの問題に適用してみる。

表1. 天候データ

No	outlook	temperature	humidity	windy	play
1	sunny	hot	high	false	no
2	sunny	hot	high	true	no
3	sunny	cool	normal	false	yes
4	sunny	mild	high	false	no
5	sunny	mild	normal	true	yes
6	rainy	cool	normal	true	no
7	rainy	cool	normal	false	yes
8	rainy	mild	high	false	yes
9	rainy	mild	high	true	no
10	rainy	mild	normal	false	yes
11	overcast	hot	high	false	yes
12	overcast	hot	normal	false	yes
13	overcast	cool	normal	true	yes
14	overcast	mild	high	true	yes

Step0. 初期化

属性をそれぞれ

$\{A_1,A_2,A_3,A_4\}$ = {outlook, temperature, humidity, windy}

$\{a_{11},a_{12},a_{13}\}$ = {sunny, rainy, overcast}

$\{a_{21},a_{22},a_{23}\}$ = {hot, mild, cool}

$\{a_{31},a_{32}\}$ = {high, normal}

$\{a_{41},a_{42}\}$ = {true, false}

とし、クラスは

$\{C_1,C_2\}$ = {yes, no}

Step1. データの読み込み

全データ(No1からNo14)までのデータをDとする。

Step2. 各属性値のsacの導出

表2. のように各属性値sacを数え上げる。

表2. 属性outlookに関するsacの数え上げ

	yes	no	total
sunny	2	3	5
overcast	4	0	4
rainy	3	2	5
total	9	5	14

Step3. 平均情報量と相互情報量の計算

Dから得られる平均情報量、相互情報量は以下の通り

$$Info(A_1,a_{11}) = -2/5\times\log_2(2/5) - 3/5\times\log_2(3/5) = 0.970950594$$
$$Info(A_1,a_{12}) = -4/5\times\log_2(4/5) = 0$$
$$Info(A_1,a_{13}) = -3/5\times\log_2(3/5) - 2/5\times\log_2(2/5) = 0.970950594$$
$$Info_{ave}(A_1) = -5/14\times Info(A_1,a_{11}) - 4/14\times Info(A_1,a_{12}) - 5/14\times Info(A_1,a_{13}) = 0.693536139$$
$$Info_0(A_1) = -9/14\times\log_2(9/14) - 5/14\times\log_2(5/14) = 0.940285958$$
$$gain(A_1) = Info0(A_1) - Info(A_1) = 0.246724568$$

以下temperature, humidity, windyに関して同様な計算を繰り返すと以下のようになる。

$$gain(A_1) = Info_0(A_1) - Info_{ave}(A_1) = 0.246724568$$
$$gain(A_2) = Info_0(A_2) - Info_{ave}(A_2) = 0.029223$$
$$gain(A_3) = Info_0(A_3) - Info_{ave}(A_3) = 0.151836$$
$$gain(A_4) = Info_0(A_4) - Info_{ave}(A_4) = 0.048127$$

Step4. 葉の決定

(10),(11)よりi_{max}=1となり(12),(13),(14)からj_0=3となる。

Step5. 全てのノードの接続先がクラスの値になるまでStep1からStep4. を繰り返す。

j_0以外の属性値は次のノードとなるためsunnyに対応したデータ集合をD',rainyに対応したデータ集合をD''とする。表3はD',D''の各属性の相互情報量を示しており、D',D''の相互情報量が最大となる属性はhumidity, windyになるということがわかる。また、humidity, windyの各要素のとる平均情報量全てが0となる。

全ての分割したテストにおいて、相互情報量を最大とする属性値はそれぞれ0になるため全ての属性値のノードはクラスに接続され木の生成は終了する。データ集合Dからえられる決定木は図3のようになる。

表3. テストD', D''の相互情報量

	D'	D''
temperature	0.570950594	0.019973094
humidity	0.9709594	0.052663566
windy	0.321928095	0.9709594

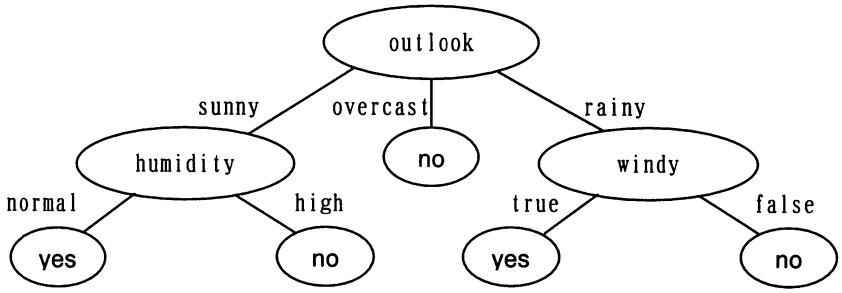

図3. Dから得られた決定木のイメージ

4-5 アルゴリズムの追加

3-4で示した利得獲得方法は属性値の数に依存しており、知識を獲得するときに必然的に属性値の数の多い属性が選ばれ易くなり、複雑な知識を得やすくなる。これに対して、利得比基準を用いて対処する。これは単純で属性自身の平均情報量を導出し、相互情報量をそれで除算を行う。ただし、属性値数の少ない属性にその除算を行わなくても良いので、適用属性は各属性の相互情報量の平均以上の属性に対して行なう。(4)(5)(6)のkをjに変えたものを(15)(16)(17)とし、各属性の相互情報量の平均を$gain_{ave}$とする。

$$S(i,j)=\sum_{j=1}^{n} sac(i,j,j) \quad \cdots \quad (15)$$

$$p(i,j,j)=\frac{sac(i,j,j)}{S(i,j)} \quad \cdots \quad (16)$$

$$Info'(i,j)=\sum_{j=1}^{n} P(i,j,j)\log_2 P(i,j,j) \quad \cdots \quad (17)$$

$$gain'(A_i)=\frac{gain(A_i)}{Info'(i,j)} \text{ (ただし、 } gain(A_i)\geqq gain_{ave}\text{)} \quad \cdots \quad (18)$$

今回扱うMushroom問題では属性値の不明という形で欠損したデータが存在しており、この対処が必要になる。ここでは不明な属性値に対処法は(4)から(14)に``不明な属性値に関する処理を行なわない''という条件を付け加えればよい。ただし(15)から(16)に関しては不明な属性値を含めた計算をしなければならない。また、(9)式を次のように変更する。Sumは不明属性を考慮しないデータの総数、$|D|$はDのデータ数を表す。

$$gain(A_i)=\frac{Sum}{|D|}(Info_0(A_i)-Info_{ave}(A_i)) \quad \cdots \quad (9)'$$

5. 対象問題

本研究では対象問題としてMushroomを使用する。これはUCIのMachine Learning Repositoryとして公開されている。Mushroom Databaseで原データはThe Aualaubon Society Field Guide to North American Mushroom(1981)から引用したものである。Mushroomの属性に関するデータが8124個あり、その形状や匂い、色などに関する22個の属性からなり、全データの51.8%の4208個が食用(edible)Mushroomであり、それ以外の3916個(48.2)が非食用(posoness)Mushroomの2クラスからなっている。この問題の属性情報を表5に表す。

本研究では、Mushroomデータから知識として求められた決定木の信頼性を10-fold cross varidationという手法で検討する。

表5.Mushroomのクラスと属性、属性値

Class
edible=e, poisonous=p
Attribute
1. cap-shape:
bell=b,conical=c,convex=x,flat=f,knobbed=k,sunken=s
2. cap-surface:
fibrous=f,grooves=g,scaly=y,smooth=s
3. cap-color:
brown=n,buff=b,cinnamon=c,gray=g,green=r,pink=p,purple=u,red=e,white=w,yellow=y
4. bruises?:
bruises=t,no=f
5. odor:
almond=a,anise=l,creosote=c,fishy=y,foul=f,musty=m,none=n,pungent=p,spicy=s
6. gill-attachment:
attached=a,descending=d,free=f,notched=n
7. gill-spacing:
close=c,crowded=w,distant=d
8. gill-size:
broad=b,narrow=n
9.gill-color:
black=k,brown=n,buff=b,chocolate=h,gray=g,green=r,orange=o,pink=p,purple=u,red=e,white=w,yellow=y
10. stalk-shape:
enlarging=e,tapering=t
11. stalk-root:
bulbous=b,club=c,cup=u,equal=e,rhizomorphs=z,rooted=r,missing=?
12. stalk-surface-above-ring:
ibrous=f,scaly=y,silky=k,smooth=s
13. stalk-surface-below-ring:
ibrous=f,scaly=y,silky=k,smooth=s
14.stalk-color-above-ring:
brown=n,buff=b,cinnamon=c,gray=g,orange=o,pink=p,red=e,white=w,yellow=y
15.stalk-color-below-ring:
brown=n,buff=b,cinnamon=c,gray=g,orange=o,pink=p,red=e,white=w,yellow=y
16. veil-type:
partial=p,universal=u
17. veil-color:
brown=n,orange=o,white=w,yellow=y
18. ring-number:
none=n,one=o,two=t
19.ring-type:
cobwebby=c,evanescent=e,flaring=f,large=l,none=n,pendant=p,sheathing=s,zone=z
20.spore-print-color:
black=k,brown=n,buff=b,chocolate=h,green=r,orange=o,purple=u,white=w,yellow=y
21.population:
abundant=a,clustered=c,numerous=n,scattered=s,several=v,solitary=y
22.habitat:
grasses=g,leaves=l,meadows=m,paths=p,urban=u,waste=w,woods=d

6. 結果と考察

図4の決定木は今回のdata miningの結果として求められたものであり、生のデータ8124個の属性値の関連性を示したものである。決定木はノード(node)と枝(edge)で構成された階層化されたグラフで最上階のノードは根(root)を示し、各階層におけるノードで枝を持つものを幹といい、枝を持たないノードを葉(leaf)という。幹は属性(attribute)が対応し、幹から派生した枝はその幹の属性における属性値が対応する。葉は枝を持たない結論としてのクラスの値(この場合はedibleかpoisonous)を示す。図4では、幹を楕円で示し、葉を円で示している。この決定木の根はodorであり、第2層の幹はspore-print-color、第3層の幹はgill-size、第4層の幹はgill-spacing、そして第5層の幹がpopulationとなっている。この決定木そのものも1つの知識表現となっているが枝や葉の属性値を記号表現している関係上、具体的な意味がわかりにくいのでこれらの決定木に対応する*if*～*then*ルールを表7に示す。

この結果、本来は8124個のexample dataが存在する場合には同数のルールが生成できることを考えれば data miningにより19個のルールに集約されたことになり、すばらしい知的処理といえる。しかしながら、このように求められた知識の質、すなわち知識の信頼性が生のデータ群の持つ情報を十分に吸収しているかどうかが問題である。そこで、求められた知識の質を検討するために10-fold varidation手法を用いてテストした結果、8124個の全てのデータに対して正解率が100%となった。

このことから今回求められた19個のルールは100%の信頼度を持つ知識となっていることが立証できた。さらに、data miningの効率化の可能性を検討するために、全データから100個、1000個のexampleをランダムに選択し、それらのデータ群から決定木を作成したものでそれらの知識の信頼度を求めた結果、表6のような結果が求められた。これによると、1000個のexampleから得られた知識でもかなり高精度(99.803%)のものが得られていることが判明した。したがって、使用目的によってはランダムに約1/80のexampleを全データより抽出して知識を生成しても約99%の信頼度をもつ知識となっていることから、全処理時間を短縮したものが期待できることがわかった。

以上の視点より、今回の分割統治法に基づいたdata miningアプローチでかなり膨大なデータ群からでも結構有益な知識獲得ができ、この手法が現実的処理方法として有効であることが確認できた。

表6.計測結果

データ数	100	1000	8124
全データの正解率(%)	98.966	99.803	
ルール数	10	20	19
計測時間(sec)	0	0.015	0.141

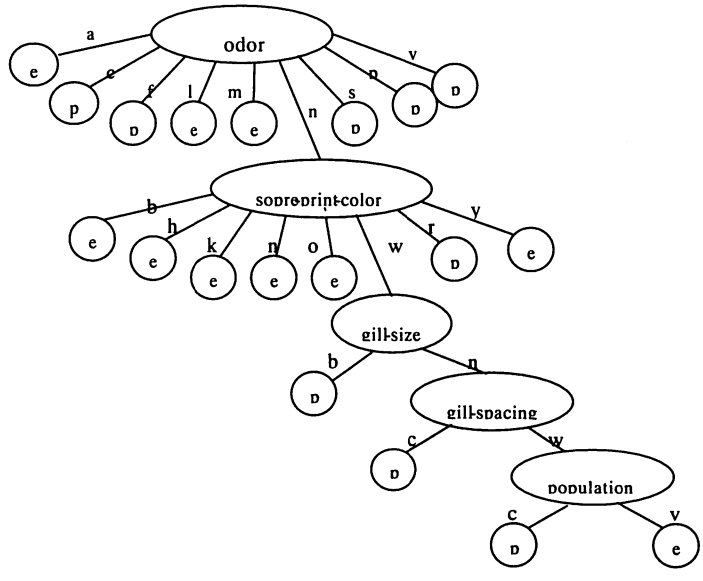

図4.Mushroomデータから得られた決定木

表7. Mushroomデータから分割統治法によって獲得した知識の詳細

rule1 : *if* (odor = almond) *then* edible
rule2 : *if* (odor = creoste) *then* poisonous
rule3 : *if* (odor = anise) *then* edible
rule4 : *if* (odor = foul) *then* poisonous
rule5 : *if* (odor = pungen) *then* poisonous
rule6 : *if* (odor = spicy) *then* poisonous
rule7 : *if* (odor = fishy) *then* poisonous
rule8 : *if* (odor = musty) *then* poisonous
rule9 : *if* (odor = none ∧ spore-print-color = black) *then* edible
rule10: *if* (odor = none ∧ spore-print-color = buff) *then* edible
rule11: *if* (odor = none ∧ spore-print-color = brown) *then* edible
rule12: *if* (odor = none ∧ spore-print-color = chocolate) *then* edible
rule13: *if* (odor = none ∧ spore-print-color = orange) *then* edible
rule14: *if* (odor = none ∧ spore-print-color = green) *then* poisonous
rule15: *if* (odor = none ∧ spore-print-color = yellow) *then* edible
rule16: *if*(odor = none ∧ spore-print-color = white ∧ gill-size = broad) *then* edible
rule17: *if*(odor = none ∧ spore-print-color = white ∧ gill-size = narrow ∧ gill-spacing = close) *then* poisonous
rule18: *if*(odor = none ∧ spore-print-color = white ∧ gill-size = narrow ∧ gill-spacing = close
∧ population = clustered) *then* poisonous
rule19: *if*(odor = none ∧ spore-print-color = white ∧ gill-size = narrow ∧ gill-spacing = close
∧ population = several) *then* edible

参考文献
[1]Ian H.Witten and Eible Frank : "Data Mining" Morgan Kaufmann Publishers(1999)
[2]Ryszad S.Michalski, Ivan Bratko, and Miroslav Kubat : "Machine Learning and Data Mining", John Wiley & LTD(1998)

Data Mining as Machine Learning

Michiaki Tsuda and Hiroyuki Narihisa*

Graduate School of Engineering
**Department of Information and Computer Engineering,*
Faculty of Engineering,
Okayama University of Science
Ridai-cho 1-1, Okayama 700-0005, Japan

(Received November 7, 2003)

Machine Learning was considered to be the computational methods that would implment various forms of learning, in particular machanisms capable of inducing knowledge from examples on data. Data mining is the extraction of implicit, previously unknown, and potentially useful informati on from raw data. Machine learning provides the technical basis of data mining.

In this paper, we present a basic concept of data mining as machine learning and show the effectiveness of knowledge discovery by adopting it to Mushroom problem that is the well-known benchmark problems each of which has 22 attributes concerning with its color, size, odor, habitant and etc. The decision tree obtained in our experiment is considerably compact and summarized information such as contains only 19 rules which represents the knowledge covering mushroom characteristics.