論 文 Technical Papers

部分空間クラスタリングと相関規則に基づく分類学習手法

A Classification Method Based on Subspace Clustering and Association Rules

中西 耕太郎 Kotaro nakanishi
大阪大学産業科学研究所高次推論方式
Institute for Scientific and Industrial Research, Osaka University
nakanishi@ar.sanken.osaka-u.ac.jp, http://www.ar.sanken.osaka-u.ac.jp/memberjp.html

鷲尾 隆 Takashi Washio
（同 上）
washio@ar.sanken.osaka-u.ac.jp, http://www.ar.sanken.osaka-u.ac.jp/~motoda/motoprjp.html

光永 悠紀*1 Yuki Mitsunaga
（同 上）
mitsunaga4@ar.sanken.osaka-u.ac.jp, http://www.ar.sanken.osaka-u.ac.jp/memberjp.html

藤本 敦*2 Atsushi Fujimoto
（同 上）
fujimoto4@ar.sanken.osaka-u.ac.jp, http://www.ar.sanken.osaka-u.ac.jp/memberjp.html

元田 浩*3 Hiroshi Motoda
（同 上）
motoda@ar.sanken.osaka-u.ac.jp, http://www.ar.sanken.osaka-u.ac.jp/~motoda/motoprjp.html

keywords: Subspace Clustering, Classification, Quantitative Frequent Itemset, Class Association Rule

Summary

Class Association Rule (CAR) based classification is known to provide high interpretability and accuracy in recent datamining study. However, most of the approaches have not addressed the issue to appropriately process numeric attributes of instances. Because the association among numeric instances is represented by a cluster in an attribute subspace, the approach named "*subspace clustering*" is expected to provide appropriate discretizations of the numeric attributes for CAR derivation. In this paper, a levelwise subspace clustering deriving interpretative hyper-rectangular clusters and a derivation scheme of quantitative and accurate CARs are proposed for CAR based classification. Significant performance of the proposed approach has been demonstrated through the tests on UCI repository data.

1. はじめに

データマイニングにおいては，事例クラスを予測するための分かりやすい知識を発見するために，人間にとって理解容易な分類規則を与える C4.5[Quinlan 93] に代表される決定木や C4.5Rules[Quinlan 87] に代表されるルールベース分類器が用いられることが多い．一方で近年，“クラスアソシエーションルール (CAR)” と呼ばれる分類規則を用いる学習分類器に関する研究が盛んになっている [Liu 98, Dong 99, Li 01]．多くの場合，これらの学習分類器は，従来の C4.5 のような決定木や決定規則によるよりもより良い分類精度を示す．特に対象事例に関係する全ての CAR の分類能力を総合的に用いる CAEP は，多くの場合に良い性能を示すことが知られている．CAR を分類に用いる多くの手法では，各数値属性について属性軸に射影した事例分布を基にエントロピーを計算して離散化を行う．しかしながら，この方法は図 1 (a) に示すように，複数属性に関するデータ分布の依存性を考慮しないため，同じクラスに対応する事例クラスタがしばしば離散化によって不適切に区分されてしまうことがある．C4.5 のような階層的離散化法によってこの問題を軽減することはできるが，なおも図 1 (b) に示すように事例クラスタを不適切区分してしまうことがあり得る．

この問題を効果的に解決する方法は，各数値属性部分空間でクラスタリングを行い分類能力の高い規則の条件部を見つけることである．このようなクラスタの存在する属性部分空間の探索及び探索された部分空間でクラスタリングを行う手法を部分空間クラスタリングといい，CLIQUE, DOC などが代表的手法である [Agrawal 98, Procopiuc 02]．これらは各属性部分空間において属性軸に平行な格子や窓の中の事例数を数えて密度の高いクラスタを見つけるが，格子や窓の向きや形状，大きさが不適切でク

†1 現在，日本電気株式会社所属
†2 現在, 富士通テン株式会社所属
†3 現在，The Asian Office of Aerospace Research and Development (AOARD) 所属.

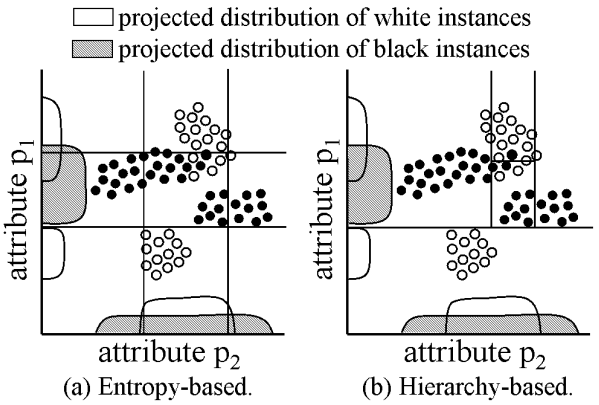


図 1 不適切な事例クラスタ区分

ラスタを見逃してしまうことがある．一方，近年開発された SUBCLU は，属性空間全体のクラスタリングを行う DBSCAN で提案された厳密な密度基準に基づいて密度ベースの部分空間クラスタリングを行う [Kailing 04, Ester 96]．密度基準に基づくクラスタでは，ある属性空間に存在するクラスタ内の事例は常にその部分空間内の何れかのクラスタにも含まれる．この性質を単調性と呼び，単調性と Apriori アルゴリズムを組み合わせることで，現実的時間内で効率的に各部分空間のクラスタを全探索可能である．しかしながら，これらの手法が導くクラスタの形状は軸平行な超直方体に限定されないため，結果の可読性に優れない．この問題に対し，定量的アソシエーションルールを発掘する手法の開発が行われてきた [Srikant 96, Wang 98]．しかし，これらの手法でもエントロピーを用いる離散化手法と同様に属性軸分布間の依存性を考慮せずに最良優先探索による離散化を行うため，図 1 (a) で説明したような不適切な離散化が起こりえ得る．CLTree も部分空間内の軸平行かつ超直方体形状のクラスタリングを行う手法であるが，階層的かつグリーディーにエントロピーを用いて事例分布の高密度部分を探索するため，C4.5 と同様に図 1 (b) に示すような不適切な離散化が起こり得る [Liu 00]．これら一連の手法のもう 1 つの問題は，数値属性からなる事例のクラスタリングにしか適用できないことである．学習分類器のための CAR を得るには，数値及びカテゴリカル属性空間でのクラスタリングを実行する必要がある．

本論文では，CAR を用いる学習分類器を構築するために全ての属性部分空間内の属性軸平行，超直方体形状のクラスタを効率的に完全探索する新たな手法を提案する．そして，このクラスタリング手法と前述の CAEP を組み合わせた学習分類器の性能評価を行う．

提案するクラスタリングアルゴリズムは幅優先探索であり，一次元属性部分空間のクラスタリングから始めて順次より高次元の属性部分空間クラスタリングへと拡張して行く．これは SUBCLU と似ているが，我々が提案する手法は標準的な Apriori アルゴリズムに幅優先探索のクラスタリングを埋め込むことで，数値アイテムと記号アイテムの両方が混在しているクラスタを導出可能である．更に，SUBCLU は 1 点のスカラー値からなる数値属性ベクトル事例のクラスタリングに限定されるのに対し，我々の手法では数値区間値をもつ数値アイテムを含む事例のクラスタリングも可能である．それらのクラスタは，高次元においても，事例の分布を確実に捉え，大きな分類能力を持つ．それに加え，提案手法により導出されるクラスタは超直方体として出力されるため人間にとって理解容易であり，データマイニング手法の適用可能性を大いに高める．本節で述べた提案手法の長所をまとめると次の様になる．

- 数値属性と記号属性が混在したクラスタを分類に使用できること．
- 数値区間値を持つ数値アイテムを含む事例に対して適用可能であること．
- 高次元においても，事例の分布を確実に捉え，大きな分類力を持つクラスタを分類に使用できること．
- 出力されたクラスタは超直方体の形状で出力され，人間にとって理解容易であること．

当稿では，はじめに次の節で CAEP を簡単に説明する．3 節ではクラスタリングの原理とそのアルゴリズムを提案する．そして 4 節で提案手法に関して性能評価を行い，5 節で QFIMiner が導出するクラスタの妥当性を論じる．

2. CAEP

2·1 CAR

CAR は “$\{< p_1 : q_1 >, ..., < p_m : q_m >\} \Rightarrow cl$” で，表される形式を有する．ここで，$< p : q >$ は “アイテム”，p は属性，q は属性値，cl はクラスを表す．アイテムの中で，数値区間値を持つアイテムを “数値アイテム”，カテゴリー値を持つアイテムを “記号アイテム” という．例えば “$\{< Age : [30, 39] >, < Married : Yes >, < NumCars : [2, 2] >\} \Rightarrow Houseowner$” は，“３０代の既婚者で自動車を２台所有する人は持ち家である” ことを表す．事例 t に含まれる数値アイテム $< p_t : q_t >$ が数値アイテム $< p : q >$ について $p_t = p$ かつ $q_t \subseteq q$ である場合，あるいは記号アイテム $< p_t : q_t >$ が記号アイテム $< p : q >$ について $p_t = p$ かつ $q_t = q$ である場合，$< p_t : q_t >$ は $< p : q >$ を “支持” するという．なお，$\subseteq$ は数値区間 q_t が数値区間 q の範囲内にあることを表す．CAR の条件部のすべてのアイテムが t に含まれる何れかのアイテムに支持される時，その事例 t のクラスは cl と予測される．従って，“$t_1 = \{< Age : [35, 37] >, < Married : Yes >, < NumCars : [2, 2] >, < Child : [3, 3] >\}$” は，上記のルール条件部を支持するので “持ち家である” と予測できる．一方，“$t_2 = \{< Age : [29, 31] >, < Married : Yes >, < NumCars : [2, 2] >, < Child : [3, 3] >\}$” は，$< Age : [29, 31] >$ が $< Age : [30, 39] >$ の範囲内にないため条件部を支持しない．今，属性とクラスからなる表形式デー

タあるいはクラスラベル付けされたトランザクションからなるデータを学習データ D，その中であるクラス cl を有するデータを D_{cl} とする．D_{cl} において “最小支持度 (*minimum support*: *minsup*)” 以上多頻度に表れるアイテム集合を “多頻度アイテム集合 (*frequent itemset*: FI)” とし，かつその中で数値アイテムを含むものを “定量的多頻度アイテム集合 (*quantitative frequent itemset*: QFI)” とする．ある CAR が適切なクラス分類を行うには，その CAR の結論部のクラスを有する事例の多くが同じく条件部を支持しなければならない．従って，本論文では CAR の条件部はその結論部のクラスに関するデータ D_{cl} において，多頻度アイテム集合 FI ないし定量的多頻度アイテム集合 QFI であるとする．

CAR を用いる学習分類器に関する最初の研究は CBA である．そこでは，各数値属性に関する値をエントロピーにより数値区間に離散化 [Fayyad 93] してカテゴリー属性として扱う方法を取り，更に各 D_{cl} に標準的な Apriori アルゴリズム [Agrawal 94] を適用して FI を求め CAR を導出する．あるトランザクション事例 t のクラスは，それによって条件部が高い “支持度 (*support*)” で支持され，かつ高い “確信度 (*confidence*)” で結論部のクラスを導く CAR から決められる．これに引き続いて提案された CMAR や CAEP という学習分類器は，1 つの事例の分類に複数の CAR を適用することで分類精度の向上を図っている．CMAR は導出された CAR を条件部の重なりの大きいもの同士グループ化し，事例 t による支持が最も大きいグループから得られるクラスに t を分類する．CAEP はあるクラス cl に関するデータ D_{cl} による支持度が，他のクラスのデータからの支持度より大きな条件部を持つ CAR を生成する．そして，事例 t はそれによって支持される条件を有する複数の CAR の重み付き投票によってクラス分類される．

2·2 CAEP

CAEP の学習過程は 2 つの段階からなる．はじめは全ての CAR 候補の条件部を導く段階である．先ず，以下の定義に従い，各クラス cl に関する定量的多頻度アイテム集合の集合 $LQFI(cl)$ を導出する．

【定義 1】（Support） アイテム集合 a のデータ D_{cl} による “支持度 (support)” を以下に定義する．

$$support_{D_{cl}}(a) = \frac{|\{t \in D_{cl} | a \subseteq t\}|}{|D_{cl}|}$$

□

【定義 2】（LQFI） あるクラス cl について，$support_{D_{cl}}(a) \geq minsup$ であるアイテム集合 a の集まりをクラス cl に関する定量的多頻度アイテム集合の集合 $LQFI(cl)$ とする． □

また，オリジナルの CAEP では，前述の CBA と同様にして定量的多頻度アイテム集合を求めるが，この論文では後述するように部分空間クラスタリングを用いる．

【定義 3】（Growth Rate） 各 $a \in LQFI(cl)$ について，以下の “*growth rate*” を計算する．ここで，$\bar{D}_{cl} = D - D_{cl}$ をクラス cl 以外の事例の集合とする．

If $support_{\bar{D}_{cl}}(a) \neq 0$,
$growth\_rate_{\bar{D}_{cl} \to D_{cl}}(a) = \frac{support_{D_{cl}}(a)}{support_{\bar{D}_{cl}}(a)}$,
if $support_{\bar{D}_{cl}}(a) = 0$ and $support_{D_{cl}}(a) \neq 0$,
$growth\_rate_{\bar{D}_{cl} \to D_{cl}}(a) = \infty$,
otherwise $growth\_rate_{\bar{D}_{cl} \to D_{cl}}(a) = 0$. □

$LQFI(cl)$ に含まれる定量的多頻度アイテム集合 a の *Growth Rate* が “*Growth Rate* の閾値” $\rho(> 1)$ より大きい，即ち $growth\_rate_{\bar{D}_{cl} \to D_{cl}}(a) \geq \rho$ であれば，a が条件部で結論部がクラス cl である CAR $a \Rightarrow cl$ が得られる．このようにして得られた規則は，クラス cl を持つ事例を他の事例からうまく区別することができる条件部を有している．これに対して，CBA や CMAR のような確信度に基づく規則生成方法では，規則がデータ D_{cl} について例え高い確信度を示しても，それ以外のデータ $\bar{D}_{cl}$ にも規則の条件部が当てはまってしまう可能性があるため，的確な分類が行えない可能性がある．

2 番目の段階では，CAR による重み付き投票に使用する index を導出する．まず，規則の条件部 a の分類能力を $support_{D_{cl}}(a)/(support_{D_{cl}}(a) + support_{\bar{D}_{cl}}(a)) = growth\_rate_{\bar{D}_{cl} \to D_{cl}}(a)/\ (growth\_rate_{\bar{D}_{cl} \to D_{cl}}(a) + 1)$ とする．これは規則の分類能力が $support_{D_{cl}}(a)$ と $support_{\bar{D}_{cl}}(a)$ の間の相対的差異によって決まると考えられるためである．これに従い，$LRB(cl)$ に含まれる全ての規則条件部によって事例 t がクラス cl に分類される可能性を表す “*Score*” を定義する．

【定義 4】（Score） $LRB(cl)$ を *Growth Rate* により $LQFI(cl)$ から選んだ全ての規則条件部の集合とする．その時，$LRB(cl)$ に含まれる全ての規則条件部によって事例 t がクラス cl に分類される可能性を “*Score*” とする．

$$score(t, cl) = \sum_{a \subseteq t, a \in LRB(cl)} \frac{growth\_rate(a)}{growth\_rate(a) + 1} * support_{D_{cl}}(a). \quad (1)$$

□

$LRB(cl)$ に含まれる規則条件部の数はクラスによって不均等なため，事例は多くの場合に条件部数の多い特定のクラスについて高い score を持ってしまう．このようなバイアスを打ち消すために，各クラス cl を重み付けする *base score* を導入する．

【定義 5】（Base Score） あるクラス cl に属する全ての事例 t を，cl に対する $\{score(t, cl) | t \in D_{cl}\}$ の値で昇順にソートした順列において，$Tail\%$ 番目の事例の score の値を $base\_score(cl)$ とする． □

尚，$Tail\%$ は人間が与えるパラメタである．ここまでで CAR の学習は終了する．

以上のようにして得られた分類器の性能をテストする際には，トレーニング過程で得られた $LRB(cl)$ に含まれる全ての条件部 a と全てのクラス cl に関する $base\_score(cl)$, $growth\_rate(a)$，$support_{D_{cl}}(a)$ の結果を用い，与えられたテスト事例 t とクラス cl に関する $score(t,cl)$ を式 (1) によって求める．そして，前述のバイアスを打ち消すために $base\_score(cl)$ によって以下のように規格化する．

【定義 6】（Normalized Score） 各クラス間における CAR の総数の差異から生じる score のバイアスを，base_score により打ち消した score を normalized_score とする．

$$norm\_score(t,cl) = \frac{score(t,cl)}{base\_score(cl)}.$$

□

定義 6 より，事例 t を最大の normalized_score を持つクラス cl に分類する．各 cl に関する $LQFI(cl)$ を導出する処理を除けば，CAEP の計算複雑性はデータを 1 回スキャンするだけなので，事例数 $N = |D|$ について $O(N)$ である．

3. CAR 条件部マイニングの提案手法

3·1 レベル幅探索による部分空間クラスタリング

はじめに数値アイテムのみからなる事例のクラスタリングに絞って説明する．分類能力の高い CAR の条件部は，各数値アイテムの属性から構成される部分空間において，あるクラスを持つ事例を多く含み他のクラス事例をあまり含まない，軸平行で超直方体の形状を持つ高密度なクラスタである．これに対し前述の SUBCLU の高密度クラスタの定義は，各事例について半径 ϵ 以内に少なくとも閾値 $MinPts$ 個以上の事例が存在する領域である [Kailing 04]．しかしこの定義では，クラスタは事例の分布に沿った高密度な多様な領域形状を有するため，それを包含する超直方体を CAR の条件部とする必要がある．更に，各事例ペア同士の距離を計算する必要があるため，本質的に計算量は $O(N^2)$ である．これに対して我々の提案手法では，部分空間内の事例密度を以下に示すように各属性軸に投影した分布上で調べ，直接に超直方体の形状を持つ高密度クラスタを導く．そのため，後に議論するように計算複雑性はほぼ $O(N \log N)$ となり，実用的なオーダーに軽減される

【定義 7】（Neighborhood） p は数値属性とし，かつ t と t' は各々区間値 q と q' を持つ p を共有する 2 つの事例とする．これら 2 事例間の p 上の距離 $Dist_p(q,q')$ を各区間 q 及び q' の最小距離，即ち $\min_{v\in q, v'\in q'} |v - v'|$ とする．更に Δ_p を属性 p 上の “許容距離 (*permissible range*)” とする．そして p 上の “Δ_p-近傍 (Δ_p-*neighborhood*)” $N_{\Delta_p}(t)$ を以下のように定義する．

$$N_{\Delta_p}(t) = \{t' \in D_{cl} | Dist_p(q,q') \le \Delta_p\}.$$

□

もし区間 q と q' が重なるならば $Dist_p(q,q') = 0$ であり，そうでないならば $Dist_p(q,q')$ は両区間が互いに向き合う境界間の距離となる．

【定義 8】（Core Instance） ある事例 $t \in D_{cl}$ について，もし Δ_p-近傍 $N_{\Delta_p}(t)$ が少なくとも “最小事例数 (*minimum points*)” $MinPts$ 個の事例を含む，即ち，

$$|N_{\Delta_p}(t)| \ge MinPts.$$

ならば，t は p 上の “*core instance*” と呼ばれる． □

【定義 9】（Direct Density-Reachability） ある 2 つの $t, t' \in D_{cl}$ について，t が p 上の core instance でかつ t' が $N_{\Delta_p}(t)$ に属するならば，t' は t から p 上で “*directly density-reachable*” であるという． □

【定義 10】（Density-Reachability） D_{cl} 内に p 上で各 t_{i+1} が各 t_i から directly density-reachable である事例の鎖 $t_1 = t, t_2, \cdots, t_{n-1}, t_n = t'$ が存在するとき，t' は t から p 上で “*density-reachable*” であるという． □

【定義 11】（Density-Connectivity） D_{cl} 内において p 上で t と t' の両方へ density-reachable である t'' が存在するとき，t と t' は “*density-connected*” であるという． □

【定義 12】（Density-Connected Set） 空でない事例集合 $C \subseteq D_{cl}$ 内の全事例が互いに p 上で density-connected であるとき，C を p 上の “*density-connected set*” という． □

【定義 13】（Dense Cluster） 数値属性の集合 S からなる部分空間内で，各数値属性 $p \in S$ 上で density-connected set である極大な事例集合 $C^S \subseteq D_{cl}$ を “*dense cluster*” という． □

【定義 14】（Quantitative Frequent Itemset） $C^S \subseteq D_{cl}$ を部分空間 S 内の dense cluster とする．そして，$\max_p(C^S)$ 及び $\min_p(C^S)$ を C^S 内の事例が p 上で取る最大及び最小の値とし，$a(C^S) = \{< p : q > | p \in S, q = [\min_p(C^S), \max_p(C^S)]\}$ とする．$a(C^S)$ は部分空間 S において C^S の最大・最小値で C^S を包含するアイテム集合である．ここでもし $|C^S| \ge minsup$，即ち $support_{D_{cl}}(a(C^S)) \ge minsup$ であれば，$a(C^S)$ を “定量的多頻度アイテム集合 (*quantitative frequent itemset*; QFI)” と呼ぶ．また S の次元を k とするとき，$a(C^S)$ を k-QFI と呼ぶ． □

QFI は部分空間において，極大な体積を有する高密度，軸平行な超直方体領域である．SUBCLU における dense cluster と同様に，QFI は以下の (anti-)monotonicity を有する．

【定理 1】（Monotonicity） S の全ての部分空間 $T \subseteq S$ について，もし $a(C^S)$ が S 内の QFI ならば，T 内に $a(C^S)$ で支持される QFI である $a(C^T)$ が存在する．即ち，T 内に $a(C^S) \subseteq a(C^T)$ である $a(C^T)$ が存在する．

□

《証明》 C^S 内の全ての事例は S の各属性軸 p 上で density-connected であるため，T の各属性軸 p 上でも

表 1 クラス $cl = Houseowner$ のトランザクションセット例; $D_{Houseowner}$

$t_1 = (\{< Age : [20,23] >, < Child : [3,3] >, < NumCars : [2,2] >\}, Houseowner)$
$t_2 = (\{< Age : [30,30] >, < Child : [4,5] >, < NumCars : [1,1] >, < Savings : [10K,10K] >\}, Houseowner)$
$t_3 = (\{< Age : [30,30] >, < Child : [2,2] >, < NumCars : [5,5] >, < Savings : [11K,11K] >\}, Houseowner)$
$t_4 = (\{< Age : [30,35] >, < Child : [5,5] >, < NumCars : [1,1] >\}, Houseowner)$
$t_5 = (\{< Age : [35,37] >, < Child : [2,2] >, < NumCars : [2,2] >, < Savings : [5K,5K] >\}, Houseowner)$
$t_6 = (\{< Age : [36,39] >, < Child : [2,3] >, < NumCars : [2,3] >\}, Houseowner)$

表 2 $D_{Houseowner}$ の幅優先探索部分空間クラスタリングの過程

1-QFIs
$(\{< Age : [30,39] >\}, \{t_2,t_3,t_4,t_5,t_6\})$,
$(\{< Child : [2,5] >\}, \{t_1,t_2,t_3,t_4,t_5,t_6\})$,
$(\{< NumCars : [1,3] >\}, \{t_1,t_2,t_4,t_5,t_6\})$,
$(\{< Savings : [10K,11K] >\}, \{t_2,t_3\})$
2-QFIs
$(\{< Age : [30,39] >, < Child : [2,2] >\}, \{t_3,t_5,t_6\})$
$(\{< Age : [30,35] >, < Child : [4,5] >\}, \{t_2,t_4\})$
$(\{< Age : [30,39] >, < NumCars : [1,3] >\}, \{t_2,t_4,t_5,t_6\})$
$(\{< Age : [30,30] >, < Savings : [10K,11K] >\}, \{t_2,t_3\})$
$(\{< Child : [2,5] >, < NumCars : [1,3] >\}, \{t_1,t_2,t_4,t_5,t_6\})$
3-QFIs
$(\{< Age : [35,39] >, < Child : [2,2] >, < NumCars : [2,3] >\}, \{t_5,t_6\})$
$(\{< Age : [30,35] >, < Child : [4,5] >, < NumCars : [1,1] >\}, \{t_2,t_4\})$

density-connectedである．従って，$C^S \subseteq C^T$ であり，T 内の各 p について $[\min_p(C^S), \max_p(C^S)] \subseteq [\min_p(C^T), \max_p(C^T)]$ が成り立つ．このことから $a(C^T)$ は $a(C^S)$ に支持され，かつこのような $a(C^T)$ が必ず存在する． ∎

このことから，QFI を全探索する際にレベル幅探索アルゴリズムを利用可能である．この探索具体例を表 1 のデータセットを使って説明する．各事例（トランザクション）t_i は 1 つの数値アイテム集合とそのクラス $cl = Houseowner$ からなる．ここではクラスタリングを $\Delta_{Age} = 5$, $\Delta_{Child} = 1$, $\Delta_{NumCars} = 1$, $\Delta_{Savings} = 1K$, $MinPts = 1$, $minsup = 2$ のパラメータの元で行うことにする．最初に各 t_i 内のアイテムを属性名で辞書順にソートするが，この表では既にソートされている．次に 1 つの属性について各事例の値でソートし，density-connected である極大な 1-QFI が探索される．属性 Age については，$\Delta_{Age} = 5$ の元で 30 から 39 に亘ってアイテムが密に分布し，かつその支持度が 5 で $minsup$ より大きいため，1-QFI $\{< Age : [30,39] >\}$ が存在する．1-QFI を表 2 にトランザクション ID リスト (TID-List) と共に示す．この例では各属性が 1 つずつ 1-QFI を有している．1-QFI の探索にはソートしか必要ないため，その計算複雑性は $O(NlogN)$ に留まる．

次のステップから k-QFIs ($k > 1$) の幅優先探索が開始される．標準的な AprioriTid アルゴリズム [Agrawal 94] 同様に，D_{cl} 内の各事例を表すインデックスリストである $TID - List$ が用いられる．今，全ての $(k-1)$-QFI が既知である時，以下の “候補生成 (*Candidate-Generation*)” 処理によって k-QFI の候補である candidate-k-QFI が全て導かれる．

候補生成

結合フェーズ (Join Phase): $k-2$ 個の属性を共有する以下の 2 つの $(k-1)$-QFI について，

$((k-1) - QFI = \{< p_1 : q_1 >, < p_2 : q_2 >, ..., < p_{k-2} : q_{k-2} >, < p_{k-1} : q_{k-1} >\}, TID - List)$,
$((k-1) - QFI' = \{< p_1 : q'_1 >, < p_2 : q'_2 >, ..., < p_{k-2} : q'_{k-2} >, < p_k : q'_k >\}, TID - List')$,
次のような結合を取る:
$(candidate - k - QFI = \{< p_1 : q^c_1 >, ..., < p_{k-1} : q^c_{k-1} >, < p_k : q^c_k >\}, TID - List^c)$.
ここで，q^c_i は $i = 1, ..., k-2$ について各 2 区間の交わり $q_i \cap q'_i$，並びに $q^c_{k-1} = q_{k-1}$，$q^c_k = q'_k$ であり，更に $TID - List^c = TID - List \cap TID - List'$ である．もし 1 つでも $q^c_i = \phi$ の場合には，これら 2 つの $(k-1)$-QFI は結合できない．

枝狩りフェーズ (Prune Phase): 導かれた candidate-k-QFI の大きさ $(k-1)$ の各部分集合 s について，

$$\forall < p_i : q^c_i > \in s,$$

$$\exists < p_i : q_i > \in (k-1) - QFI, q^c_i \cap q_i \neq \phi, \quad (2)$$

を満たす $(k-1)$-QFI が存在するならば，この candidate-k-QFI は候補に留まり，$TID - List^c$ は dense cluster 候補 $\hat{C}^S$ ($|S| = k$) となる．そうでなければ，candidate-k-QFI は除去される．

この枝狩りフェーズは定理 1 に基づいている．式 (2) において q^c_i が q_i と交わる限り，s と $(k-1) - QFI$ が $minsup$ 以上の事例を共有する可能性は排除できず，単調性を満たす可能性は失われない．従ってこの場合，candidate-k-QFI は残される．表 2 に示す 2 つの 1-QFI の $\{< Age : [30,39] >\}$ と $\{< Child : [2,5] >\}$ からは，$TID - List^c = \{t_2, t_3, t_4, t_5, t_6\}$ を持つ candidate-2-QFI$\{< Age : [30, 39] >, < Child : [2,5] >\}$ が導かれる．この候補は，全ての部分集合 $\{< Age : [30,39] >\}$, $\{< Child : [2,5] >\}$ が QFI なので，枝狩りフェーズを通過する．

以上で得られた各 candidate-k-QFI とその $TID - List^c$ について，図 2 に示されるアルゴリズム “*QFI-Count*” が，定義 13 及び 14 に基づいて，dense cluster $C^S = TID - List$ とその k-QFI を探索する．当然，候補に属する事例がこれらの定義に合致するクラスタを構成しなければ，$C^S = TID - List$ と k-QFI は 1 つも出力されない．図中のステップ (7) から (9) にかけての内部ループでは，指定された許容距離 Δ_p と最小事例数 $MinPts$ 及び定義 12 の下で関数 $MDCS$ において，はじめに dense cluster 候補 $\hat{C}^S = TID - List^c$ 中である属性軸 p 上で事例の極大な density-connected set C が探索される．$\hat{C}^S$ 内に複数の dense cluster が含まれていれば，複数の C が見つかることになる．$MDCS$ は，同ループの前回の反復において $TIDLS.temp$ から導かれた各 C の内部で，新たな p について C を求めることを繰り返す．最小支持度 $minsup$ 未満の事例数しか含まない C は，$MDCS$ 内で除去される．この更新操作は外側のステップ (5) から (10)

QFI-Count($candidate-k-QFI, TID-List^c$);
(1) $k-QFIS=\phi, TIDLS=\phi$;
(2) If $|TID-List^c| < minsup$ return $k-QFIS$;
(3) $S=\{p| < p:q >$
$\in candidate-k-QFI, p$ is numeric.$\}$;
(4) $TIDLS.temp=\{TID-List^c\}$;
(5) while $TIDLS \neq TIDLS.temp$ do begin
(6) $TIDLS=TIDLS.temp$;
(7) forall $p \in S$ do begin
(8) $TIDLS.temp=$
$MDCS(TIDLS.temp, p)$;
(9) end
(10) end
(11) forall $TID-List \in TIDLS$ do begin
(12) $k-QFIS=k-QFIS+$
$(QFI(S, TID-List), TID-List)$;
(13) end
(14) return $k-QFIS$;

図 2 QFI-Count のアルゴリズム

(1) For each numeric attribute, create an index list sorted with the ascending order of D. Sort items in each $t \in D$ lexicographically.
(2) $L_1=\{(1-QFI, TID-List)\}$;
(3) for (k=2; $L_{k-1} \neq \phi$; $k++$) do begin
(4) $C_k=$
$\{(candidate-k-QFI, TID-List^c)\}=$
$Extended-Candidate-$
$Generation(L_{k-1})$;
(5) forall $(candidate-k-QFI,$
$TID-List^c) \in C_k$ do begin
(6) $L_k=L_k \cup$
$QFI-Count(candidate-k-QFI,$
$TID-List^c)$
(7) end
(8) end
(9) Answer $L=\bigcup_k L_k$;

図 3 全体アルゴリズム

のループで，各 C が dense cluster $C^S=TID-List$ に収斂するまで続けられる．クラスタの単調性より，探索経路の依存しないで各 C^S を求めることができる．ステップ (11) から (13) のループでは，関数 QFI において定義 14 によって各 C^S に対応する QFI が計算される．

表 2 の例で，candidate-2-QFI $\{< Age:[30,39] >, < Child:[2,5] >\}$ と $TID-List^c=\{t_2,t_3,t_4,t_5,t_6\}$ が QFI-Count に与えられる場合を考えると，ステップ (7) から (9) の内部ループで $MDCS$ は，属性 Age について $\Delta_{Age}=5$ の下で極大な density-connected set として $TIDLS.temp=\{\{t_2,t_3,t_4,t_5,t_6\}\}$ を求める．次にこの $TIDLS.temp$ について，$\Delta_{Child}=1$ の下で属性 $Child$ 上で $TIDLS.temp=\{\{t_3,t_5,t_6\},\{t_2,t_4\}\}$ が求められる．そしてこれ以上 $MDCS$ を適用しても $TIDLS.temp$ は変わらず，かつこれらの支持度は $minsup=2$ 以上なので，これらの dense cluster に対応する 2 つの 2-QFI である $(\{< Age:[30,39] >, < Child:[2,2] >\}, \{t_3,t_5,t_6\})$ と $(\{< Age:[30,35] >, < Child:[4,5] >\}, \{t_2,t_4\})$ が導かれる．

3·2 数値と記号アイテムからなる QFI の導出

候補生成処理は数値と記号両方のアイテムからなる QFI を導けるように拡張可能である．結合されるアイテム集合中の記号アイテムの値は，標準的な AprioriTid アルゴリズムと同様に与えられる．

拡張候補生成

結合フェーズ (Join Phase): $k-2$ 個の属性を共有する以下の 2 つの $(k-1)$-QFI について，
$((k-1)-QFI=\{< p_1:q_1 >, < p_2:q_2 >, ...,$
$< p_{k-2}:q_{k-2} >, < p_{k-1}:q_{k-1} >\}, TID-List)$,
$((k-1)-QFI'=\{< p_1:q'_1 >, < p_2:q'_2 >, ...,$
$< p_{k-2}:q'_{k-2} >, < p_k:q'_k >\}, TID-List')$,
次のような結合を取る:
$(candidate-k-QFI=\{< p_1:q^c_1 >, ...,$
$< p_{k-1}:q^c_{k-1} >, < p_k:q^c_k >\}, TID-List^c)$.
ここで q^c_i は $i=1,...,k-2$ について，数値アイテムの場合には各 2 区間の交わり $q_i \cap q'_i$ であり，記号アイテムの場合は $q^c_i=q_i=q'_i$ である．また数値，記号何れのアイテムの場合も $q^c_{k-1}=q_{k-1}$，$q^c_k=q'_k$ であり，更に $TID-List^c=TID-List \cap TID-List'$ である．もし数値アイテムの 1 つでも $q^c_i=\phi$ または記号アイテムの 1 つでも $q_i \neq q'_i$ の場合には，これら 2 つの $(k-1)$-QFI は結合できない．

枝狩りフェーズ (Prune Phase): 導かれた candidate-k-QFI の大きさ $(k-1)$ の各部分集合 s について，
$\forall < p_i:q^c_i > \in s, \exists < p_i:q_i > \in (k-1)-QFI$,
全ての数値アイテムについて $q^c_i \cap q_i \neq \phi$
かつ全ての記号アイテムについて $q^c_i=q_i$,
を満たす $(k-1)$-QFI が存在するならば，この candidate-k-QFI は候補に留まり，$TID-List^c$ は数値アイテムで構成される部分空間内の dense cluster 候補 $\hat{C}^S$ $(|S|=k)$ となる．そうでなければ candidate-k-QFI は除去される．

図 2 の QFI-Count アルゴリズムも拡張が必要である．candidate-k-QFI が記号アイテムのみからなる場合には，ステップ (5) から (10) のループをスキップし $TIDLS=TIDLS.temp$ とする．また，ステップ (12) の関数 QFI において，記号アイテムについては値を $q^c_i=q_i=q'_i$ とする．

図 3 にデータ D から QFI を導くアルゴリズム全体を示す．必要な入力パラメータは各数値属性の許容距離 Δ_p と最小事例数 $MinPts$，最小支持度 $minsup$ である．はじめに拡張候補生成と QFI-Count で効率的に処理を行うために，幾つかの必要なインデックスを作成する．そして，AprioriTid アルゴリズムの結合処理を拡張候補生成処理に置き換えたアルゴリズムによって，全ての QFI を計算しリスト L に格納する．実装レベルでは $(candidate-k-QFI, TID-List^c)$ という対応インデックスではなく，標準の AprioriTid アルゴリズムと同様に各事例 t_i についてそれが含む candidate-k-QFI を示す逆引きインデックス $(t_i, \{candidate-k-QFI\})$ を用いる．これら一連の処理中でもっとも計算複雑性の高い処理は，事例数 N について $O(N \log N)$ である最初のステップのソート及び QFI-Count の dense cluster 導出処理である．$MDCS$ では，最初のステップで作成されるインデッ

クスリストを使い $TIDLS.temp$ を 1 回スキャンするだけで，各数値属性軸 p 上で極大な density-connected set を簡単に導出でき，この処理は $O(N)$ である．しかし，図 2 に示す QFI-Count のステップ (5) から (10) までの外周ループの反復回数は，属性部分空間内の事例分布に強く依存する．最悪は各ループ反復で 1 つの事例のみが C から除かれる場合であり，ループ全体の計算複雑性は $O(N^2)$ となる．しかし，各ループ反復で平均して C からそこに属する事例数の一定割合 $0 < r < 1$ が除かれる状況が，もっともあり得る場合である．この場合にはループ反復回数を m とした時，$r^m N$ が $minsup$ より小さくなった時にループが終了し，$minsup \simeq r^m N$ より m は $O(\log N)$ 程度となる．従って，アルゴリズム全体の期待時間計算複雑性は $O(N \log N)$ となる．

4. 性 能 評 価

4·1 計 算 複 雑 性

QFI 導出過程の計算時間及びメモリ消費量に関する効率性について，以下の手順によって作成した人工データを用いて評価する．まず，無作為に全体の r_n% が数値アイテムであり，残りのアイテムが記号アイテムである SSI(set of seed items) を作成する．次に，平均 $|\overline{QFI}|$ を持つ範囲 $[0.7|\overline{QFI}|, 1.3|\overline{QFI}|]$ の一様乱数分布から決められる個数の seed item を無作為に選択し QFI 候補を生成する．これを繰り返し多数の QFI 候補の集合 SQFI を作成する．更に，SQFI から無作為に幾つかの QFI 候補を抽出し，それぞれに SSI から無作為に抽出した $2|\overline{QFI}|$ 個の seed item を加えて事例 t を生成する．各 QFI 候補から最小支持度 ($minsup$) 以上多数回に亘り事例 t を生成し，事例集合 D を作成する．従って，各 QFI 候補は D において定量的多頻度集合 QFI となる．またこれにより，平均的な事例のサイズ $|\bar{t}|$ は $3|\overline{QFI}|$ となる．最後に，それぞれの事例の数値アイテムの値に対して，5% のガウス雑音を加えることによって，部分空間クラスタを形成する事例 t の分布にいくらかばらつきを与える．我々が提案した QFI を導出する部分空間クラスタリングアルゴリズムを用いて，CPU が Pentium4 2.7GHz，メモリが 2GB の計算機を用いて性能検証を行った．その際のデフォルトパラメータは人工データについて $|SSI| = 1000$, $r_n = 50\%$, $|SQFI| = 10$, $|\bar{t}| = 12$, $N = |D| = 40000$, $minsup = 5\%$, QFIMiner の解析条件について $MinPts = 1$, $\alpha_\Delta = 20\%$ とした．但し，α_Δ はデータ D 内のそれぞれの数値属性の最小値と最大値の間隔に対する Δ の相対的な幅を示す．

表 3 にこれらのパラメータを個別に変化させたときの計算時間とメモリ消費量の依存性に関する定性的傾向を示す．我々のアルゴリズムは $|SSI| = 20000$ といった高次元の事例に対しても 12 分で計算できた．このことは標準的な AprioriTid アルゴリズムに類似している．r_n に対する依存性が弱いことが示されているが，これは．数値

表 3 クラスタリングの計算複雑性

Para-meter	Range of Assessment	Dependency of Comp.Time	Dependency of Mem.Cons.
$\|SSI\|$	$[20, 20000]$	constant	constant
r_n	$[0\%, 100\%]$	constant	constant
$\|SQFI\|$	$[1, 50]$	$O(\|SQFI\|)$	$O(\|SQFI\|)$
$\|\bar{t}\|$	$[8, 100]$	exp. inc.	exp. inc.
$minsup$	$[0.2\%, 10\%]$	exp. dec.	exp. dec.
α_Δ	$[0.1\%, 100\%]$	inc. const.	inc. const.
$MinPts$	$[1, 8000]$	dec. const.	dec. const.
N	$[200, 10^6]$	$O(N \log N)$	$O(N)$

exp. inc./dec. : exponential increase/decrease.
inc./dec. const. : increase/decrease and saturation.

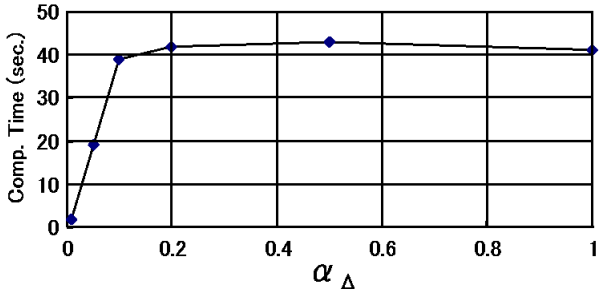


図 4 α_Δ と計算時間の関係

アイテムと記号アイテムに関して，基本的な手続きにおいて大きな差異はがないためと考えられる．また，$MinPts$, Δ_p, N を除き，その結果は従来の Apriori-Tid アルゴリズムに類似している．図 4 に示すように，Δ_p が事例間の平均間隔に相当するほど極めて小さい場合は，Δ_p に対してクラスタの数が非常に敏感であるため，計算時間とメモリ消費量に急激な増加が観測された．図 5 に示された N に対する計算時間の関係は，事例数 100 万のデータまでほぼ $O(NlogN)$ である．一方，3·1 節の最初で述べたように SUBCLU の計算時間は $O(N^2)$ となる．但し，SUBCLU は記号アイテムを扱うことができないため，ここでは計算時間の直接的な比較は行わない．各事例とそれが含む candidate-k-QFI を対応づける逆引きインデックスの大きさが N に比例するため，メモリ消費量は N に比例する．

4·2 分類手法の性能評価

我々が提案した部分空間クラスタリング手法によって得られた QFI を CAR の条件部として CAEP に導入した．この手法を省略して LSC-CAEP(Levelwise Subspace Clustering-CAEP) と呼ぶ．QFI はそれぞれのクラスを有するデータの属性部分空間中で導出されるため，QFI はクラスを強く特徴づける条件となっている．一方，従来の CAEP は，前処理において数値属性個別に値の離散化を行うため，クラスを特徴づけるには不適切な相関規則の条件部が多く生成される可能性が高い．

表 4 に示された UCI repository に公開されているデータを用いて C4.5，CBA，LSC-CAEP の性能比較実験を

表 4 各種データに対する精度比較

dataset	num. of records	num. of attributes(numeric)	num. of classes	C4.5	CBA	LSC-CAEP	SD of C4.5	SD of CBA	SD of LSC-CAEP
Australian	690	14(6)	2	.8608	.8538	**.8666**	.0362	.0422	.0347
Cars	392	7(6)	3	.9617	.9744	**1.0000**	.0347	.0318	0
Cleve	303	13(5)	2	.7656	.8283	**.8383**	.0899	.0539	.0422
Crx	690	15(6)	2	.8608	.8538	**.8715**	.0333	.0466	.0442
Diabetes	768	8(8)	2	.7226	**.7445**	.7229	.0567	.0410	.0681
Ecoli	336	8(7)	8	**.8422**	.7018	.7794	.0750	.2613	.0992
German	1000	20(7)	2	.7070	**.7350**	.7173	.0323	.0411	.0517
Heart	270	13(6)	2	.7666	8187	**.8222**	.0724	.0930	.0694
Hepatitis	155	19(6)	2	**.8387**	.8182	.8236	.0724	.0870	.1062
Horse	368	22(8)	2	.6933	.8236	**.8394**	.0409	.0458	.0488
Hypo	3163	25(7)	2	**.9889**	.9826	.9793	.0045	.0067	.0071
Iris	150	4(4)	3	.9600	.9467	**.9733**	.0562	.0611	.0466
Labor	57	16(8)	2	.7368	.8633	**.9500**	.2252	.1327	.1124
Led7	3200	7(0)	10	.7337	.7206	**.7400**	.0204	.0157	.0117
Lymph	148	18(2)	4	.7635	.8040	**.8157**	.1005		.1189
Nursery	12960	8(0)	5	**.9705**	.8289	.9408	.0046	.1107	.0048
Pima	768	8(8)	2	**.7382**	.7290	.7141	.0566	.0500	.0338
Sonar	208	60(60)	2	**.7884**	.7746	.6681	.0632	.1307	.1288
Tae	151	5(1)	3	**.5099**	.4717	.5067	.0858	.2315	.1470
Tic-Toc-Toe	958	9(0)	2	.8507	.9959	**1.0000**	.0449	.0072	0
waveform	5000	21(21)	3	.7664	**.7968**	.7886	.0158	.0152	.0153
Wine	178	13(13)	3	.9382	.9496	**.9833**	.0552	.0612	.0374
Zoo	101	16(0)	7	.9207	**.9709**	.9309	.0637	.0469	.0477
Average				.8123	.8264	**.8379**	.0583	.0733	.0555

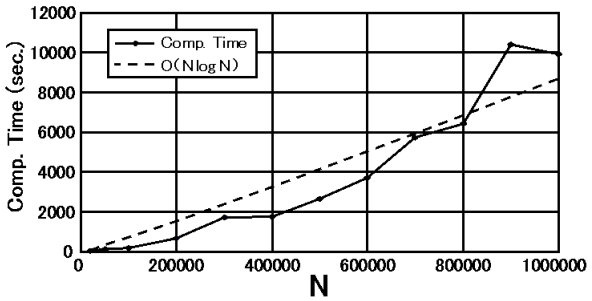


図 5 データ数 N に対する計算時間の依存性

行った．C4.5, CBA[*1]の分類精度は 10 fold cross validation で評価した．但し，C4.5 は weka に実装されている J48 を用いて評価した．尚，LSC-CAEP において，デフォルトパラメータが未定であるため，事前に各々のデータに対して 10 fold cross varidation を実行することにより最適なパラメータを学習し，そのパラメータでオリジナルのデータをシャッフルしたデータに対して 10 fold cross validation を実行して性能評価した．以下にパラメータの学習方法を記す．

LSC-CAEP で必要とされるパラメータは support の閾値, Growth Rate の閾値, α_Δ, Tail である．CAR の条件部を導出する過程で必要とされるパラメータは support の閾値と α_Δ であるが，4･1 節で述べたように CAR の導出過程には $O(NlogN)$ の計算時間がかかる．一方，CAEP で必要とされるパラメータは Growth Rate の閾値と Tail であるが，CAEP の計算時間は $O(N)$ であり，その計算時間は無視できるほど小さい．そこで，Growth Rate の閾値と Tail それぞれについて探索するパラメータのリストを作成し，support の閾値と α_Δ によって生成された条件部に対して，リストに記された Growth Rate と Tail の全ての値に対して分類精度を導出した．尚，support の閾値と α_Δ については，前述した通り計算時間の問題から多数のパラメータの組み合わせを評価することは困難である．そこで，パラメータの学習過程ではこれらのパラメータについては，最適である可能性が高いと思われる探索範囲に絞った．本稿では計算を繰り返し，最終的に最高分類精度を達成したパラメータの組み合わせを最適なパラメータとして学習した．

表 4 では太字で書かれた手法が最高の分類精度を達成したことを示している．表 4 の最下段に性能評価実験で使用した全てのデータにおける各分類手法の分類精度の平均を記した．また，表 4 の右端に提案手法の 10 fold cross validation における 10 回の試行における分類精度の標準偏差を記した．表 4 における各データに対する最高分類精度とその次に良い分類精度の差と誤差分類精度を比較すると car, nursery, tic-toc-toe 以外のデータでは誤差分類精度の方が大きい．故に，分類手法の絶対差から言えば，提案手法と表 4 で性能比較した手法間には個別ではその性能に顕著な差がないといえる．しかしながら，提案手法は表 4 に記された平均分類精度において CBA，C4.5 を上回っている．また，もう一つの評価基準として，各データにおいて最高分類精度を達成した手法に対して 2 点，二番目の分類精度を達成した手法に対して 1 点を与えるという点数評価を行うと，C4.5 が 19 点，CBA が 18 点，LSC-CAEP が 32 点となった．

∗1 CBA の公開実行形式プログラムのバグにより Lymph に対して計算できなかったため，CBA の Lymph に対する分類精度は [Liu 98] を参照した．尚，標準偏差は当該論文不掲載のため空欄とした．

更に，提案手法は表 4 に記された 23 件のデータの内，12 件のデータで最高分類精度を達成した．これに基づいて，提案手法の他手法に対する統計的優位性について考察する．各データ $i(i=1,...,N)$ において，LSC-CAEP の分類精度が他の 2 手法を上回る確率を p_i とする．各データ i における分類精度の比較は一回しか行わないため，各データ i において，LSC-CAEP が他の 2 手法を上回る確率分布は二項分布 $B(1,p_i)$ に従う．この二項分布において期待平均は p_i，期待分散は $p_i(1,p_i)$ となる．ここで，各 p_i は独立であるため，N 種類のデータのうち，LSC-CAEP が 3 種類の手法のうち最高の分類精度を達成する回数は，各データ i における二項分布 $B(1,p_i)$ を重ね合わせた混合二項分布 $\sum_{i=1}^{N} B(1,p_i)$ に従うといえる．しかしながら，p_i は未知なため，混合二項分布 $\sum_{i=1}^{N} B(1,p_i)$ を直接用いずに，二項分布 $B(N,p)$ において，2σ 検定により LSC-CAEP の分類性能における優位性を示す．何故ならば，混合二項分布 $\sum_{i=1}^{N} B(1,p_i)$ の標準偏差は，p_i が一定 p であるとき，即ち，混合二項分布 $\sum_{i=1}^{N} B(1,p_i)$ が二項分布 $B(N,p)$ となるとき最大となる性質を有するためである．このとき，二項分布 $B(N,p)$ において，2σ 検定により LSC-CAEP の優位性が証明できれば，より小さな標準偏差を持つ混合二項分布 $\sum_{i=1}^{N} B(1,p_i)$ において，LSC-CAEP は余裕を持って CBA，C4.5 と比較して優位であるといえる．

個々の手法が対等な精度であると仮定すると，二項分布 $B(N,p)$ において，LSC-CAEP が各データにおいて CBA，C4.5 を上回る分類精度を達成する確率は 1/3 となる．すると，LSC-CAEP が分類精度において C4.5 と CBA を上回るデータの個数 x の期待値は $23\times\frac{1}{3}=7.66$ となる．このとき，x の標準偏差は $\sqrt{23\times\frac{1}{3}(1-\frac{1}{3})}=2.26$ となる．故に，二項分布 $B(23,\frac{1}{3})$ において，平均値から 2σ の区間上限値は $7.66+2.26\times 2=12.18$ となる．それに対して，表 4 より，LSC-CAEP は 12 個のデータにおいて最高の分類精度を達成した．LSC-CAEP が 3 手法の中で最高の分類精度を達成する回数は整数値をとるため，この場合，2σ 検定により，個々の手法の分類性能は同等であるという仮説を棄却できる水準であるといえる．故に，LSC-CAEP の分類性能は CBA，C4.5 を上回っているといえる．実際のデータでは，p_i が一定の値をとることは無いが，前述した通り，標準偏差が最大値をとる二項分布において証明できたため，混合二項分布に対して，余裕を持って LSC-CAEP が優位であるといえる．また，実験に用いたデータの妥当性に関しては，UCI に掲載されているデータを利用することにより，我々が故意に LSC-CAEP が有利なデータのみ実験することが無いようにした．従って，上記の計算による検定結果は妥当であるといえる．

一方，4．1 節図 4 において α_Δ と計算時間の関係について評価したが，ここでは分類精度との関係を分析した．幾つかのデータに対して support の閾値を 0.1 に固定して，α_Δ を変化させ分類精度に対する依存性を 10CV で検証した．その結果を図 6 に示す．図 6 に示したデータのうち，Cleve，Australian，Crx では，α_Δ の変化に対して分類精度が急激に変化することは無く，α_Δ による分類精度への影響は小さいといえる．Iris データでは，粒度が細かすぎる α_Δ をとった場合，事例数が少なすぎるため不適切な許容距離になり，適切クラスタが得れなかった．Iris データは数値属性のみであるため，CAR を得ることができず，分類不可能になる場合すらあった．同様の現象は Wine や Labor に関してもいえる．とりわけ，Labor データでは α_Δ の値が小さくなるにつれて分類精度が著しく減少している．これは，数値属性に関して適切な CAR が得れず，記号属性のみからなる CAR のみで分類を行ったためであると考えれられる．これらの結果より，データによっては α_Δ の設定が分類精度に大きく影響することがわかる．故に，LSC-CAEP を適用する際は適切な α_Δ を設定する必要があるといえるが，いずれの場合も実験的には概ね $\alpha_\Delta=0.15$〜0.20 程度が良好な精度を得る目安であると言える．

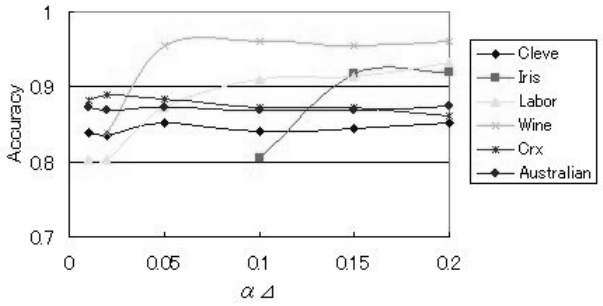


図 6 α_Δ と分類精度の関係

5. 考　　察

本章では，LSC-CAEP で用いる超直方体のクラスタの妥当性について考察する．はじめに，超直方体のクラスタが分類において，任意形状のクラスタに対して不利である場合について述べる．図 7 において，二次元の数値属性に関して (a) に超直方体のクラスタ，(b) に任意形状のクラスタを示した．図 7.(a) において超直方体のクラスタの右上部分に疎である領域が見られるが，(b) の任意形状のクラスタは事例が密である領域のみからなる．このように，超直方体のクラスタは，部分的に疎である領域を持つ可能性がある．故に，超直方体のクラスタは任意形状のクラスタと比較して，分類性能で劣る可能性があるといえる．しかしながら，LSC-CAEP は 4・2 節の分類評価実験において，UCI repository に公開されている 2 3 種類に亘るデータに関して全体として良好な分類精度を示した．これらのデータは多様な性質を有するベンチマーク用のものであり，故に，前述したような問題事象は現実の多様なデータにおいて顕著に発生しない

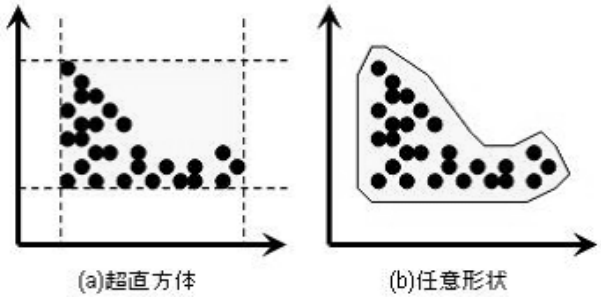


図 7 超直方体のクラスタが有効でない場合

といえる．ただし，本問題については今後の研究において更に検討すべき課題であると考える．

ここまで，超直方体のクラスタが持つ短所について述べたが，超直方体のクラスタは任意形状のクラスタと比較して2つの長所を持つ．1つは，計算複雑性である．4･1節において，LSC-CAEP でクラスタを導出する QFIMiner の計算複雑性が $O(NlogN)$ であることを示した．それに対して任意形状のクラスタを導出する SUBCLU の計算複雑性は $O(N^2)$ である．従って，本手法では超直方体のクラスタを採用することにより，より高速なクラスタリングを可能にしたといえる．2つ目は，クラスタの理解容易性である．QFIMiner が導出するクラスタはその次元数が増大したとしても，各数値属性における多頻度領域の和集合として，人間にとって理解しやすい形式で表現できる．その理解容易性はデータマイニング手法の実践において，高い利用可能性をもたらすといえる．一方，SUBCLU 等で導出される任意形状のクラスタは超直方体形状のクラスタと異なり，一定の形状をとらないため，人間にとって理解しやすい形式の表現にすることは困難である．

以下に，LSC-CAEP の長所である単純かつ明確な結果出力を示し，その理解容易性について述べる．第一の例として，LSC-CAEP を UCI repository に公開されている Labor Relations データに対して適用した．Labor データとは，カナダの 1987 年と 1988 年の前 4 半期に少なくとも 500 人の会員 (教師，警察など) がいる支部のビジネスや個人的サービスの部門の労働契約を含むデータであり，雇用条件とその契約が受け入れられたかどうかが記載されている．データ数は 57 個であり，アイテムの種類数は 16 個，そのうち記号アイテム，数値アイテムはそれぞれ 8 個である．クラスを契約が受け入れられたか否かとして，LSC-CAEP を Labor データに適用した結果，大きな support を持つ次の二つの QFI が得られた．

support=19%: {class:good, duration-years:[2,2],
working-hours:[33,40], wage-inc.-2nd-year(%):[4.0,5.8]}.
support=16%: {class:good, duration-years:[3,3],
working-hours:[35,40], wage-inc.-2nd-year(%):[3.5,5.0]}.

これらの QFI から契約年数が 3 年の人は労働時間が同じ場合，契約年数が 2 年の人と比べて 2 年目の昇給に対する要望が 0.5%から 0.8%低いということがわかる．これは就業条件の選択において，就業者は昇給率が低くても安定した雇用契約期間があれば満足を得るためと考えられる．次に LSC-CAEP を同じく UCI repository に公開されている Iris データに適用した．Iris データとは，アヤメの萼片と花弁についてそれぞれの幅と長さとその品種が記載されたデータである．データ数は 150 個であり，アイテムの種類数は 5 個，そのうち記号アイテムは 1 個，数値アイテムは 4 個である．setosa, versicolour, virginica の 3 種類のアヤメの品種をクラスとして，LSC-CAEP を Iris データに適用した結果，高い Growth Rate を持つ以下の 2 つの CAR を得た．

growth rate=4.5: petal width:[1.4-2.5]) → class:virginica
growth rate=1.9: sepal length:[4.9-7.0],
sepal width:[2.0-3.4] → class:setosa.

これらの CAR から Iris の種別に対する知識を得ることができる．例えば，2 番目の CAR から，萼の長さが 4.9〜7.0cm で萼の幅が 2.0〜3.4cm であれば品種が setosa であることがわかる．このように，事前に設定した離散化区間と異なり，きめ細かい数値属性区間の境界粒度を有し，かつ Growth Rate の高い相関規則は理解しやすい．

先述したクラスタの理解容易性，クラスタリングの低い計算複雑性を確保する立場から，我々は以上のように実際的なクラスタリング手法を開発する立場を優先した．これまでの性能評価より，我々が提案した LSC-CAEP は，高次元で尚且つ高密度なクラスタを分類に適用することにより，高い分類精度と分類の理解容易性を実現したといえる．

6. 結　　論

クラスタリングの基準は手法によって大きく異なるので，我々の提案する LSC-CAEP が用いる部分空間クラスタの良さを簡単に他の手法と比較することは困難である．しかし，提案方法では各属性部分空間内の高密度事例領域を常に包接する超直方体の QFI が導かれる．分類精度の高さから，各クラスに依存する事例分布を区別可能な各クラスデータ D_{cl} 内の高密度クラスタが導かれることがわかる．本論文で提案された原理は，単調性を利用したボトムアップ型のアルゴリズムによって，数値とカテゴリカルアイテムの混合データの効率的部分空間クラスタリングを可能にする．この枠組みに基づけば，様々な学習手法を開発することができ，LSC-CAEP はその一例である．この方向性に沿ったクラスタリングや分類に関する新たな手法開発が期待される．

◇ 参 考 文 献 ◇

[Agrawal 94] Agrawal, R. and Srikant, R.: Fast Algorithms for Mining Association Rules, *Proc. of 20th Int. Conf. on Very Large Data Bases (VLDB)*, pp. 487–499 (1994)

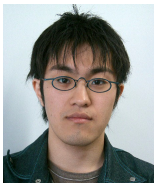
[Agrawal 98] Agrawal, R., Gehrke, J., Gunopulos, D., and Raghavan, P.: Automatic subspace clustering of high di-

mensional data for data mining applications, *Proc. of the 1998 ACM SIGMOD international conference on Management of data*, pp. 94–105 (1998)
[Dong 99] Dong, G., Zhang, X., Wong, L., and Li, J.: CAEP: Classification by Aggregating Emerging Patterns, *Proc. of Second International Conference on Discovery Science, Lecture Notes in Computer Science*, Vol. 1721, pp. 30–42 (1999)
[Ester 96] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X.: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, *In Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining*, pp. 226–231 (1996)
[Fayyad 93] Fayyad, U. M. and Irani, K. B.: Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning, *Proc. of 13th Int. Joint Conf. on Artificial Intelligence (IJCAI)*, Vol. 2, pp. 1022–1027 (1993)
[Kailing 04] Kailing, K., Kriegel, H.-P., and Kroger, P.: Density-Connected Subspace Clustering for High-Dimensional Data, *Proc. Fourth SIAM International Conference on Data Mining (SDM'04)*, pp. 246–257 (2004)
[Li 01] Li, W., Han, J., and Pei, J.: CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules, *Proc. of First IEEE International Conference on Data Mining*, pp. 369–376 (2001)
[Liu 98] Liu, B., Hsu, W., and Ma, Y.: Integrating Classification and Association Rule Mining, *Proc. of Fourth International Conference on Knowledge Discovery and Data Mining* (1998)
[Liu 00] Liu, B., Xia, Y., and Yu, P. S.: Clustering Through Decision Tree Construction, *Proc. of the Ninth International Conference on Information and Knowledge Management*, pp. 20–29 (2000)
[Procopiuc 02] Procopiuc, C. M., Jones, M., Agarwal, P. K., and Murali, T. M.: A Monte Carlo algorithm for fast projective clustering, *Proceedings of the 2002 ACM SIGMOD international conference on Management of data*, pp. 418–427 (2002)
[Quinlan 87] Quinlan, J. R.: Simplifying Decision Tree, *International Journal pf Man-Machine Studies*, pp. 221–234 (1987)
[Quinlan 93] Quinlan, J. R.: *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, London ,England (1993)
[Srikant 96] Srikant, R. and Agrawal, R.: Mining Quantitative Association Rules in Large Relational Tables, *Proc. of 1996 ACM SIGMOD Int. Conf. on Management of Data*, pp. 1–12 (1996)
[Wang 98] Wang, K., Hock, S., Tay, W., and Liu, B.: Interestingness-Based Interval Merger for Numeric Association Rules, *Proc. of 4th Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, pp. 121–128 (1998)

〔担当委員：津本 周作〕

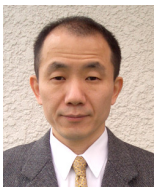
2006 年 1 月 12 日 受理

著 者 紹 介



中西 耕太郎（正会員）

1982 年生．2005 年，大阪大学工学部電子情報エネルギー工学科通信工学専攻卒業．同年，人工知能学会研究会優秀賞を受賞．現在，同大学院工学研究科電気電子情報工学専攻博士前期課程に在籍中．人工知能学会会員．



鷲尾 隆（正会員）

1960 年生．1983 年東北大学工学部原子核工学科卒業．1988 年東北大学大学院原子核工学専攻博士課程修了．工学博士．1988 年から 1990 年にかけてマセチューセッツ工科大学原子炉研究所客員研究員．1990 年（株）三菱総合研究所入社．1996 年退社．大阪大学産業科学研究所助教授（知能システム科学研究部門）．2006 年大阪大学産業科学研究所教授（知能システム科学研究部門）．現在に至る．原子力システムの異常診断手法に関する研究，定性推論に関する研究を経て，現在は人工知能の基礎研究，に科学的知識発見，データマイニングなどの研究に従事．人工知能学会，計測自動制御学会，日本知能情報ファジイ学会，情報処理学会，AAAI，IEEE Computer Society 各会員．

光永 悠紀（正会員）

1981 年生．2004 年大阪大学工学部電子情報エネルギー工学科卒業．2005 年 ICDM"05: The Fifth IEEE International Conference on Data Mining 国際会議にて発表．2006 年大阪大学工学部電子情報エネルギー工学科通信工学専攻修了．同年，日本電気株式会社に入社．現在に至る．

藤本 敦（正会員）

2003 年，大阪大学工学部電子情報エネルギー工学科卒業．2005 年，大阪大学大学院通信工学専攻修了．同年，富士通テン株式会社入社．現在に至る．

元田 浩（正会員）

1943 年生．1965 年東京大学工学部原子力工学科卒業．1967 年東京大学大学院原子力工学専攻修士課程修了．同年（株）日立製作所入社．同社中央研究所，原子力研究所，エネルギー研究所，基礎研究所を経て 1995 年退社．1996 年大阪大学産業科学研究所教授（知能システム科学研究部門）．2006 年退官．米国空軍科学技術局アジア宇宙航空研究開発事務所 (AOARD) 科学顧問．現在に至る．原子力システムの設計，運用，制御に関する研究，診断型エキスパート・システムの研究を経，現在は人工知能の基礎研究，特に機械学習，知識獲得，知識発見などの研究に従事．工学博士．認知科学会，人工知能学会，情報処理学会，日本ソフトウエア科学会，AAAI, IEEE Computer Society, 各会員．