論 文

Technical Papers

適応的密度基準に基づく部分空間クラスタリングを用いた定量的多頻度アイテム集合のマイニング

Mining Quantitative Frequent Itemsets Using Adaptive Density-based Subspace Clustering

光永 悠紀 Yuki Mitsunaga
大阪大学産業科学研究所
The Institute for Scientific and Industrial Research,Osaka University
mitsunaga@ar.sanken.osaka-u.ac.jp

鷲尾 隆 Takashi Washio
（同　上）
washio@ar.sanken.osaka-u.ac.jp

元田 浩 Hiroshi Motoda
（同　上）
motoda@ar.sanken.osaka-u.ac.jp

keywords: subspace clustering, quantitative frequent itemset, quantitative association rule, Apriori algorithm, data mining

Summary

A novel approach to subspace clustering is proposed to exhaustively and efficiently mine quantitative frequent itemsets (QFIs) from massive transaction data for quantitative association rule mining. The numeric part of a QFI is an axis-parallel and hyper-rectangular cluster of transactions in an attribute subspace formed by numeric items. For the computational tractability, our approach introduces adaptive density-based and Apriori-like subspace clustering. Its outstanding performance is demonstrated through the comparison with the past subspace clustering approaches and the application to practical and massive data.

1. はじめに

従来の相関規則マイニング手法の多くは，記号アイテムからなるトランザクションデータ中で，最小支持度 “*minimum support* ($minsup$)” 以上の頻度で現れる多頻度アイテム集合を導出し，更にそれらの共起関係を導出する．この重要な拡張として，数値アイテムと記号アイテム両方を含むトランザクションデータセットから，定量的多頻度アイテム集合 “*Quantitative Frequent Itemset* (QFI)” をマイニングし，それに基づいて定量的相関規則 “*Quantitative Association Rules* (QARs)” を導出することが挙げられる [Srikant 96]．定量的アイテム集合 “*Quantitative Itemset* (QI)” は，“$\{< p_1 : q_1 >, ..., < p_m : q_m >\}$” という形式で記述される．$< p : q >$ が一つのアイテムを表し，p が属性，q がその値である．例えば，“$\{< Age : [30, 39] >, < NumCars : [2, 2] >, < Married : Yes >\}$” は “年齢が 30 代で，車を 2 台所持しており，結婚している.” ことを表す．数値アイテムは区間値を持ち，記号アイテムはカテゴリカルな値を持つ．2 つの数値アイテム $< p : q >, < p_t : q_t >$ が与えられた時，$p_t = p$ でかつその区間値 q_t が区間値 q の範囲内である時，数値アイテム $< p : q >$ は数値アイテム $< p_t : q_t >$ に支持される．トランザクション t が与えられた時，ある定量的アイテム集合 (QI) に含まれる各アイテムが t 内のいずれかのアイテムに支持されるなら，この QI は t に支持される．例えば，“$t_1 = \{< Age : [35, 37] >, < Married : Yes >, < NumCars : [2, 2] >, < Child : [3, 3] >\}$” は前述の QI を支持し，一方，“$t_2 = \{< Age : [29, 31] >, < Married : Yes >, < NumCars : [2, 2] >, < Child : [3, 3] >\}$” は，$< Age : [29, 31] >$ が $< Age : [30, 39] >$ の範囲内ではないために支持しない．トランザクションデータセット D が与えられた時，QI が最小支持度 ($minsup$) 以上の数のトランザクションに支持されていれば，それを定量的多頻度アイテム集合 (QFI) と呼ぶ．QFI の数値アイテムで構成される部分集合は，D の全属性空間の部分空間中の各軸に平行な超直方体に相当する．また，定量的相関規則 (QAR) は，“$\{< p_1 : q_1 >, ..., < p_i : q_i >\} \Rightarrow \{< p_{i+1} : q_{i+1} >, ..., < p_m : q_m >\}$” という形式で記述され，2 つの QFI“$\{< p_1 : q_1 >, ..., < p_i : q_i >\}$” と “$\{< p_1 : q_1 >, ..., < p_i : q_i >, < p_{i+1} : q_{i+1} >, ..., < p_m : q_m >\}$” の共起関係を表す．例えば，“$\{< Age : [30, 39] >, < NumCars : [2, 2] >\} \Rightarrow \{< Married : Yes >\}$” は，“

年齢が 30 代で車を 2 台所持している人ならば，それに加えて結婚している." ことを表す.

QAR マイニングの先駆的研究は Srikant と Agrawal によって行われた QARM である [Srikant 96]．彼らは，前処理で各数値アイテムが取る領域を同じ個数のトランザクションが含まれる空間に分割し，更に含まれるトランザクションの個数が最大支持度に達するまで隣り合う区間を結合し，その後，従来用いられてきた Apriori アルゴリズムを適用して多頻度アイテム集合をマイニングした．この手法計算の複雑さは，$N = |D|$ とした時に $O(N)$ である．Wang は共起性に関する興味深さを定義し，それを基準にして隣接する区間を結合する十分実用的な効率性を有する $O(N \log N)$ の手法 M-tree を提案した [Wang 98]．しかし，これらの手法では最良優先探索戦略の下で各属性を他属性とは独立に離散化するので，トランザクションが複数属性に複雑に依存して分布している場合，しばしば適切な離散化を行うことができない．この問題を解決する方法として，与えられた最小支持度以上を有しかつその中で最大の確信度をもつ QAR が導出されるように，各数値アイテムの値域を離散化しながら QAR をマイニングする最適アルゴリズムである SONAR [Fukuda 01]，Opt.AR [Rastogi 98] の研究が行われた．しかし，この最適化は一般に NP-完全な問題であることが知られている [Wijsen 98]．即ち，D で扱う数値属性の次元数を限定しないと，各離散化区間の上限と下限の候補は指数関数的に増加するため実用的でない.

一方，現実的な計算量で最良の QAR を導出するために，データ中の数値アイテムによって構成される数値属性空間内の各部分空間上でトランザクションが密集した軸平行超直方体クラスタを，QFI として探索する部分空間クラスタリングの導入が考えられた．ある属性空間内で高密度のクラスタに含まれるトランザクションは，その部分属性空間内の高密度クラスタにも常に含まれるという単調性がある．この性質と記号アイテム集合の支持度の単調性を利用して，効率的な定量的多頻度アイテム集合 (QFI) 探索アルゴリズムが設計できる．QLIQUE では，はじめに数値属性空間を軸平行な格子状に離散化して，次に密度の濃い隣接するブロックを，階層的に結合していくことによりクラスタを見つける [Agrawal 98]．ENCLUS [Cheng 99]，MAFIA [Goil 99]，SCHISM [Sequeira 04] は，可変な格子幅と可変な密度の閾値を導入し，更にエントロピーを用いたクラスタの興味深さの評価指標を取り入れている．DOC では各属性部分空間において，超立方体の窓を動かすことによりトランザクションの密度を測る手法を導入している [Procopiuc 02]．CLTree では，エントロピーに基づく密度基準を用いて階層的な最良優先探索を行い，軸に平行な超直方体の部分空間クラスタを探索する [Liu 00]．これらの手法の計算量は $O(N) \sim O(N \log N)$ と少ないが，用いる格子や窓の形及び大きさ，向きが不適切で探索戦略も不十分なため，クラスタを見逃すことがあった．最近開発された SUBCLU は，DBSCAN で提案された厳密な密度基準の下で部分空間クラスタを探索する [Kailing 04, Ester 96]．各トランザクションから半径 ϵ の近傍に，閾値 $MinPts$ 個以上の他のトランザクションが存在する場合，その密度の濃い領域を "*density-connected set*" と呼ぶ．この手法は，高密度クラスタの単調性を用いて Apriori に似た幅優先探索アルゴリズムで全属性部分空間を完全探索し，非常に高次元な部分空間のクラスタ以外は見逃さない．しかし，全てのトランザクション間の距離を計算するため，計算量は $O(N^2)$ である．また，この手法は数値アイテムしか扱えない.

表 1 QFI マイニング手法の性能比較

Method	IT	SR	$O()$	SZ	CS
QARM	NC	C	N	∞	RT
M-tree	N	C	$N \log N$	∞	RT
SONAR	N	C	N^3	2	RC
Opt.AR	NC	C	C^N	∞	RT
QLIQUE	N	C	N	∞	UR
ENCLUS	N	C	N	∞	UR
MAFIA	N	C	N	∞	UR
SCHISM	N	C	N	∞	SP
CLTree	N	G	N	∞	UR
DOC	N	R	N	∞	RT
SUBCLU	N	C	N^2	∞	FS
QFIMiner	NC	C	$N \log N$	∞	RT

IT: Items to be mined,
N; Numeric only, NC; Numeric and cathegorical,
SR: Search, C; Complete, G; Greedy, R; Random,
$O()$: Computational time complexity for N,
SZ: Size limit of QFI to be mined,
CS: Cluster shape, RT; Rectangle,
RC; Rectilinear convexity, UR; Union of Rectangles,
SP; Subspace, and FS; Free shape.

本論文では以上の背景に基づき，定量的多頻度アイテム集合 (QFI) のマイニングに焦点をあて，新たな手法 "*QFIMiner*" を提案する．QFIMiner と従来手法の違いを表 1 にまとめる．表から分かるように QFIMiner には，以下の特徴がある.

(1)対象トランザクションデータから数値及び記号アイテムで構成される QFI をマイニング可能である.
(2)属性の密度判定が正確な場合には，全ての QFI を探索可能なアルゴリズムを採用している.
(3)効率的な密度推定方法により，$minsup$ 以上のトランザクションに支持される QFI を従来手法より効率的に探索可能な事が実験的に示されている.
(4)計算量が実用的に十分な実質 $O(N \log N)$ である.
(5)探索対象とする QFI の大きさ，即ち，含まれるアイテム数に制限がない.
(6)数値アイテム部分空間クラスタの形状は軸平行超直方体をである.

特徴 (2) の蓋然的仮定に関しては後に詳述する．最後の特徴は，QFIMiner が導くクラスタ形状が単純なものに限

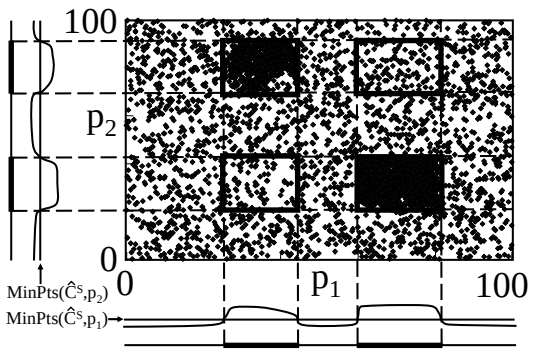


図 1 クラスタとその射影

られることを表すが，これにより QFI や QAR を区間値を有する数値アイテムの集合で表すことができ，マイニング結果の高い可読性が得られる．このように QFIMiner は，従来手法と異なり表 1 の全ての項目に亘り，高い性能と機能を有する．更に表に含まれない QFIMiner の新たな機能として，

(7)対象トランザクションデータに含まれる数値アイテムに区間値を許す．

が挙げられる．従来手法では，対象トランザクションに例えば $< Age : 36 >$ という数値アイテムが含まれることはあっても，$< Age : [35, 37] >$ のような区間値を有する数値アイテムが含まれる場合を扱うことができなかった．

次節では，提案する QFIMiner についての概要を説明する．そして，3 節では手法の詳細について説明し，最後に 4 節で QFIMiner の優れた評価性能を示す．

2. QFIMiner の概要

QFIMiner は数値アイテムと記号アイテムからなるトランザクションの集合であるデータセット D から QFI を探索する．トランザクションの数値部分は，D の全属性空間の部分空間 S に含まれる軸平行な超直方体領域を表す．今，トランザクションが高密度に集まったクラスタが，それ以外のバックグラウンドノイズの上に存在する場合を考える．図 1 は，$S = \{p_1, p_2\}$ で構成される属性部分空間において，各トランザクションの数値アイテムが区間値ではなくスカラー値を持っており，バックグラウンドノイズの中に二つの高密度のクラスタが存在する例を表している．

QFIMiner は，トランザクション分布の密度評価のために前もって決めた格子や窓を用いるのではなく，DBSCAN と同様な密度の定義を使用する．この手法は，適切な密度の閾値を用いることでクラスタを見逃す可能性を大きく減らすことができる．QFIMiner では，一次元部分空間のクラスタ探索からはじめ，$(k-1)$ 次元のクラスタを結合して k 次元の部分空間 S からクラスタの候補 $\hat{C}^S$ を順次生成していく幅優先探索アルゴリズムを用いる．これは SUBCLU と似ているが，QFIMiner は，幅優先部分空間クラスタリングを標準的な Apriori アルゴリズムに埋め込むことで数値アイテムと記号アイテム両方からなるクラスタを導出することができる．即ち，数値アイテムと記号アイテムで構成される属性部分空間に存在する，最小支持度 ($minsup$) 以上の数のトランザクションに支持されたクラスタを効率良く探索できる．

$O(N^2)$ の計算時間を回避するために，QFIMiner では各トランザクションペア間の高次元属性空間での距離を計算しない．そのかわりに，部分空間 S のそれぞれの属性軸上で，高密度クラスタの候補 $\hat{C}^S$ に含まれるトランザクション間の距離を計算する．図 1 は，$\hat{C}^S$ が $[0, 100] \times [0, 100]$ の領域をとる場合である．今，ある属性に関する $density-connected\ set$ を，その属性軸上で $\pm\Delta$ 以内の距離に少なくとも $MinPts$ 個以上の他のトランザクションが存在する連続して連なったトランザクションの集合とする．更に，$maximal\ density-connected\ set$ を，他の何れの $density-connected\ set$ にも含まれない $density-connected\ set$ とする．QFIMiner は数値属性部分空間において，このような全ての $maximal\ density-connected\ set$ を探索する．密度の単調性により，属性部分空間 S 内の全ての属性に対する $maximal\ density-connected\ set$ の共通部分が新しいクラスタ候補 $\hat{C}^S$ となる．図 1 では，黒枠のボックスで表わされる 4 つの共通部分が新たな $\hat{C}^S$ である．各 $\hat{C}^S$ について S のどの軸に射影しても高密度クラスタ C^S に収束するまで，射影と $maximal\ density-connected\ set$ の探索を繰り返し行う．図 1 では，この繰り返し操作により高密度の条件に合致する 2 つのクラスタが残り，他は削除される．各属性軸について，ソートを行ったトランザクションの一回の走査により密度を評価可能なので，このアルゴリズムの平均計算量は $O(N \log N)$ である．

ある軸に対して $maximal\ density-connected\ set$ を探索する時，密度閾値 $MinPts$ がバックグラウンドノイズより小さいと高密度クラスタがノイズに埋もれてしまう可能性がある．一方，$MinPts$ が大きすぎてもクラスタを見逃してしまう可能性がある．そこで，$\hat{C}^S$ が属性部分空間 S の平均の密度を持つと仮定した場合に各属性軸 p において $\pm\Delta_p$ 近傍に存在するトランザクション数の期待値を $MinPts(\hat{C}^S, p)$ として，これを属性軸 p に関する閾値 $MinPts$ として用いる．$MinPts(\hat{C}^S, p)$ は常に高密度クラスタの密度とバックグラウンドノイズの密度の間の値をとる．図 1 では，$MinPts(\hat{C}^S, p)$ により効率的に高密度クラスタに対応する $maximal\ density-connected\ set$ を抽出している．この適応的に密度の閾値を変化させる手法では，$MinPts(\hat{C}^S, p)$ は低次元では高くなりノイズレベル以下のより多くの余分な $maximal\ density-connected\ set$ が取り除かれるため，QFIMiner は高速化される．各属性軸上の射影密度と適応的密度閾値の使用が，QFIMiner の出力品質の維持と計算時間の低減を両立させる鍵になっている．

3. 手法とアルゴリズム

3·1 部分空間クラスタリング

はじめに，数値アイテムのみで構成されたトランザクションに関する部分空間クラスタリングに焦点を当てる．$MinPts$ の密度閾値は一般性を失わずに一定値のままとし，適応的な $MinPts(\hat{C}^S, p)$ への拡張については後で説明する．

【定義 1】（Neighborhood） p を数値属性とし，かつ t と t' は各々区間値 q と q' をもつ p を共有するトランザクションとする．また，この二つのトランザクションの軸 p 上での距離 $Dist_p(q,q')$ を区間 q と q' の間の最小の距離 $\min_{v\in q, v'\in q'}|v-v'|$ とする．そして，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 ($\Delta_p - neighborhood$)” $N_{\Delta_p}(t)$ を以下の式で定義する．

$$N_{\Delta_p}(t) = \{t' \in D | Dist_p(q,q') \leq \Delta_p\}.$$

q と q' の区間に互いに重なる部分がある場合は $Dist_p(q,q') = 0$ となり，それ以外の場合は $Dist_p(q,q')$ は両区間が互いに面している境界間の距離となる．

【定義 2】（Core transaction） データセット D 中のあるトランザクション $t(\in D)$ がその Δ_p-近傍 $N_{\Delta_p}(t)$ に少なくとも “最小トランザクション数 ($minimum points$)” $MinPts$ のトランザクションを含む，即ち

$$|N_{\Delta_p}(t)| \geq MinPts$$

である時，そのトランザクション t を属性 p 上の “*core transaction*” と呼ぶ．

【定義 3】（Direct Density-Reachability） ある 2 つのトランザクション $t, t'(\in D)$ について，t が属性 p 上で $core transaction$ であり，そして t' が $N_{\Delta_p}(t)$ に含まれる時，属性 p に関して t' は t から “*directly density-reachable*” であると言う．

【定義 4】（Density-Reachability） D 内で属性 p 上で t_i から t_{i+1} が $direct\, density-reachable$ であるトランザクションの連なり $t_1, t_2, \cdots, t_{n-1}, t_n$, があり，$t_1 = t$、$t_n = t'$ である時，t' は t から属性 p 上で “*density-reachable*” であると言う．

【定義 5】（Density-Connectivity） D 内で属性 p 上で t と t' の両方へ $density-reachable$ であるトランザクション t'' が存在する時，t と t' は互いに “*density-connected*” であると言う．

【定義 6】（Density-Connected Set） 空でないトランザクション集合 $C(\in D)$ に含まれる全てのトランザクションが互いに属性 p 上で $density-connected$ である時，C を属性 p 上の “*density-connected set*” と呼ぶ．

【定義 7】（Dense Cluster） 数値属性の集合 S で構成される部分空間内で，D においてトランザクション集合 $C^S(\subseteq D)$ が全ての数値属性 $p(\in S)$ 上で空でない $density-connected\, set$ である時，C^S を S 上の $density-connected\, set$ という．更に C^S が如何なる他の S 上の $density-connected\, set$ にも含まれない極大 ($maximal$) な集合である時，C^S を “高密度クラスタ ($dense-cluster$)” と言う．

表 2 トランザクションデータセットの例 D

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

【定義 8】（Quantitative Frequent Itemset） C^S $(\subseteq D)$ を部分空間 S における高密度クラスタとする．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 を包略するアイテム集合である．ここでもし $|C^S| \geq minsup$ 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 とした時，$a(C^S)$ を k-QFI と呼ぶ．

QFI は部分空間内で極大な体積を持ち軸に平行で高密度な超直方体の単一領域である．SUBCLU が探索対象とする高密度クラスタと同様に，QFI は以下のような単調性を持つ．

【補題 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 において QFI であれば，T 内に $a(C^S)$ によって支持される QFI である $a(C^T)$ が存在する．つまり，T 内に $a(C^S) \subseteq a(C^T)$ である $a(C^T)$ が存在する．

Proof. C^S 内の全てのトランザクションは全ての属性 $p \in S$ 上で $density-connected$ であるため，全ての属性 $p \in T$ 上でも $density-connected$ である．従って，$C^S \subseteq C^T$ が成り立ち，全ての $p \in T$ 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． ∎

よって，低次元部分空間からボトムアップに幅優先探索を適用して，全ての QFI を探索できる．この探索の具体例を 表 2 に示すデータセットを用いて，$\Delta_{Age} = 5$, $\Delta_{Child} = 1, \Delta_{NumCars} = 1, \Delta_{Savings} = 1K, MinPts = 1$ and $minsup = 2$ というパラメータ設定の条件下で説明する．まず，各 t_i 内のアイテムを属性名で辞書順に整列する．表はすでにこの操作をすませた状態である．次に各属性について，$maximal\, density-connected\, set$ である 1 次元の QFI(1-QFI) を探索する．$\Delta_{Age} = 5$ の条件下で，属性 Age には 30 から 39 に亘って高密度の領域があり，かつその支持度は 5 で $minsup$ より大きいため，Age について一つの 1-QFI$< Age : [30;39] >$ が存在する．表 3 には 1-QFI とその “*transaction id list* (TID-List)” が示されている．この例では各属性は一つ

表 3 D に対する幅優先部分空間クラスタリング

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 $(\{< Age : [35,39] >, < Child : [2,2] >,$ $< NumCars : [2,3] >\}, \{t_5, t_6\})$ $(\{< Age : [30,35] >, < Child : [4,5] >,$ $< NumCars : [1,1] >\}, \{t_2, t_4\})$

ずつ 1-QFI を持つ．

次のステップでは，k-QFIs $(k > 1)$ の幅優先探索が開始される．標準的な AprioriTid algorithm と同様に，D 内の各トランザクションを表すインデックスリストである $TID - List$ が用いられる [Agrawal 94]．今，全ての $(k-1)$-QFI が既知であると仮定すると，以下の “候補生成 $(Candidate - Generation)$” 処理によって全ての $candidate - k - QFI$ が導出される．

【定義 9】(Candidate-Generation)

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，

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

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．

この枝狩りのフェーズは補題 1 に基づいている．式 (1) において，q^c_i が q_i と交わる限り，s と $(k-1)$-QFI が $minsup$ 以上の数のトランザクションを共有する可能性は否定できず，単調性を満たす可能性は失われない．よって，$candidate - k$-QFI は残される．表 3 では，$candidate -$ 2-QFI$\{< Age : [30,39] >, < Child : [2,5] >\}$ とその $TID - List^c = \{t_2, t_3, t_4, t_5, t_6\}$ が，二つの 1-QFI$\{< Age : [30, 39] >\}$ と $\{< Child : [2,5] >\}$ から導出される．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．

```
QFI-Count(candidate − k − QFI, TID − List^c);
/* Notions of input arguments follow Definition 9.*/
 (1)k − QFIS = φ, TIDLS = φ;
 (2)If |TID − List^c| < minsup return k − QFIS;
 (3)S = {p| < p : q >∈ candidate − k − QFI, p is numeric.};
 (4)TIDLS.temp = {TID − List^c};
 (5)while TIDLS ≠ TIDLS.temp do begin
 (6)    TIDLS = TIDLS.temp;
 (7)    forall p ∈ S do begin
 (8)      TIDLS.temp = MDCS(TIDLS.temp, p);
 (9)    end
(10)end
(11)forall TID − List ∈ TIDLS do begin
(12)    k − QFIS = k − QFIS + (QFI(S, TID − List), TID − List);
(13)end
(14)return k − QFIS;
```

図 2 QFI-Count アルゴリズム

補題 1 より，高密度クラスタ C^S $(|S| = k)$ は k-QFI に対応し，かつ $C^S \subseteq C^T$ と $C^S \subseteq C^{T'}$ に従う．ただし，C^T と $C^{T'}$ $(|T| = |T'| = k-1)$ は共に高密度クラスタでそれぞれ $(k-1) - QFI$ と $(k-1) - QFI'$ に対応するものとする．$\hat{C}^S$ は C^T と $C^{T'}$ の共通部分なので，$C^S \subseteq \hat{C}^S$ である．従って，高密度クラスタ C^S とその k-QFI は，もしそれらが存在すれば，$\hat{C}^S$ に含まれるトランザクションの密度を評価することで導出できる．この際，図 2 に示される “QFI-Count” のアルゴリズムが，定義 7 と定義 8 に基づいて，高密度クラスタ $C^S = TID - List$ とその k-QFI を探索する．

まずはじめに，候補生成で生成された $TID - List^c = \hat{C}^S$ を持つ k-QFI の候補が与えられる．もし $|\hat{C}^S|$ が $minsup$ より小さい場合，ステップ (2) で $k - QFIS = \phi$ が返される．図中のステップ (7) から (9) にかけての内周ループでは，関数 $MDCS$ によって各数値属性軸 p 上で極大な $density - connected\,set\,C$ が探索される．はじめに $MDCS$ には，高密度クラスタ候補 $TIDLS.temp = \{TID - List^c\} = \{\hat{C}^S\}$ と何れの数値属性軸上で極大な $density - connected\,set\,C$ を探すかを指定する数値属性 p が引数として与えられる．また，実際にはこれに加えて密度判定に必要な許容距離 Δ_p と最小トランザクション数 $MinPts$ も外部から指定される．そして，これらの Δ_p，$MinPts$ の下で定義 6 に沿って，高密度クラスタ候補 $\hat{C}^S$ 中である p 上極大な $density - connected\,set\,C$ が探索される．$\hat{C}^S$ に複数の高密度クラスタが含まれている場合は複数の C が見つかり，$MDCS$ からそれらのリスト $TIDLS.temp$ が出力される．同ループ 2 回目からは，$MDCS$ は前回のループにおいて $TIDLS.temp$ 内に導出された各 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．この際，削減や細分化によって $minsup$ 以下のトランザクションしか含まない C は，$MDCS$ 内部で除去される．内周ループで S の全ての数値属性 p についてこの処理を行った後も，外周ループであるステップ (5) から (10) を通じて上記内周ループ処理は繰り返され，$TIDLS.temp$ 内の各 C が高密度クラスタ C^S に収束するまで続けられる．各 C^S はその単調性により，探索経路に依存しないで導出することができる．外周ルー

(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 全体のアルゴリズム

プを抜けた段階で，$TIDLS$ 中の各 $TID-List$ が各高密度クラスタ C^S に対応する．

ステップ (11) から (13) までのループでは，関数 QFI において定義 8 によって各 C^S に対応する QFI が計算される．部分空間を表す各数値属性の集合 S とある高密度クラスタ C^S に属するトランザクションの ID リスト $TID-List$ から，定義 8 に従って各数値属性 $p(\in S)$ 軸上のトランザクションの最大値 $max_p(C^S)$ 及び最小値 $min_p(C^S)$ が求められ，これから $a(C^S)$，すなわち QFI が導かれる．

ここで表 3 の例で，$TID-List^c = \{t_2, t_3, t_4, t_5, t_6\}$ が与えられた時，その $candidate-2$-QFI$\{< Age : [30,39] >, < Child : [2,5] >\}$ について考える．内週ループでは，$MDCS$ は $\Delta_{Age} = 5$ という条件での属性 Age に対する C として $TIDLS.temp = \{\{t_2, t_3, t_4, t_5, t_6\}\}$ を導出する．次に，この $TIDLS.temp$ に対して $\Delta_{Child} = 1$ の条件下で属性 $Child$ について $MDCS$ を適用すると，$TIDLS.temp = \{\{t_3, t_5, t_6\}, \{t_2, t_4\}\}$ が得られる．しかし，これ以上 $MDCS$ を適用しても，$TIDLS.temp$ は変わらない．それぞれの候補のサイズが $minsup = 2$ 以上であるので，2 つの 2-QFI($\{< Age : [30,39] >, < Child : [2,2] >\}, \{t_3, t_5, t_6\}$) と ($\{< Age : [30,35] >, < Child : [4,5] >\}, \{t_2, t_4\}$), が導出される．

3·2 QFI の導出

候補生成処理は数値アイテムと記号アイテムからなる QFI を導出す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 アルゴリズムと同じように与えられる．

【定義 10】（Extended-Candidate-Generation）
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:

$(candidate-k-QFI = \{< p_1 : q_1^c >, ..., < p_{k-1} : q_{k-1}^c >, < p_k : q_k^c >\}, TID-List^c)$.

ここで q_i^c 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 $q_i^c = q_i = q'_i$ である．また数値，記号何れのアイテムの場合も $q_{k-1}^c = q_{k-1}$，$q_k^c = q'_k$ 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 $q_i^c = \phi$ 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，

$\forall < p_i : q_i^c > \in s, \exists < p_i : q_i > \in (k-1)-QFI$,
全ての数値アイテムについて $q_i^c \cap q_i \neq \phi$
かつ全ての記号アイテムについて $q_i^c = q_i$,

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 アルゴリズムも変更が必要である．$candidate-k$-QFI が記号アイテムのみからなる場合は，ステップ (5) から (10) までのループを飛ばし，$TIDLS = TIDLS.temp$ 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 $q_i^c = q_i = q'_i$ とする．

D から QFI を導出するアルゴリズムの全体を 図 3 に示す．必要となる入力パラメータは，全ての数値属性に関する許容距離 Δ_p と最小トランザクション数 $MinPts$，最小支持度 $minsup$ である．はじめに $Extended$-$Candidate$-$Generation$ と QFI-$Count$ で効率的な探索を行うためにインデックスリストを作成する．次に，AprioriTid アルゴリズムの $Join$ を $Extended$-$Candidate$-$Generation$ 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トランザクション 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，トランザクション数 N について $O(NlogN)$ 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．図 3 のはじめの操作で作成したインデックスリストを用いることで，$MDCS$ では全ての数値属性軸 p 上の $maximal\ density-connected\ set$ を $TIDLS.temp$ を一度スキャンするだけで簡単に導出することができる．この操作には最大で $O(N)$ かかる．図 2 に示す QFI-Count における外周ループのステップ (5) から (10) の反復回数は，部分空間におけるトランザクションの分布に強く影響をうける．一度の反復で一つのトラ

ンザクションしか取り除かれないという最悪の場合では，合計で $O(N^2)$ のコストがかかってしまう．しかし，もっともよくあると考えられるケースは，一度の反復で C に含まれるトランザクション数の一定割合 $0 < r < 1$ が除かれる場合である．この場合には m をループ反復回数としたとき，$r^m N$ が $minsup$ より小さくなるとループが終わる．従って，$minsup \simeq r^m N$ より，m はおよそ $O(\log N)$ になる．よって，アルゴリズム全体の計算時間の期待値は $O(N \log N)$ となる．

3·3 適応的密度閾値

どこまでのトランザクションが高密度クラスタに属し，どこからがはずれ値になるかの判断はほとんど主観的な問題であるので，バックグラウンドノイズの中から高密度クラスタを正しく識別するための密度閾値の最適性を一般的に定義することは難しい．そこで最適性よりもむしろ以下の考察に基づいた頑健な密度閾値を提案する．トランザクションが空間に一様に分布してる場合を除いて，高密度クラスタは空間の他の領域より相対的に高密度の領域に存在すると考えられる．これは次の蓋然的仮定が成立することを意味する．

［仮定 1］（Average Density） はずれ値が空間内のトランザクション密度の平均よりも密度の薄い領域に位置するのに対し，高密度クラスタは空間内のトランザクション密度の平均よりも高密度の領域に存在する．

D^S を部分空間 S を構成する全ての属性に対応するアイテムを含む，データセット D 内のトランザクションの集合とする．つまり，D^S の各トランザクションは S の中に存在している．トランザクションが存在している S 内の領域における D^s の全トランザクションに関する平均密度 $\bar{d}^S$ を考える．上の仮定から，高密度クラスタのほとんどは $\bar{d}^S$ 以上の密度をもつ領域に位置し，はずれ値のほとんどはこの領域の外に位置する．一方，高密度クラスタ C^S は先に説明した高密度クラスタの候補である $\hat{C}^S$ の中に存在する．これは $\bar{d}^S$ より高密度の C^S の領域が，$\hat{C}^S$ 領域の中に存在することを意味している．図 2 の $MDCS$ が一つの軸 p について $maximal\,density-connected\,set$ を探索するとき，$\hat{C}^S$ に含まれるトランザクション t は軸 p に射影され各 t について，定義 1 の Δ_p-近傍 $N_{\Delta_p}(t)$ が計算される．もし $\hat{C}^S$ の中に高密度クラスタ C^S が存在するならば，$\hat{C}^S$ の中の多くの t の $N_{\Delta_p}(t)$ が，$\hat{C}^S$ の密度が $\bar{d}^S$ と等しい場合に想定される値より大きくなる．よって，以下の戦略を考える．

戦略 1（Average Density Threshold） $\hat{C}^S$ の密度が $\bar{d}^S$ であるときの $N_{\Delta_p}(t)$ の期待値 $MinPts(\hat{C}^S, p)$ を適応的な密度閾値として用いる．

［補題 2］（Average Minimum Points (MinPts)） $MinPts(\hat{C}^S, p)$ は次のように与えられる．

$$MinPts(\hat{C}^S, p) = |D^S| r_p,$$
$$\text{ここで } r_p = \frac{2\Delta_p}{R_p} \prod_{p' \in S, p' \neq p} \frac{C_{p'}}{R_{p'}},$$

C_p は p 上の $\hat{C}^S$ の幅であり，R_p は D^S に含まれるトランザクションの p に関する範囲である．

Proof. 部分空間 S にトランザクションが存在している領域の体積は $\prod_{p' \in S} R_{p'}$ であり，従ってその平均密度 $\bar{d}^S$ は $|D^S| / \prod_{p' \in S} R_{p'}$ である．$\hat{C}^S$ は超直方体であるので，S 内のその体積は $\prod_{p' \in S} C_{p'}$ である．よって，平均密度 $\bar{d}^S$ を持つ $\hat{C}^S$ に含まれるトランザクションの平均個数は，$|D^S| \prod_{p' \in S} C_{p'} / R_{p'}$ となる．これらのトランザクションは，軸 p 上の範囲 C_p に存在するので，トランザクションの内 $2\Delta_p / C_p$ の割合は p 上の $\pm\Delta_p$ の範囲に存在する．よって，$N_{\Delta_p}(t)$ の期待値は；

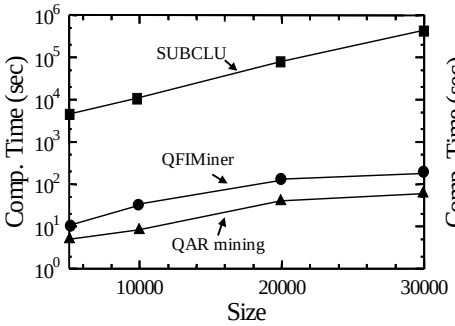
$$MinPts(\hat{C}^S, p) = \frac{2\Delta_p}{C_p} |D^S| \prod_{p' \in S} \frac{C_{p'}}{R_{p'}} = |D^S| \frac{2\Delta_p}{R_p} \prod_{p' \in S, p' \neq p} \frac{C_{p'}}{R_{p'}} \blacksquare$$

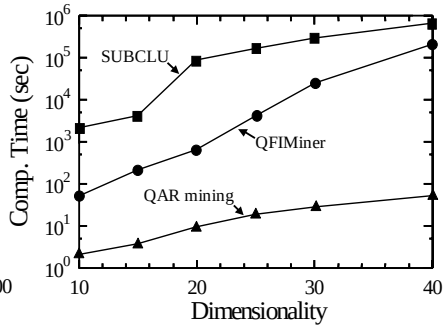
この $MinPts(\hat{C}^S, p)$ は，図 2 の $MDCS$ の関数の中で全ての $maximal\,density-connected\,set$ を導出する際に適用される．

QFIMiner に入力されるパラメータは Δ_p と $minsup$ であるが，Δ_p は各 p について範囲 R_p に対する一定の相対幅 Δ によって定義される．Δ と $minsup$ により $MinPts(\hat{C}^S, p)$ は効率的に多くのバックグラウンドノイズ中の小さなクラスタも抽出可能である．例えば $|D| = 4100$ 個のデータが 2 次元空間 $\{p_1, p_2\}$ にあり，その内 4000 個のトランザクションが空間中の領域 $[0, 100] \times [0, 100]$ にバックグラウンドノイズとして一様に分布していて，残りの 100 個のトランザクションを高密度クラスタを作るべく $[40, 60] \times [40, 60]$ の領域に加えたとする．$\Delta = 2\%$ (ただし $\Delta_{p_i} = \Delta R_{p_i}$) の条件下では，$k = |S| = 1$ である $i = 1, 2$ について $MinPts(\hat{C}^{\{p_i\}}, p_i)$ は $2\Delta_{p_i} / R_{p_i} |D^{\{p_i\}}|$ であるので，$R_{p_i} = 100$ 及び $|D^{\{p_i\}}| = |D| = 4100$ より，その値は 164 となる．一方，$[40, 60]$ の領域に存在するノイズトランザクションの数の期待値は 800 個で，そこに加えて 100 個のトランザクションが存在している．よって，$[40, 60]$ 内の $N_{\Delta_p}(t)$ の範囲に存在するトランザクション数の期待値は $2\Delta_{p_i} / (60 - 40) \times 900 = 180$ となる．これに対してこの他の領域で $N_{\Delta_p}(t)$ の範囲に存在するトランザクション数の期待値は 160 となる．$MinPts(\hat{C}^S, p)$ とこれらの $N_{\Delta_p}(t)$ の違いは大きくはないが，高密度クラスタに含まれるトランザクションのほとんどが $core\,transaction$ となる．また，これらのトランザクションは各 p_i 上の領域 $[40, 60]$ で $maximal\ density-connected\ set$ を形成する．一方，これ以外の領域の大部分はこのような集合を形成しないか，またはノイズによりたまたま $maximal\,density-connected\,set$ が形成されてもそのほとんどが $minsup$ によって削除される．

4. 評 価 実 験

QFIMiner の効率・クラスタリング品質・大規模問題適用性・実際的有用性などについての性能評価を行った．実験には，CPU が 2.7GHz Pentium 4 で RAM が 2GB のパーソナルコンピュータを用いた．





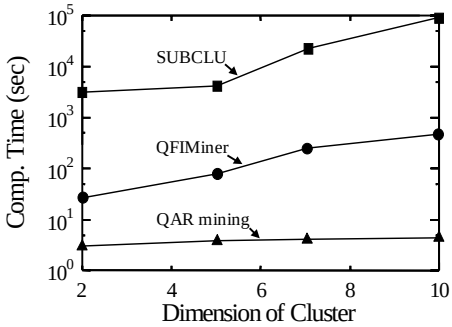


図 4 データセットサイズと計算時間の関係 **図 5** データセットの次元数と計算時間の関係 **図 6** クラスタの次元数と計算時間の関係

4·1 他の手法との比較

QFIMiner を，表 1 に掲げた従来手法の中でも性能や機能が比較的近い SUBCLU [Kailing 04] 及び QARM [Srikant 96] の 2 手法と性能比較した．SUBCLU は求めるクラスタ形状や記号アイテムを扱えない機能面を除くと，QFIMiner に近い原理を使用し似た性能が期待される．更に，クラスタリング品質の点で SUBCLU が従来手法中で最も優れていることが知られている [Kailing 04]．また，QARM は QFIMiner よりクラスタ検出能力の点で不利であると考えられるが，データ数に関する計算量の点では QFIMiner より有利である．SUBCLU ではパラメータとして ϵ(QFIMiner の Δ に相当) と $MinPts$ をとり，これを用いて密度評価を行う．QARM はパラメータとして $minsup$ と最大支持度 ($maxsup$) を必要とする [*1].SUBCLU については，その開発者から直接に実行形式コードを入手した．また，QARM については，その論文 [Srikant 96] に基づいてプログラムを再構築した．実験に用いるデータセットについても SUBCLU と対等な条件で比較するために，その開発者から入手した．このデータは異なる次元組み合わせの複数の部分空間に，いくつかの異なった密度をもつクラスタを埋め込むことで作成された人工データセットである．各クラスタの大きさは部分空間の各軸の範囲 R_p の 0.2〜50%となっており，大量のバックグラウンドノイズが全ての属性空間に一様に加えられている．

QFIMiner と SUBCLU に共通するパラメータとして $\Delta = 2\%$ を用いた．これが大きければ密度評価の統計的ばらつきを減少できるが，逆に広がりが 0.2%のような微小なクラスタを検出できなくなる恐れがある．そこで，想定しているクラスタの広がり 0.2〜50%の中間程度のオーダー分解能とした．また，3 手法に共通するパラメータとして $minsup = 2\%$ を用いた．これは今回対象とするクラスタの殆どのトランザクション数が支持度 2%以上であることと，あまり最小支持度を低く設定するとバックグラウンドノイズをクラスタとして拾ってしまう恐れがあるためである．また，SUBCLU は QFIMiner と異なり $MinPts$ が適応的に自動設定されないため $MinPts = 8$ とした．これは $\Delta = 2\%$ と併せると，今回対象とする最も薄いクラスタの密度を若干下回り，原理的にはこれによってすべてのクラスタを検出可能である．更にこの値が小さすぎるとバックグラウンドノイズをクラスタとして拾ってしまう．また，QARM で用いる $maxsup$ は 20% に設定した．この値は前処理で数値属性を離散化する際に，1 つの離散化ブロックに属するトランザクション数の最大許容値である．これが大きすぎれば離散化の粒度が大きくなり過ぎて小さいクラスタを見逃してしまい，小さすぎれば粒度が細かすすぎて大きなクラスタを細分化して見失う．今回対象とするクラスタは，データ総数の約 1%〜90%程度までの幅広い割合を含み最適値を設定することが困難であるため，試行錯誤の上で最も良い中間程度のオーダー値に設定した．

はじめに，図 4, 図 5 図 6 に示すように，データセットの大きさやデータセットの属性次元数，クラスタが存在する最大の部分空間次元数に関して，計算時間を評価した．特に図 4 は，3·2 節の最後に述べた QFIMiner のデータ数に関する計算複雑性が $O(N \log N)$ であるという予想を検証するための評価である．ここでは，データの属性次元数が 20 次元であり，4 〜 7 次元の部分空間に埋め込まれた 5 つのクラスタを含むデータセットを用いた場合に，データセットの大きさに対する計算時間の依存性を示した．データセットの大きさ N に対する QFIMiner の計算時間の増加は SUBCLU より少なく，QARM に近い．それぞれの曲線は，QFIMiner の $O(N \log N)$，SUBCLU の $O(N^2)$，QARM の $O(N)$ に整合している．このことから QFIMiner の計算複雑性の予想が裏付けられたと考える．図 5 では，3967 個のトランザクションを有し，7 次元の部分空間クラスタを含むデータセットを用いた場合に，データの属性次元数に対する計算時間の依存性を示したものである．この図によると QFIMiner の計算時間の増加は他の手法より大きいが，補題 2 に示される値より $MinPts(\hat{C}^S, p_i)$ を若干大きく取ると大幅に計算時間が減少することが観測された．これは，提案する適応的密度閾値の下で，属性の組み合わせが多くなると，低次元の部分空間においてノイズによる高密度クラスタ候補の数が大きく増えるためである．図 6 は，4000 個のトラ

*1 他に部分的完全性というパラメータを $K = 0.1$ として使用している．しかし，これは QFI の探索にはほとんど影響はない．

ンザクションを有する，15 次元属性のデータセットに関して，クラスタの存在する部分空間次元数と計算時間の関係を示しているが，計算時間はクラスタの次元の増加にはあまり大きな影響は受けないことが分かる．即ち，低次元部分空間で如何に効率的で適切に不要なクラスタの候補を枝狩りするかが，計算スピードの鍵を握っている．全ての図において，QFIMiner の計算速度は，SUBCLU より 1,2 桁速く，QARM より 1,2 桁遅いことが分かる．

表 4 は，3 つの手法について正しく発見することができたクラスタ数を示している．評価に用いた 6 つのデータセットは異なる部分空間に異なる数のクラスタを含んでいる．各クラスタも異なるトランザクション数，異なる大きさを有している．ただし，各クラスタは部分空間上の各属性軸上で等方的な広がりを与えてある．N はデータ中のトランザクション総数，dim. of cluster はクラスタ各々の埋め込み部分空間次元数（QFI に含まれるアイテム数），括弧内の# of transactions はクラスタ各々を構成するトランザクション数，size of clusters はクラスタ各々の等方的広がりの幅で R_p の何%であるかを表す．各手法名称の列数字が正しく検出されたクラスタ個数である．QFIMiner は他の 2 つの手法で探索されたクラスタはもちろん，探索されなかったクラスタも見つけることに成功している．ただし，上から 3 番目のデータの 1 つのクラスタと最後のデータのクラスタの導出には失敗している．これらのクラスタは密度が他に比較して薄いために，データ分布の統計的ばらつきによって QFIMiner の低次元空間探索の段階で枝狩りされてしまったと考えられる．例えば，3 番目データの 50%の広がりを持ちデータ数 256 個からなる 5 次元クラスタの場合，$N = 4469$, $R_p = 100\%$，$C_p = 50\%$, $\Delta = 2\%$ と補題 2 より 1 次元空間での密度閾値 $MinPts(\hat{C}^S, p)$ は約 179 となる．これに対してこのクラスタの 1 次元空間への射影密度の期待値は約 189 ± 13.7(std) であり，統計的ばらつきによって容易に密度閾値を下回ってしまい枝狩りされてしまう．また，最後の 24%の広がりを持ちデータ数 49 個からなる 10 次元クラスタの場合も，$N = 3986$, $C_p = 24\%$ より同じく 1 次元空間での密度閾値 $MinPts(\hat{C}^S, p)$ は約 159 となる．これに対してこのクラスタの 1 次元空間への射影密度期待値は約 166 ± 12.9(std) であり，容易に統計的ばらつきによって枝狩りされてしまう．従って，前節で設けた仮定 1 を満たすクラスタであっても，密度が薄過ぎると統計的ばらつきによって低次元空間探索段階で枝狩りされ導出困難となることがある．しかし，それでもなおかつ QFIMiner の検出性能が SUBCLU より上回っている理由は，各属性軸に対してトランザクションの射影を行うことにより密度の評価を行っているためと考えられる．各軸に射影して一つの軸に $\hat{C}^S$ のトランザクションを集めることで，密度の統計的なエラーを減らすことができ探索の精度があがる．これに対して，SUBCLU のように高次元の部分空間のままでトランザクション間の距離を計算すると，トランザクションの分布が疎になってしまう．それによって，トランザクション間の距離を直接 ϵ を用いて密度評価するとエラーが増えると考えられる．以上を要約すると，QFIMiner の性能は他の手法に対して精度と効率の面について優れていると言える．

表 4 他手法との比較

N	dim. of cluster (# of transactions)	size of clusters (%)	QFIMiner	SUBCLU	QARM
4324	3(711) 5(256) 7(92)	20 20 20	3	2	1
4057	5(256) 5(256) 5(256)	20 20 20	3	3	2
4469	5(256) 5(256) 5(256) 5(256) 5(256)	0.2 2 10 20 50	4	2	1
4045	2(3673)	30	1	1	1
3967	7(1024)	20	1	1	1
3986	10(49)	24	0	0	0

The dimensionality of all data sets is 15.

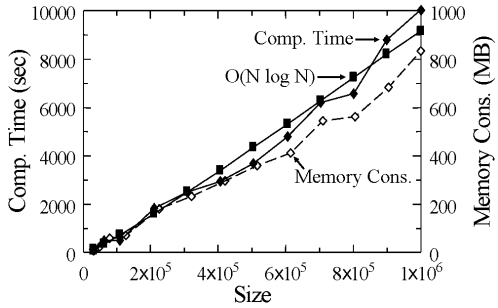


図 7 データセットサイズに対する計算時間とメモリ使用量の関係

4·2 大規模データへの適用性

数値アイテムと記号アイテムからなる大きな人工データを用いて，計算時間とメモリ使用量の観点から，QFIMiner の大規模データへの適用性評価を行った．あらかじめ定義された QFI の種の集合から無作為に QFI を選択し，更にそれに余分にランダムに生成したアイテムを追加して一つのトランザクションを生成し，それを集めてデータ集合とした．直径が各 R_p の 10% の部分空間クラスタとバックグラウンドノイズを生成するために，$\pm 5\%$ の振幅をもつガウス雑音を導入することによって，各トランザクションの数値アイテムの値にばらつきを与えた．生成されたアイテム数は 1000 個で, その半分が数値アイテムである．また，QFI の種の数は 10 で，トランザクションの平均サイズは 24 である．この実験では $\Delta = 5\%$ と $minsup = 5\%$ を使用した．計算時間とメモリ使用量のデータ数 N に対する依存性を 図 7 に表す．$O(NlogN)$ の基準線と比較して，計算時間は 100 万トランザクションに至るまでおよそ $O(N \log N)$ である．これは 3·2 節の最後に述べた QFIMiner の計算複雑性に関する予想と一致する．また, メモリ使用量はおよそ $O(N)$ である．ほとんど $TID - List$ を記憶することにのみにメモリを使用するので, メモリ使用量はデータ数に比例する．これらの結果は，大規模データに関する QFIMiner の使いやすさを示している．

4·3 実データに対するマイニング

QFIMiner が, 対象データをうまく特徴づける分かりやすい相関規則を発掘可能かどうかの検証を行った．データは，UCI ML Repository [U.C.Irvine 04] の adult census というデータセットを用いた．テストに用いたデータセットには 10 万件のデータが含まれており，そのうちの 6180 件が年収が 5 万 USD 以上の人々のデータで，残りが年収 5 万 USD 以下の人々のデータである．比較を行うために，5 万 USD を閾値としてデータを二分割した．$\Delta = 2\%$ 及び最小支持度としては高い $minsup = 50\%$ 用いたが，多くの事例に共通して現れるアイテムが多かったため，2000 個以上の QFI が探索された．これらから考察により，年収の大小で興味深い差異が見られた 2 つの QAR を選択した結果を示す．

>50K, support=69.6%, conf=95.9%

$\{< age : [29,62] >, < birth-country : US >,$

$< veteran : Yes >\} => \{< dividends : [0.00,6500.00] >\}.$

<50K, support=56.2%, conf=99.4%

$\{< age : [0,45] >, < birth-country : US >,$

$< business : No >\} => \{< dividends : [0.00,750.00] >\}.$

一つ目のルールは，5 万 USD 以上の収入がある人々のデータから得られたもので，二つ目のルールは，5 万 USD 以下の収入の人々のデータから得られたものである．同じくらいの支持度と確信度で，後者が若く無職な人が多いのに比べて，前者はより年齢が高く退役軍人である人が多い．また，株式所有者としての配当の上限も前者のほうが高い．次の 2 つのルールは，収入の多いグループから得られたものである．

>50K, support=65.6%, conf=85.5%

$\{< veteran : Yes >, < marital-stat : Married >\} =>$

$\{< weeks-worked : [50,52] >, < dividends : [0.00,6500.00] >\}.$

>50K, support=65.6%, conf=70.6%

$\{< veteran : Yes >, < dividends : [0.00,6500.00] >\} =>$

$\{< weeks-worked : [50,52] >, < marital-stat : Married >\}.$

前者は，既婚の退役軍人は, そのほとんどが労働時間が長く裕福であることを示している．一方，後者は確信度が相対的に低く，裕福な退役軍人であるからといって必ずしも労働時間が長くて結婚しているわけではないことを示している．このように QFIMiner は，大規模データに対してもきめ細かな各数値属性区間値を導出し，かつ解釈容易なルールを導出可能である．

5. 考 察 と 結 論

本研究で提案した QFIMiner の手法上の概要は，以下のようまとめられる．

（1）軸射影密度ベースクラスタリング
データ中の数値属性空間において，各事例近傍の他事例の存在密度を評価し，事例が高密度分布している空間領域をクラスタと見なす密度ベースクラスタリングを用いている．しかも，事例を数値各属性軸に射影して事例存在密度を評価する．

（2）部分空間クラスタリング
データ中の数値属性空間において，事例の高密度部分が存在する部分空間とそこに存在するクラスタを同時に導出する．

（3）軸平行直方体クラスタリング
クラスタ形状把握容易性を確保するため，クラスタを各数値属性軸に平行な直方体領域に限定する．

（4）定量的多頻度アイテム集合マイニング
記号アイテムと数値属性部分空間上の軸平行直方体クラスタの共起を統合して，記号アイテムと区間値を有する数値アイテムからなる多頻度アイテム集合を導出する．

実験で明らかになったように，SUBCLU と異なりクラスタリングに軸射影密度を用いることで，大規模データに関しても実用的平均計算量 $O(N \log N)$ を実現できた．また，軸射影による密度集積効果によって，高次元クラスタについても統計精度を落とさずに検出精度の高いクラスタリングが可能となった．一方，部分空間クラスタリングや記号アイテムを含めた多頻度アイテム集合のマイニングでは，Apriori アルゴリズムや SUBCLUE と同様な幅優先探索を行うため，データ中の属性種類数やクラスタ次元数に対しては，通常のバスケット分析と同様な所要計算量の傾向を示すことが実験的にも確かめられた．更に，軸平行直方体クラスタリングとそれに基づく定量的多頻度アイテム集合の出力結果は，記号アイテムに加えて数値アイテムに関しても区間値が列挙された理解容易な形式で表現され，実験からも対象データの性質を反映する知見を読み取れることが確認できた．

今後，クラスタリング品質及び計算効率を更に改善する可能性として，密度閾値 $MinPts$ の一層の最適化が考えられる．例えば，従来手法の 1 つである SCHISM でも，QFIMiner と同様に全ての部分空間で平均密度を基準にした可変の密度閾値を採用しているが，その際に閾値に Chernoff-Hoeffding の統計的マージンを加えることによって枝狩りを高速化している [Sequeira 04]．SCHISM は CLIQUE と同様にクラスタリングの初めの段階で生成された等幅の格子に基づいて閾値を決めるため，QFIMiner のように各部分空間高密度クラスタ候補毎の柔軟な平均密度評価ができない問題があり，SUBCLU に比べても十分な検出性能は出せない．しかし，QFIMiner にも適切な統計的マージンによる密度閾値のきめ細かい調整を取り入れれば，クラスタ検出能力を維持・向上させつつ，枝狩り効率も向上させることができる可能性があり，手法研究が望まれる．

また，4·3 節に示されたように，QFIMiner を大規模データに適用すると最小支持度などの設定によっては数千を超える多量の QFI が導出される．各々の QFI の可読性は高いが，多量の QFI から如何にして解析者にとって

興味深い共起関係を見つけるかという問題は残る．これはバスケット分析一般にも共通する問題であり，“*closed itemset*” [Pasquier 99] のみを発掘するなど様々な取り組みがなされているが，今後，定量的多頻度アイテム集合の性質を考慮した方法の追求が必要になると思われる．現在，以上のような課題解決に向け研究中である．

◇ 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 dimensional data for data mining applications, *Proc. of the 1998 ACM SIGMOD international conference on Management of data*, pp. 94–105 (1998)

[Cheng 99] Cheng, C.-H., Fu, A. W., and Zhang, Y.: Entropy-based subspace clustering for mining numerical data, *Proc. of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 84–93 (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)

[Fukuda 01] Fukuda, T., Morimoto, Y., Morishita, S., and Tokuyama, T.: Data Mining with Optimized Two-Dimensional Association Rules, *ACM Transactions on Database Systems (TODS)*, Vol. 26, No. 2, pp. 179–213 (2001)

[Goil 99] Goil, S., Nagesh, H., and Choudhary, A.: MAFIA: Efficient and Scalable Subspace Clustering for Very Large Data Sets, *Tech. Report No. CPDC-TR-9906-010, Center for Parallel and Distributed Computing, Dept. of Electrical and Computer Engineering, Northwestern University* (1999)

[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)

[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)

[Pasquier 99] Pasquier, N., Bastide, Y., Taouil, R., and Lakhal, L.: Discovering Frequent Closed Itemsets for Association Rules, *LNCS: Proc. the 7th International Conference on Database Theory (ICDT99)*, Vol. 1540, pp. 398–416 (1999)

[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)

[Rastogi 98] Rastogi, R. and Shim, K.: Mining Optimized Association Rules with Categorical and Numeric Attributes, *Proc. of 14th Int. Conf. on Data Engineering, IEEE Computer Society*, pp. 503–512 (1998)

[Sequeira 04] Sequeira, K. and Zaki, M.: SCHISM: A New Approach for Interesting Subspace Mining, *Proc. of Fourth IEEE International Conference on Data Mining*, pp. 186–193 (2004)

[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)

[U.C.Irvine 04] U.C.Irvine, : *UCI Machine Learning Repository*, University California Irvine, `http://www.ics.uci.edu/~mlearn/MLRepository.html` (2004)

[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)

[Wijsen 98] Wijsen, J. and Meersman, R.: On the Complexity of Mining Quantitative Association Rules, *Data Mining and Knowledge Discovery*, Vol. 2, No. 3, pp. 263–281 (1998)

〔担当委員：小野田 崇〕

2006 年 2 月 13 日　受理

◇ 付　　録 ◇

著 者 紹 介

光永　悠紀(学生会員)

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．現在，日本電気株式会社に勤務．

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．現在，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）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，日本ファジィ学会，情報処理学会，AAAI，IEEE 各会員．

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．現在，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）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．