2A2-2

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

Development of a Classifier by Using QAR Analysis and Application to Mutagenesis Datasets

中西 耕太郎[*1] 鷲尾 隆[*1] 光永 悠紀[*1] 藤本 敦[*1] 元田 浩[*1]
Kotaro Nakanishi Takashi Washio Yuki Mitsunaga Atsushi Fujimoto Hiroshi Motoda

[*1]大阪大学産業科学研究所高次推論方式
Department of Advanced Reasoning, Osaka University. The Institute of Scientific and Industrial Research

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

データマイニングの分野においては，クラスと呼ばれる目的属性を含む大量のデータから，クラスを予測する分類規則を作成するクラス分類問題に関する手法の研究開発が主流の１つになっている．特に近年，"クラスアソシエーションルール (CAR)" に基づいて分類を行う手法の研究が盛んになっている．この規則は "$\{< p_1 : q_1 >, ..., < p_m : q_m >\} \Rightarrow cl$" 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台所有する人は家を所有している" 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 "家を持っている" と予測できる．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事例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 を分類に用いた初期の研究としては，Liu，Hsu，Ma による CAR の support と confidence を用い，最も優先度の高い CAR を分類に使用する CBA[Liu98] がある．この手法では各クラス値を持つ事例数によって CAR の数に差が出来てしまい事例数の少ないクラスについて正確な分類が出来ないという問題がある．この問題を解決するために Dong，Zhang，Wong，Li によって CAEP[Dong99] と呼ばれる手法が開発された．この手法ではあるクラス値についてのみ多頻度でかつ他のクラス値の事例においては相対的に多頻度でない属性集合である Emerging Pattern（EP）を分類に用いる．これらの CAR を分類に用いる手法は，C4.5 のような決定木や決定規則による手法に比べより良い分類精度を示すことが知られている．しかし，これらの分類手法は数値を含むデータを扱う際に何らかの記号離散化前処理を必要とし，数値を含むデータを直接的に解析することができない．この方法では，次に述べるように離散化の際の区間境界が不適切だと，属性やその値同士の共起性を QFI としてうまく捉えられない場合を生じ，それが分類規則の精度向上を阻む可能性がある．また数値属性の離散化については Srikant と Agrawal が「catch-22」と呼ばれる２つの問題を述べた．[Agrawal96] １つは，数値データの離散化幅が狭すぎると１つ１つの区間に含まれる要素数が少なく，支持度が低くなってしまう「最小支持度問題」である．そのため，最小支持度を下回らないようにある程度以上の粒度で離散化する必要がある．もう１つは，離散化幅が広すぎると支持度は増加するが属性やその値間の共起性を捉える粒度が粗くなりすぎて，確信度が低くなってしまう「最小確信度問題」である．そのために最小確信度を下回らないように離散化幅を狭くする必要がある．CAR を用いる分類手法は高性能の分類規則を作り出す可能性を秘めているが，前述の欠点によって広範囲の適用が阻まれている．この問題を解決するために多くの研究が行われてきたが，何れも数値を含む事例データに関する実用的かつ最適な CAR を現実的時間で求めることができない．そこで，我々は現実的な時間計算量で属性やその値の間の共起が多頻度になる各数値区間とそれらの QFI を同時に完全探索する QAR mining by Monotonic INTerval（QARMINT）[Washio04] という手法を開発した．我々は QARMINT に前述の CAEP と呼ばれる CAR に基づく分類手法を適用することにより，数値属性を含むデータに関

連絡先: 大阪大学産業科学研究所
〒 567–0047 大阪府茨木市美穂ヶ丘 8-1
E-mail: nakanishi@ar.sanken.osaka-u.ac.jp

して新たな CAR に基づく分類手法（CAQEP）[中西 05] を開発した．この手法は前述の CAR を分類に用いる手法の利点に加え，数値属性を含むデータから共起性を捉える数値属性離散化区間を直接的に導出して CAR に反映することにより，従来の分類手法には無い種々利点を有する．しかしながら，空間計算量，時間複雑性の問題から，広範囲への適用が阻まれていた．そこで，当報では，我々によって開発された前述の欠点を克服した部分空間クラスタリング手法 QFIMiner[Washio05A] に CAEP を適用することにより，より広範囲への適用を可能とする CAR に基づいた分類手法 Levelwise Subspace Clustering - CAEP(LSC-CAEP)[Washio05B] を開発し，その性能評価を行った．

2. CAEP

本節では CAR に基づく分類手法 CAEP について述べる．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．ここで，アイテム集合 a のデータ D_{cl} による支持度を $support_{D_{cl}}(a) = |\{t \in D_{cl} | a \subseteq t\}|/|D_{cl}|$ とする．また，各クラス cl について，$support_{D_{cl}}(a) \geq minsup$ を満たす QFI である a の集まりを $LQFI(cl)$ とする．オリジナルの CAEP では，前述の CBA と同様にして QFI を求めるが，LSC-CAEP では後述するように部分空間クラスタリングを用いる．そして，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．

Growth Rate

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 $growth\_rate_{\bar{D}_{cl} \to D_{cl}}(a) = 0$.

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 CAR 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，CAR 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 CAR 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．

次の段階では，後述する CAR による投票を行う際の重みを計算するために，各クラス cl に関する “*base score*” を計算する．まず，CAR の条件部 a の分類能力を $support_{D_{cl}}(a)/(support_{D_{cl}}(a) + support_{\bar{D}_{cl}}(a))=$ $growth\_rate_{\bar{D}_{cl} \to D_{cl}}(a)/\ \ (growth\_rate_{\bar{D}_{cl} \to D_{cl}}(a) + 1)$ と定義する．これは規則の分類力が $support_{D_{cl}}(a)$ 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．ここで $LRB(cl)$ を前段階で述べた *Growth Rate* で $LQFI(cl)$ から選んだ全ての CAR 条件部の集合とする．その時，$LRB(cl)$ に含まれる全ての CAR 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 “*aggregate score*” を定義する．

Aggregate score:

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

$LRB(cl)$ に含まれる CAR 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スコアを持ってしまう．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* を計算する．

Base score:

$base\_score(cl)$ is the aggregate score where the number of t having their aggregate scores less than this score is $Tail\%$ of all t in $\{score(t, cl) | t \in D_{cl}\}$.

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．

Normalized score:

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

これにより事例 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. 部分空間クラスタリング手法 QFIMiner

部分空間クラスタリング手法 QFIMiner は QFI の候補を作成する際，apriori アルゴリズムを拡張したアルゴリズムを用いる．apriori アルゴリズムは，要素数 1 の多頻度アイテム集合からはじめて，ボトムアップ的に要素数 k の QFI から要素数 $k+1$ の QFI の候補を作り出すことを行う．具体的に述べると $k-1$ 個の要素が共通な要素数 k の 2 つの QFI

$$X_k = \{p_1, p_2, \cdots, p_{k-1}, p_k\}$$
$$Y_k = \{p_1, p_2, \cdots, p_{k-1}, p'_k\}$$

より要素数 k+1 の QFI の候補を上記の 2 つの QFI の和集合

$$\begin{aligned} Z_{\mathrm{k+1}} &= X_k \cup Y_k \\ &= \{p_1, p_2, \cdots, p_{k-1}, p_k, p'_k\} \end{aligned}$$

とする．さらに候補を絞り込むために，このようにして候補にあがった要素数のアイテム集合に対して，要素数 k のアイテム集合からなる各部分集合が全て QFI として存在しているかどうかをチェックする．これは QFI であるためには，その部分集合は全て QFI でなければならないためである．要素数の QFI は前段階の処理で全て取り出されており既知であるので，このチェックは効率的に実行することが可能である．このようにして残った各候補についてのみ，実際の支持度をデータから計算して，*minsup* 以上の支持度を有する QFI を効率的に求めることができる．ここで，多頻度領域の性質を用いることにより，QFIMiner はより効率的に候補生成を実現する．通常では，候補を作成した後に要素数の一つ少ないアイテム集合が全て QFI であるかをチェックする．しかし，数値属性アイテムを含む QFI の候補を作成する場合を考えると，X_k , Y_k に共通する数値アイテムそれぞれについて，そのアイテムを属性とする数値軸上に X_k 及び Y_k の多頻度領域をそれぞれ正射影した区間が事例を含まない場合，X_k , Y_k を結合してできる多頻度領域は存在しないという性質がある．よって，この性質を用いて QFI の候補生成に対して枝狩りを行うこと

表 1: 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 LSC-CAEP
Australian	690	14(6)	2	.8608	.8538	**.8666**	.0347
Cars	392	7(6)	3	.9617	.9744	**1.0000**	0
Cleve	303	13(5)	2	.7656	.8283	**.8383**	.0422
Crx	690	15(6)	2	.8608	.8538	**.8715**	.0442
Diabetes	768	8(8)	2	.7226	**.7445**	.7229	.0681
Ecoli	336	8(7)	8	**.8422**	.7018	.7794	.0992
German	1000	20(7)	2	.7070	**.7350**	.7173	.0517
Heart	270	13(6)	2	.7666	8187	**.8222**	.0694
Hepatitis	155	19(6)	2	**.8387**	.8182	.8236	.1062
Horse	368	22(8)	2	.6933	.8236	**.8394**	.0488
Hypo	3163	25(7)	2	**.9889**	.9826	.9793	.0071
Iris	150	4(4)	3	.9600	.9467	**.9733**	.0466
Labor	57	16(8)	2	.7368	.8633	**.9500**	.1124
Led7	3200	7(0)	10	.7337	.7206	**.7400**	.0117
Lymph	148	18(2)	4	.7635		**.8157**	.1189
Nursery	12960	8(0)	5	**.9705**	.8289	.9408	.0048
Pima	768	8(8)	2	**.7382**	.7290	.7141	.0338
Sonar	208	60(60)	2	**.7884**	.7746	.6681	.1288
Tae	151	5(1)	3	**.5099**	.4717	.5067	.1470
Tic-Toc-Toe	958	9(0)	2	.8507	.9959	**1.0000**	0
waveform	5000	21(21)	3	.7664	**.7968**	.7886	.0153
Wine	178	13(13)	3	.9382	.9496	**.9833**	.0374
Zoo	101	16(0)	7	.9207	**.9709**	.9309	.0477
Average				.8123	.8264	**.8379**	.0555

により計算時間を短縮することができる．こうして作成された QFI の候補に対して，全てのアイテムを記号アイテムとした場合に X_k，Y_k を両方含む事例数をカウントしこれが最小支持度を超えない場合は更なる枝狩りを行う．このとき QFI の候補として残った各属性集合が数値属性を含んでいた場合，その数値属性群に対して，密集度基準 MinPts と許容距離 *Delta* を用い，QFI の候補における数値属性値の密集領域を導出する．このとき，QFI の候補を支持する事例数が最小事支持度以上であれば，その QFI の候補を QFI とする．このアルゴリズムによって，計算複雑性 $O(NlogN)$，空間計算量 $O(N)$ で，で QFI を導出することが可能である．詳細については，[Washio05A, Washio05B] を参考されたい．

4. 性能評価

4.1 分類手法の性能評価

表 1 に示された UCI repository[UCirvine] に公開されているデータを用いて C4.5，CBA，LSC-CAEP の性能比較実験を行った．性能比較実験では，CPU が Pentium4 2.7GHz，メモリが 2GB の計算機を用いた．尚，これらの分類精度は 10 fold cross validation で評価した．

表 1 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 1 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 1 の右端に提案手法の 10CV 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 1 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 1 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 1 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．更に，提案手法は表 1 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．個々の手法が対等な精度であるならば，各データについて提案手法が最高精度になる確率は 1/3 であるので，2 項分布より 23 個のデータ中 12 ケースで最高精度になる確率は ${}_{23}C_{12}(1/3)^{12}(2/3)^{11} = 2.9\%$ である．これらのことから，C4.5, CBA と比べ有意に LSC-CAEP が優れているといえる．

4.2 CAR の理解容易性

データマイニングの手法を開発する上では，データに埋もれた知識を発見するために，人間にとって理解しやすい出力結果を得ることが重要である．但し，結果の理解し易さは主観的な問題である．そこでここでは，LSC-CAEP の出力表現の長所を示す単純かつ明確な解析結果例を示すに留める．LSC-CAEP を UCI repository に公開されている Iris データに対して適用した．その結果，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 の高い相関規則であれば理解しやすい．従来の分類手法ではこのように単純で理解しやすい解析結果を得ることは難しい．

4.3 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．

表 2: クラスタリングの計算複雑性

Para-meter	Range of Assessment	Dependency of Comp.Time	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.

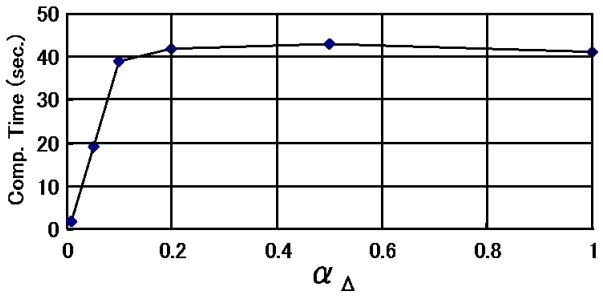

図 1: α_Δ と計算時間の関係

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 2 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．これは標準的な AprioriTid アルゴリズムとほぼ同等である．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．数値アイテムと記号アイテムに関して，基本的な処理に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 1 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 2 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．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．

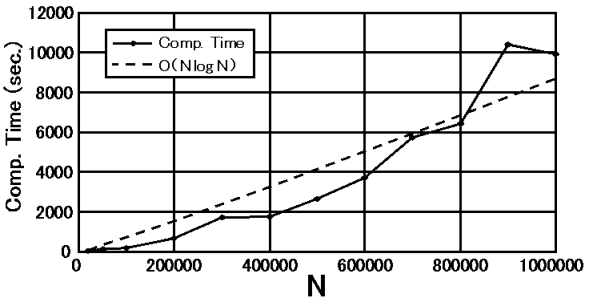

図 2: データ数 N に対する計算時間の依存性

5. むすび

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，QFIMiner が導出する QFI の品質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人間にとって理解容易な超直方体の QFI が導かれる．しかも，上記に記された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的確に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

[Liu98] B.Liu, W.Hsu, Y. Ma, Integrating Classification and Association Rule Mining. SIGKDD, pp, 80-86 (1998).

[Dong99] G.Dong, X.Zhang, L.Wong, J.Li: CAEP:Classification by Aggregating Emerging Patterns, Int'l Confidence on Discovery Science, pp, 30-42 (1999).

[中西 05] 中西 耕太郎，鷲尾 隆，藤本 敦，元田 浩: 定量的相関規則を用いたクラス分類手法の開発，第 69 回知識ベースシステム研究会，pp，143-150 (2005).

[Agrawal96] R.Srikant and R.Agrawal, Mining Quantitative Association Rules in Large Relational Tables, Proc. of the 1996 ACM SIGMOD International Conference on Management od Data, pp. 1-12 (1996).

[Washio04] T.Washio, A.Fujimoto and H.Motoda, Extention of Basket Analysis and Quantitative Association Rule Mining, 人工知能学会　知識ベースシステム研究会（第 67 回）SIG-KBS-A403, pp. 117-122 (2004).

[Washio05A] Takashi Washio, Yuki Mitsunaga, Hiroshi Motoda, "Mining Quantitative Frequent Itemsets Using Adaptive Density-Based Subspace Clustering," Proc. of Fifth IEEE International Conference on Data Mining (ICDM'05), pp.793-796 (2005).

[Washio05B] Takashi Washio, Koutarou Nakanishi, Hiroshi Motoda, "Deriving Class Association Rules Based on Levelwise Subspace Clustering ," Proc. of PKDD2005: 9th European Conference on Principles and Practice of Knowledge Discovery in Databases, LNAI 3721, pp.692-700 (2005).

[UCirvine] http://www.ics.uci.edu/˜mlearn/MLRepository .html, University of California Irvine Machine Learning Repository.