1F1-02

定量的相関規則クラス分類手法とその変異原性データへの適応

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

中西 耕太郎[*1] 鷲尾 隆[*1] 藤本 敦[*1] 元田 浩[*1] 岡田 孝[*2]
Koutarou Nakanishi Takashi Washio Atsushi Fujimoto Hiroshi Motoda Takashi Okada

[*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

[*2]関西学院大学理工学部
Department of Informatics, Kwansei Gakuin University

The present association-rule-based classifier needs pre-discretization of a dataset that includes numeric attributes. In this paper, we propose a new approach based on an association-rule-based classification method called CAEP[Dong99] and QAR mining by Monotonic INTerval[Washio04] developed in our previous work. It can directly and completely mine significant combinations of attributes for classification from datasets consisting of many categorical and numeric attributes and derive a classification rule set based on the combinations. The result of our classification shows better performance than the results of conventional approaches. The application of the method to mutagenesis datasets has been made to get knowledge that is benefit for experts in the domain of quantitave structure-activity relationship(QSAR) analysis.

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，相関規則集合に基づいて分類を行う手法の研究が盛んになっている．従来の分類手法はアイテム間の相関関係を用い，高い精度の分類規則を作成できる．その初期の研究としては Liu，Hsu，Ma による相関規則の support と confidence を用い最も優先度の高い相関規則を分類に使用する 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相関規則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．この他にも Apriori アルゴリズムに比べ，大量のデータや属性数の多いデータに対してより短い時間で相関規則を導出できる FP-growth と呼ばれるアルゴリズムを用いた CMAR[Li01] といった手法も開発されている．これらの相関関係を分類に用いる手法の主要な利点には次が挙げられる，

1. 多数の属性を含むデータから分類に有用な属性の組み合わせを自動的に完全導出できること．
2. 少数例のクラスに対する分類規則を網羅でき分類精度を落とさないこと．
3. パラメータの設定により過学習を起こり難くできること．

これらの理由で相関関係を分類に用いる手法は，C4.5 をはじめとする従来の手法に比べ高い精度の分類を達成できる可能性がある．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多頻度アイテム集合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．相関関係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．また数値属性の離散化については Srikant と Agrawal が「catch-22」[Agrawal96] と呼ばれる 2 つの問題を述べた．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．

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，何れも数値を含むトランザクションデータに関する実用的な最適相関ルールを現実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多頻度アイテム集合を同時に完全探索する QAR mining by Monotonic INTerval（QARMINT）[Washio04] という手法を提案した．

我々は QARMINT に前述の CAEP と呼ばれる相関関係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する新たな相関規則に基づく分類手法（CAQEP）[中西 05] を開発した．この手法は前述の相関関係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分類規則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．当報では CAQEP を変異原性データに適応し，生成された分類規則より専門家にとって有益な知識を導出することを目的とする．

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

2. CAQEP

2.1 多頻度アイテム集合と多頻度領域

本節では QARMINT における多頻度アイテム集合と多頻度領域について述べる．あるデータベース D において，最小支持度を超える頻度で各トランザクションに共起するアイテム集合を「多頻度アイテム集合（frequent itemset）」と呼ぶ．上記 Apriori アルゴリズムを数値アイテムの数値情報を含めて多頻度なアイテム集合を探索できるように拡張する．ここで，記号アイテムは記号属性 A_{s_i} とその値 v_{s_i} のペア $\langle A_{s_i}, v_{s_i}\rangle$，数値アイテムは同様に，$\langle A_{n_i}, v_{n_i}\rangle$ と表されるものとし，従ってトランザクション $t = \{\langle A_{s_1}, v_{s_1}\rangle, ..., \langle A_{s_{n_s}}, v_{s_{n_s}}\rangle, \langle A_{n_1}, v_{n_1}\rangle, ..., \langle A_{n_{mn}}, v_{n_{mn}}\rangle\}$ となる．次に，各数値アイテムの値 v_{n_i} を無視し $t^\dagger = \{\langle A_{s_1}, v_{s_1}\rangle, ..., \langle A_{s_{n_s}}, v_{s_{n_s}}\rangle, \langle A_{n_1}\rangle, ..., \langle A_{n_{m_n}}\rangle\}$ とする．このような変換を施したデータに対して最小支持度を越える多頻度アイテム集合 f を探索する．次に，得られた f に含まれる全ての数値属性 A_{n_i} に注目し，f を含む各トランザクション t の持つ数値属性の値 v_{n_i} が密集している区間を抽出する．つまり，ある密集基準を満たし

$$\frac{|\{t|\bigwedge_{\forall A_{n_i}\in f} l_i \le v_{n_i} \le u_i, f \subseteq t\}|}{|D|} \ge minsup$$

を満足する領域を抽出すればよい．ただし $l_i < u_i$，$A_{n_i} \in f$ である．これによって得られた範囲を本稿では「多頻度領域（frequent region）」と呼ぶ．

多頻度領域を探索するためには，トランザクションが密集していると判断する基準が必要である．そこで本稿では，各数値属性 A_{n_i} ごとに「それぞれのトランザクション t の持つ数値属性の値 v_{n_i} 同士が特定の値 Δ_i より離れていなければ密集しているとみなす」という基準を設定する．この Δ_i を数値属性 A_{n_i} に関する「許容距離（permissible range）」と呼ぶ．そして，多頻度アイテム集合に含まれる全ての数値属性に対して許容距離 Δ_i 以内であるトランザクション同士を全て結合する．即ち，トランザクション t の近傍集合 $E_f(t)$ を

$$E_f(t) = \{t' \in D| \wedge_{\forall A_{n_i}\in f} |v'_{n_i} - v_{n_i}| \le \Delta_i\} \cup \{t\}$$

として，すべての t について求まった近傍集合 $E_f(t)$ の集合を ES_f とする．但し，$\langle A_{n_i}, v_{n_i}\rangle \in t$ である．ここで，任意の $E_f(t_i)$，$E_f(t_j)$（$\in ES_f$）の積集合が空集合でないとき，「$E_f(t_i)$ と $E_f(t_j)$ は直接結合している（directly connected）」と呼ぶ．また ES_f について，任意の $E_f(t_i)$，$E_f(t_j)$ に対して

$$E_f(t_{1(i\to j)}), E_f(t_{2(i\to j)}), ..., E_f(t_{m(i\to j)})$$

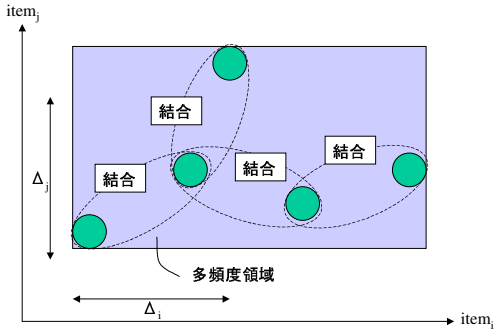
が存在し，

$$E_f(t_i), E_f(t_{1(i\to j)}), ..., E_f(t_{m(i\to j)}), E_f(t_j)$$

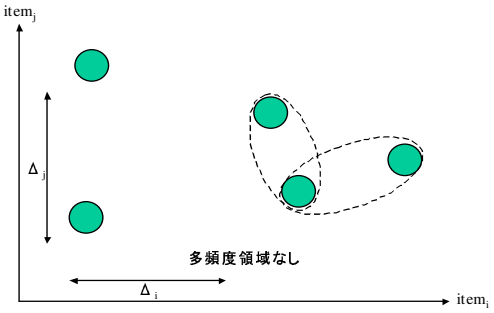
がこの順序で直接結合している場合，「$E_f(t_i)$ と $E_f(t_j)$ は結合している（connected）」と呼ぶ．この近傍集合の集合 ES_f をもとにして得られる極大な多頻度集合の候補 C_f は

$$C_f = \{E_f(t) \in ES_f | E_f(t) s\, are\, mutually\, connected\, in\, ES_f\}$$

となる．このような C_f は ES_f から複数得られる．これによって生成された多頻度集合の候補 C_f に含まれる要素数が最小支持度以上であるとき，それを多頻度集合 F_f と呼び，F_f における各数値属性 A_{n_i} の最大値と最小値で囲まれた範囲が多頻度領域 Id_f となる．



（1）密集している場合



（2）密集していない場合

図 1: $item_i, item_j$ における多頻度領域

$$F_f = \{C_f | \frac{|C_f|}{|D|} \ge minsup\}$$

$$Id_f = \{\, [\underset{t\in F_f}{min}(v_{n_i}), \underset{t\in F_f}{max}(v_{n_i})] \,|\, A_{n_i} \in f\}$$

例として 2 種類の数値アイテム $item_i, item_j$ の多頻度領域を求めると図 1 のようになる．ただし，最小支持度は 4 とする．図におけるそれぞれの点をトランザクションとすると，（1）では各数値アイテムに対して許容距離 Δ 以下であるトランザクションを結合してゆけば多頻度集合の候補の要素数は 5 となる．よって，最小支持度を上回るので図のような長方形の多頻度領域が得られる．また，（2）においてはトランザクションの結合はなされるが，多頻度集合の候補の要素数が 3 であり，最小支持度を下回るために多頻度領域は得られない．

2.2 多頻度領域抽出アルゴリズム

本節では QARMINT によって多頻度領域を抽出するアルゴリズムについて述べる．まず属性数が 1 の多頻度アイテム集合 $F_{f(1)}$ の多頻度領域抽出法を述べる．多頻度アイテム集合が記号のみからなる場合，そのアイテムに関してはこの処理はしなくてよい．数値アイテムである場合は，それぞれの数値属性 A_{n_i} に対応する許容距離 Δ_i を用いて近傍集合を生成し，多頻度領域を抽出する．この多頻度アイテム集合における多頻度領域を抽出するために最も計算時間を要する部分はクイックソートであり，計算時間は $N = |D|$ とした時に最悪 $O(N \log N)$ である．

属性数が 2 つ以上の多頻度アイテム集合 $F_{f(l)}(l \ge 2)$ の多頻度領域を探索するには，単一属性のみに密なだけでなく関連する全ての数値属性に関して値が密な必要がある．このとき支持度の単調性より，$F_{f(l)}$ に含まれる数値アイテムからなる任意の集合のうち 1 つでも多頻度領域を持たないものがあれば，その $F_{f(l)}$ は多頻度領域を持たない．よって，$F_{f(l)}$ に含まれ

る全ての数値属性が張る空間上で前述の C_f を直接求めるよりも，$l=1$ から始めてボトムアップ的に最小支持度未満の多頻度領域の候補を切り捨てながら多頻度領域を生成した方が効率が良い．そこで，我々は数値アイテム集合に関して Apriori アルゴリズムで多頻度領域を生成する手法を用いる．

まず，ある要素数 l の多頻度アイテム集合 $F_{f(l)}$ に含まれる属性の集合を $f(l)$ とすると，$f(l)$ に含まれる全ての属性 A_{n_i} について Δ_i 以内の間隔に密集した 1 次元（属性数 1）の多頻度領域を求める．これによって生成された各数値属性ごとの多頻度領域をもとに，以下のように複数の数値アイテムを組み合わせてゆき，l 次元（属性数 l）の多頻度領域を生成する．

まず，l 次元の $F_{f(l)}$ を生成するには，$l-2$ 次元の要素が共通な $l-1$ 次元の 2 つの多頻度集合 $F_{f(l-1)}$,$F_{f'(l-1)}$ ($|f(l-1)\cap f'(l-1)|=l-2$) を用い，$f(l)=\{f(l-1)\cup f'(l-1)\}$ として $A_i\in f(l)$ である全ての数値属性 A_{n_i} について一次元の多頻度領域を求める．ここで，$f(l)$ 内の各数値アイテム A_{n_i} の多頻度領域 $Id_{\{A_{n_i}\}}$ は前述の通り複数存在しえるので，その集合を $\{Id_{\{A_{n_i}\}}\}$ とする．そして，全ての $A_{n_i}\in f(l)$ に関する $\{Id_{\{A_{n_i}\}}\}$ の直積

$$PId_{f(l)}=\underset{A_{n_i}\in f(l)}{\otimes}\{Id_{\{A_{n_i}\}}\}$$

をとりその各要素に番号 $m_{f(l)}(=1,...,|PId_{f(l)}|)$ をふる．更に，$PId_{f(l)}$ の各要素は $f(l)$ に関する多頻度領域である可能性を有するのでそれらを多頻度集合の候補 $I_{f(l)m_{f(l)}}$ とする．このとき，$I_{f(l)m_{f(l)}}$ の要素数が最小支持度未満のものは切り捨てる．こうして生成された $I_{f(l)m_{f(l)}}$ は多頻度集合であるが，1 次元増えたため内部に含まれるトランザクションの分布が疎になる場合がある．そのために多頻度集合かつ密集している部分を探索する．密集した部分を探すためには，2.1 節と同様に $I_{f(l)m_{f(l)}}$ において $f(l)$ に含まれる各数値アイテム A_{n_i} に関して許容距離 Δ_i 以下であるトランザクション t の近傍集合 $E_{f(l)m_{f(l)}}(t)$ すなわち

$$E_{f(l)m_{f(l)}}(t) = \{t'\in I_{f(l)m_{f(l)}}|\wedge_{\forall A_{n_i}\in f(l)}|v'_{n_i}-v_{n_i}|\le\Delta_i\}\}$$

を求めればよい．これを $I_{f(l)m_{f(l)}}$ に含まれる全てのトランザクションに対して行い，近傍集合の集合を $ES_{f(l)m_{f(l)}}$ とする．この近傍集合の集合 $ES_{f(l)m_{f(l)}}$ をもとにして得られる多頻度集合の候補 $C_{f(l)m_{f(l)}}$ は

$$C_{f(l)m_{f(l)}}=\{E_{f(l)m_{f(l)}}\in ES_{f(l)m_{f(l)}}| \\ E_{f(l)m_{f(l)}}\,s\,are\,mutually\,connected\,in\,ES_{f(l)m_{f(l)}}\}$$

となる．このような $C_{f(l)m_{f(l)}}$ は $ES_{f(l)m_{f(l)}}$ から複数得られる．これによって生成された多頻度集合の候補 $C_{f(l)m_{f(l)}}$ に含まれる要素数が最小支持度以上であるとき，それらの各 $C_{f(l)m_{f(l)}}$ における各数値属性 A_{n_i} の最大値と最小値で囲まれた超直方体の範囲を多頻度領域 Id_f とする多頻度領域 $F_{f(l)}$ が導出される．上記を，l を 1 つずつ増やしながら反復する．

このアルゴリズムで最も時間計算量を要するところは近傍集合を求める部分である．各近傍集合を求めるために各トランザクションの数値アイテムの値を比較するため，トランザクション数を N とすると，時間計算量は最悪 $O(N^2)$ である．しかし実際は全トランザクションが同じ数値アイテム集合を含み，かつ密集しているということは少なく，$O(N^2)$ より少なくなる．

2.3 CAEP

CAQEP では，QARMINT によって各分類対象クラス毎に得られた多頻度集合について，CAEP と呼ばれる手法を用いて分類規則を生成する．まず得られた多頻度アイテム集合について，各クラスを持つデータの中で多頻度アイテム集合を含むデータの割合（support）を求める．あるクラス C_i に属するデータ t の集合を D_i としてアイテム集合 e についての support を以下の式で計算する．

$$sup_{C_i}(e)=\frac{|\{t\in D_i\ e\subseteq t\}|}{|D_i|}$$

今，クラス C_i について C_i 以外のクラスラベルを持つデータに $\neg C_i$ という新たなクラスを何与する．次にアイテム集合 e について対象クラス C_i がクラス $\neg C_i$ に対して相対的に多頻度であることを示す指標である growth_rate を以下の式で計算する．

$$growth\_rate(e)= \begin{cases} \frac{sup_{C_i}(e)}{sup_{\neg C_i}(e)} & (sup_{\neg C_i}(e)\neq 0) \\ \infty & (sup_{\neg C_i}(e)=0,\quad sup_{C_i}(e)\neq 0) \end{cases}$$

上記 C_i と e について growth_rate と support が閾値を上回った場合，e を C_i を導く EP（Emerging Pattern）と呼ぶ．尚，growth_rate の閾値は 1 より大きい値とする．e を含むトランザクションは C_i である可能性が高いので $e\Rightarrow C_i$ という相関規則が導出される．下記の式によってこのような相関規則の確信を表す指標である aggregate_score を求める．尚，growth_rate の閾値は 1 より大きい値とする．

$$aggregate\_score(e)=\frac{growth\_rate(e)}{growth\_rate(e)+1}*sup_{C_i}(e)$$

各分類対象クラス C_i に対応する EP の集合を $E(C_i)$ とし，C_i に存在する全てのクラスが既知である t について $E(C_i)$ に属し t に含まれる EP の aggregate_score の総和を t の C_i に対する score とする．その定義は下の通りである．

$$score(t,C_i)=\underset{e\subseteq t,e\in E(C_i)}{\Sigma}aggregate\_score(e)$$

各 C_i 毎に score 値のスケールが異なるので，score の規格化のために C_i の EP を持つトランザクションの score のメディアンを，その C_i の base_score とし，score の正規化に用いる．

こうして得られた EP と base_score を用いてクラスを分類する．クラスが未知のトランザクション t が各分類対象クラス C_i を持つ確信を表す指標として，以下のようにその t に含まれる EP の score の総和を base_score で割ったものを normalize_score とする．

$$normalize\_score(C_i)=\frac{ratio\_score(C_i)}{base\_score(C_i)}$$

全てのクラスについて normalize_score を求め normalize_score が最大のクラスにトランザクションを分類する．

3. 性能評価実験

本報で用いる CAQEP を，UC Irvine Machine Learning Repository に公開されているデータを用いて 10 fold cross validation を実行し，従来のクラス分類手法と性能比較した．その結果を下の表 1 に示す. 提案手法は実験の結果，Pima を除くデータについて従来の手法を超える分類精度を達成することに成功した．

表 1: 本手法と他手法との性能比較

データ名	データ数	属性数 (数値属性数)	クラス数	NB	C4.5	TAN	CBA	LB	CAEP	CAQEP
Pima	768	8(8)	2	**.759**	.711	.7577	.7303	.7577	.75	.669
Heart	270	13(6)	2	.8222	.7669	.8333	.8187	.8222	.8370	**.8444**
Iris	150	4(4)	3	-	.953	-	.929	-	.9467	**.9667**
Wine	178	13(13)	3	-	.927	-	.916	-	.9711	**.9722**
Hepatitis	155	19(6)	2	.8392	.8	-	.802	.845	.8303	**.852**
Ecoli	336	8(7)	8	-	.824	-	-	-	-	**.831**

表 2: 使用する属性

descriptor	meaning
diameter	The largest entry of A
radius	If r_i is the largest entry in row i of A, then the radius is defined as the smallest value of r_i
PEOE_PC+	Sum of the positive p_i
PEOE_PC-	Sum of the negative p_i
PEOE_RPC+	Magnitude of the largest positive p_i divided by the sum of the positive p_i
PEOE_RPC-	Magnitude of the smallest negative p_i divided by the sum of the negative p_i
vdw_area	The area of the van der Waals surface
vdw_vol	Van der Waals volume
logP(o/w)	Log of the octanol/water partition coefficient

表 3: クラス値とその領域

class	region
inactive	activity = -99
low	-99 < activity < 0
medium	0 ≤ activity < 3
high	3 ≤ activity

4. 変異原性データからの知識獲得

これまでに変異原性と発癌性は高い相関関係を持っていることがわかっている．しかしながら，変異原性データの数が多いため，専門家の考察によって変異原性を誘発する化学的条件を網羅的に抽出することが困難であった．そのため網羅的なデータの評価が望まれる．提案手法を変異原性データに関する HP[Mutagenicity00] に公開されている MOE データに対して適応し，その結果得られた相関規則より専門家にとって有益な知識を得るという実験を行った．MOE データはデータ数が 230，属性数が 102，それら全てが数値属性であるが実験に使用した計算機のメモリの制約上，専門家から意見を聞き表 2 に記された重要属性に限定した．また提案手法は数値属性を直接的にクラス属性として取り扱うことができないので，専門家の意見に従い activity という数値属性を離散化し，クラスとした．各クラス値の activity 領域を表 3 に示す．提案手法によって得られたクラス inactive に関する相関規則の例を下に示す．

例 1:
logP(o/w):[1.69 - 2.63]⇒class:inactive
[support=0.63,growth_rate=4.01,aggregate_score=0.50]

例 2:
PEOE_PC+:[0.65 - 1.11], PEOE_PC-:[-1.11 - -0.65] ⇒ class:inactive
[support=0.72, growth_rate=1.46, aggregate_score=0.432]

これらの相関規則に加え，クラス inactive においていくつかの aggreegate_score の値が高い相関規則について，専門家から『クラス inactive に関しては Log，vdw_area，vdw_vol，radius，PC の相関が高い．即ち，疎水性であり，分子のサイズが小さく，電荷の分離も小さめである場合に変異原性が低いという示唆が得られる.』という知見が述べられた．これより提案手法によって，専門家にとって有益な知識を導出することに成功したといえる．

5. むすび

本研究では数値アイテムを直接的に扱える相関関係を用いた分類手法をプログラム実装し，UC Irvine Machine Learning Repository に公開されているデータを用いて性能評価実験を行った．その結果，幾つかのデータによって従来の分類手法を越える分類精度を達成した．また，提案手法を変異原性データに適応し，得られた相関規則から専門家にとって有益な情報を抽出することに成功した．今後取り組むべき課題としては，使用メモリの低減や計算の高速化が挙げられる．

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).

[Li01] W.Li, J.Han, J.Pei, CMAR:Accurate and Efficient Classification Based on Multiple Class-Association Rules, IEEE Int'l Conf. on Data Mining, pp, 369-376 (2001)

[Mutagenicity00] http://www.clab.kwansei.ac.jp/mining/datasets/PAKDD2000/okd.htm, Mutagenicity data descption.

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