対称比率規則の抽出手法

Mining Symmetric Ratio Rules

濱本 雅史♥　北川 博之♦

Masafumi HAMAMOTO　Hiroyuki KITAGAWA

数値属性間で成り立つ線形関係を表した比率規則は，データの理解補助，欠損値の補完などの幅広い有用な応用が可能である．われわれの先行研究である，２属性の数値データに対する比率規則には，相関ルールマイニングと対応づけられたサポートと確信度の概念が導入されている．しかしその定義において２属性は非対称に扱われるため，属性の役割を入れ替えて抽出された比率規則は一致せず，それらが表す線形関係も異なりうる．本論文では２属性を対称に扱うような比率規則である，対称比率規則の定義を行い，その抽出手法を提案する．この提案手法について実データを用いた実験を行い，提案手法の妥当性を示す．

Ratio Rules represent linear relationships among numeric attributes, and are applicable to data understanding support, filling in missing values, and related issues. Our previous work introduced some concepts such as support and confidence, which are used in association rule mining, to Ratio Rules for two attributes. However, extracted Ratio Rules or linear relationships represented by them do not coincide when the attributes are swapped, because two mining target attributes are dealt asymmetrically in the definition. In this paper we define the Symmetric Ratio Rules, which deal target attributes symmetrically, and propose a method to extract them. We show appropriateness of the proposed method using real data.

1. はじめに

近年，大量のデータから重要な情報を抽出するデータマイニング手法として様々なものが検討されている．これらは扱うデータの種類や目的により，例えば相関規則マイニング，クラスタリング，分類，テキストマイニング，時系列マイニング，Webマイニングなどといった手法が提案されている[1]．本研究では，数値属性を持つデータより線形関係を表した比率規則[2]を抽出する問題を考える．比率規則は属性間における，属性値の典型的な割合を表したものである．比率規則は[2]で示されているように，単にデータを理解する補助になるだけでなく，欠損値の埋め合わせ，予測，外れ値検出，可視化など様々な応用が可能である．

既存の比率規則抽出手法として，行列計算を用いて比率規則を捉える手法がある[2][3][4]．いずれの手法も，各タプルが複数の比率規則の線形結合で表されると考え，比率規則を行列分解により得られる特徴ベクトルとして表す．言い換えると原点を通る直線として比率規則が表される．それゆえ一部分でのみ成立する線形関係などは捉えることが難しい．

一方でわれわれは先行研究において，比率規則を線分とその周辺領域として定義し，比率規則の抽出を相関ルールマイニングの問題と対応づけた[5]．また相関ルールマイニングで用いられるサポートと確信度の概念を，比率規則に従うタプル数などの関係から定義した．最終的にはユーザより与えられた最小サポートと最小確信度を満たし，かつサポートあるいは確信度が最大となるような比率規則を抽出する．

ただしこの先行研究における定義では，対象となる２種類の数値属性は非対称に扱われる．これは比率規則が表す線分を，一方の属性における区間で表しているためである．こうすることでマイニングが高速に行える利点などを持つが，数値属性を入れ替えた結果は一致せず，得られた比率規則が表す線形関係も一致しない．

本論文では比率規則の抽出対象となる２属性を対称に扱う，対称比率規則を定義する．この対称比率規則は比率規則が成り立つ部分を，抽出対象の２属性両方を用いて表す．これにより抽出対象の２属性の役割によらず同様の線形関係を表すことが出来る．この対称比率規則について諸概念の定義および抽出手法の提案を行い，実験によりその手法の妥当性を示す．

本論文は以下のように構成される．２章では関連研究について述べ，それらと比較した対称比率規則の特徴を示す．３章では先行研究である非対称比率規則の定義をもとに，対称比率規則とその関連概念を定義する．４章では条件を満たす対称比率規則の抽出手法を提案する．５章では実データを用いた実験結果を示す．最後にまとめと今後の課題について述べる．

2. 関連研究

比率規則の抽出に関するアプローチとして，各データは比率規則の線形和によって表されるという仮定を元に，行列計算で比率規則を抽出する方法がある．具体的な行列分解手法としては，主成分分析を用いる手法[2]と，非負行列分解を元にした手法[3][4]がある．主成分分析を用いた手法では，全体の分布を最大にする軸である主成分ベクトルを比率規則とする．この手法は全体を一つの主要な比率規則で表し，続いてそれを補足する比率規則でデータを表現する．このとき各比率規則は直交するという制約を持っている．また第一主成分ベクトルが表す比率規則は，線形回帰分析[6]により得られる線形関係と本質的には同一である．非負行列分解を元にした手法では与えられたデータが非負の実数で表され，かつ比率規則が負の相関を持たないことを仮定している．このような場合には各比率規則は直交しないが，非負行列分解を用いることで妥当な比率規則を得ることができる．

異なるアプローチとしてわれわれの先行研究である，２次元空間中の線分とその周辺領域として比率規則を捉える手法がある[5]．得られる線形関係は２属性間のみに限られるが，比率規則を線分として考えることで前者のアプローチより一般化された定義となっている．またサポートや確信度の概念を数値属性に関する相関ルールマイニング[7]と対応付けて導入している．これにより得られる規則がどのような性質を持っているかの説明づけがなされるほか，サポートや確信度の最小値を調節することで，ユーザの意向

♥ 学生会員　筑波大学大学院システム情報工学研究科
hamamoto@kde.cs.tsukuba.ac.jp

♦ 正会員　筑波大学大学院システム情報工学研究科 / 筑波大学計算科学研究センター
kitagawa@cs.tsukuba.ac.jp

を結果に反映させることができる．

ただしこの定義上，扱う２属性は非対称である．つまり２属性の組<X, Y>から得られる規則と，役割を入れ替えた２属性の組<Y, X>から得られる規則は同一とならない．またマイニング対象の２属性のうち，一方の分布は考慮されない．この理由については次章で述べる．

本論文では先行研究における非対称性の問題を取り除くため，定義の拡張を行った．定義される対称性を持った比率規則は，属性の役割を入れ替えても同様の線形関係を表すことができる．

3. 比率規則

本章では対称比率規則の定式化を行う．以下ではまず本研究の対象データおよび先行研究の非対称比率規則を説明し，続いて対称比率規則とその関連概念を定義する．

3.1　対象とするデータ

本論文が対象とするデータは，数値属性を持つタプルの集合である．ただし各属性には欠損値は存在しないと仮定する．比率規則の抽出対象としては，ドメインが区間[-0.5, 0.5]となるよう正規化された２属性を考える．以下，比率規則を抽出する対象とする２属性をX, Yとし，それぞれの属性値をx, y ($-0.5 \leqq x, y \leqq 0.5$)と表現する．

3.2　非対称比率規則

比率規則は１章で述べたように，属性間の線形関係を表したものである．したがって分析対象の２属性X, Yで張られる空間を考えたとき，直線$y = ax + b$ ($a, b \in R$)として比率規則を考えることが最も単純である．

しかし，直線$y = ax + b$上に厳密な意味で複数のタプルが存在する状況は少ない．またパラメータa, bの取り得る値はどちらも$(-\infty, \infty)$の範囲における任意の実数であるため，Y軸にほぼ平行な直線を取り扱う際にa, bが無限大に発散する．それだけでなく，一般的に線形関係は全区間で成り立つとは限らない．

そこで１点目の問題には，パラメータに対する許容幅を設定し，許容幅内で異なる直線上に存在するタプルも，同一の比率規則に従うとする．２点目の問題についてはHough変換[8]により，パラメータが有限区間を取るよう変数変換を行う．Hough変換を用いると直線$y = ax + b$は$\rho = x\cos\theta + y\sin\theta$(ただし$\rho = b\sin(\tan^{-1}(-1/a))$, $\theta = \tan^{-1}(-1/a)$)と表現される．属性値x, yが区間[-0.5, 0.5]を取るよう正規化されているので，ρ, θの値はそれぞれ有限の区間$[0, \sqrt{2}/2]$, $[0, 2\pi]$で押さえられる．３点目については比率規則を線分として表現することで対処する．ここで非対称比率規則では属性Xにおける区間Iを与えて線分を表現する．

以上の点から非対称比率規則は次のように定義される．

> タプルt (x_t, y_t) ($x_t \in I$, $I \subseteq [-0.5, 0.5]$)が以下の式を満たす値ε_t, δ_tを持つとき，tは比率規則$RR_{x \in I}(\rho \pm \varepsilon, \theta \pm \delta)$に従う．
> $$\rho + \varepsilon_t = x_t\cos(\theta + \delta_t) + y_t\sin(\theta + \delta_t)$$
> ただし$|\varepsilon_t| \leqq \varepsilon$, $|\delta_t| \leqq \delta$

3.3　対称比率規則

前節で述べた非対称比率規則の定義では，線分を表す際に１属性のみに注目している．しかし２属性XとYは対称ではなく，両者の役割を入れ替えると得られない規則がある．例としてX軸に平行な規則を考える．この規則は２属性の役割を入れ替えると，Y軸に平行で$y \in I$を満たす区間で成り立つ規則となるべきであるが，非対称比率規則ではこの規則を表すすべが無い．

この原因は，定義中に属性Yのどの区間で成り立つか示されていないためである．そこで非対称比率規則の定義を拡張し，属性を入れ替えても同様に表すことのできる比率規則を，対称比率規則として次のように定義する．

> タプルt (x_t, y_t) ($x_t \in I$, $y_t \in J$, $I, J \subseteq [-0.5, 0.5]$)が以下の式を満たす値$\varepsilon_t$, δ_tを持つとき，tは対称比率規則$SRR_{<x \in I, y \in J>}(\rho \pm \varepsilon, \theta \pm \delta)$に従う．
> $$\rho + \varepsilon_t = x_t\cos(\theta + \delta_t) + y_t\sin(\theta + \delta_t)$$
> ただし$|\varepsilon_t| \leqq \varepsilon$, $|\delta_t| \leqq \delta$

以下誤解のない限り，短縮した形$SRR_{<I, J>}(\rho, \theta)$として対称比率規則を表現する．

3.4　対称比率規則の諸概念

非対称比率規則では，相関ルールマイニングと対応付けられたサポートや確信度などの概念を持っている．これらの概念は対称比率規則にも同様に導入できる．以下にその諸概念を定義する．

- サポート：領域<I, J>のサポートは，全タプルに対する，領域<I, J>に含まれるタプルの割合と定義し$support$(<I, J>)で表す．
- 確信度：対称比率規則$SRR_{<I, J>}(\rho, \theta)$の確信度を，$support$(<$I$, J>)に対する$support(SRR_{<I, J>}(\rho, \theta))$の割合 $support(SRR_{<I, J>}(\rho, \theta))/support$(<$I$, J>)で定義し，$\mathrm{conf}(SRR_{<I, J>}(\rho, \theta))$で表す．
- 最小サポート，最小確信度：抽出される対称比率規則に対してユーザから与えられる，最低限満たすべきサポートおよび確信度．以下ではそれぞれ$minsup$, $minconf$で表す．
- 最適確信度対称比率規則：$support$(<I, J>)が$minsup$を満たし，かつ$conf(SRR_{<I, J>}(\rho, \theta))$が$minconf$を満たした上で最大となるような対称比率規則$SRR_{<I, J>}(\rho, \theta)$．最大値を与える領域<$I$, J>を最適確信度領域と呼ぶ．
- 最適サポート対称比率規則：$conf(SRR_{<I, J>}(\rho, \theta))$が$minconf$を満たし，かつ領域<$I$, J>が$minsup$を満たした上で最大となるような対称比率規則$SRR_{<I, J>}(\rho, \theta)$．最大値を与える領域<$I$, J>を最適サポート領域と呼ぶ．

最適確信度対称比率規則は，領域<I, J>に含まれるタプル数が一定数以上である条件のもと，領域<I, J>に含まれれば確実に成り立つような対称比率規則である．また最適サポート対称比率規則は，領域<I, J>中で対称比率規則に従うタプルの割合が一定以上である条件のもと，なるべく多くのタプルが含まれる領域<I, J>を持つ対称比率規則である．

以下，最適確信度対称比率規則と最適サポート対称比率規則，および最適確信度領域と最適サポート領域をまとめてそれぞれ最適対称比率規則，最適領域と呼ぶ．

4. 提案手法

本章では最適対称比率規則を抽出するための手法を提案する．本手法ではすべてのタプル(全タプル数を N とする)がメインメモリに乗ることを仮定し，その仮定のもとで最

適領域の厳密な解を求める．またパラメータ ρ, θ はそれぞれ 2ε, 2δ ごとの離散値 $\rho_1, \rho_2, \ldots, \rho_R$ および $\theta_1, \theta_2, \ldots, \theta_T$ ($\rho_1 = 0, \theta_1 = 0$) として扱う．

求める最適対称規則は，サポートや確信度が全体で最も大きい１つではなく，各パラメータ組 (ρ_p, θ_q) ごとに求める．

アルゴリズム全体を，図１に擬似コードとして示す．アルゴリズムは，枝刈り，対称比率規則生成，対称比率規則統合の３フェーズから構成される．この手法では，対称比率規則の生成に関して $O(RTN^3)$ で行うことができる．以下各フェーズについて説明する．

```
/* 枝刈りフェーズ */
パラメータ組（ρ, θ）の候補の枝刈り
/* 対称比率規則生成フェーズ */
for each 残りの候補（ρi, θj）do
    候補集合 = {}
    for each 区間 Jk do
        最適サポート/最適確信度区間 I を抽出
        if SRR<I, Jk>(ρi, θj) が minsup および minconf を
          満たす then
              SRR<I, Jk>(ρi, θj) を候補集合に追加
        end
      end
      候補集合中，確信度/サポートが最大の対称比率
      規則を出力
end
/* 対称比率規則統合フェーズ */
for each 類似した組
          (SRR<Ia, Jb>(ρi, θj), SRR<Ic, Jd>(ρk, θl)) do
    SRR<Ia, Jb>(ρi, θj) と SRR<Ic, Jd>(ρk, θl) を同じ集合
    Sq に統合
end
対称比率規則集合 {S1, S2, ...}を出力
```

図１　対称比率規則抽出の提案手法

Fig.1　Symmetric Ratio Rule Mining Algorithm

4.1　枝刈りフェーズ

本提案手法では対称比率規則の候補を削減するため，以下に挙げる２種類の枝刈りを行う．

- 対称比率規則に従うタプル数による枝刈り

この枝刈りは非対称比率規則の抽出手法[5]で用いられている枝刈りとほぼ同様に，パラメータ ρ, θ に対する枝刈りを行う．抽出すべき対称比率規則 $SRR_{<I, J>}(\rho, \theta)$ に従うタプル数は，与えられた最小サポート *minsup* と最小確信度 *minconf* に対して次の式を満たす．

$$
\begin{aligned}
& N \times support(SRR_{<I, J>}(\rho, \theta)) \\
& \quad = N \times conf(SRR_{<I, J>}(\rho, \theta)) \times support(<I, J>) \\
& \quad > N \times minsup \times minconf
\end{aligned}
$$

１行目の最大値 $N \times support(SRR_{<[-0.5, 0.5], [-0.5, 0.5]>}(\rho, \theta))$ は全空間中で対称比率規則に従うタプル数を表す．この値が最小サポートと最小確信度の積未満の場合，それよりも小さな空間 $<I, J>$ の比率規則は条件を満たすことはないので，このときのパラメータ組 (ρ, θ) は枝刈りできる．

- 一定幅未満の候補に対する枝刈り

列挙を行う側の区間が非常に狭い場合，もう一方の区間を最大限広くしても最小サポートを満たさないことが考えられる．そのような狭すぎる区間に対する枝刈りを行う．

具体的には属性 Y でソートを行い，$j-i+1 \leqq N \times minsup$ ($i \leqq j$) を満たさない区間 $[y_i, y_j]$ を枝刈りする．本論文ではすべてのタプルがメインメモリに収まる程度のタプル数を想定しているので，手法全体におけるソート処理のコストは大きくないと想定する．

4.2　対称比率規則生成フェーズ

対称比率規則生成フェーズでは条件を満たす最適対称比率規則を生成する．このとき２次元数値属性相関ルールマイニング[9]を用いることで，最適対称比率規則を $O(RTN^3)$ で求めることができる．

２次元数値属性相関ルールマイニングでは，``２種類の数値属性 $X$, $Y$ が各数値属性で張られる２次元空間中の領域 $L$ に含まれるならば，条件 $C$ を満たす"という２次元数値属性相関ルールを考える．本研究においては $L$ として矩形領域を，$C$ として``対称比率規則を満たす"という条件をそれぞれ与えることで，２次元数値属性相関ルールマイニングを適用することが出来る．

4.3　対称比率規則統合フェーズ

対称比率規則をそのまま結果として出力すると，規則数が多い場合ユーザの結果の理解が困難になると考えられる．そこで類似した対称比率規則の統合を行う．

対称比率規則統合フェーズでは，２つの対称比率規則 $SRR_{<Ia, Jb>}(\rho_i, \theta_j)$, $SRR_{<Ic, Jd>}(\rho_k, \theta_l)$ の類似度を，以下の式で表される Jaccard 係数で測る．

$$
\frac{|\, SRR_{<Ia, Jb>}(\rho_i, \theta_j) \cap SRR_{<Ic, Jd>}(\rho_k, \theta_l) \,|}{|\, SRR_{<Ia, Jb>}(\rho_i, \theta_j) \cup SRR_{<Ic, Jd>}(\rho_k, \theta_l) \,|}
$$

$|SRR_{<I, J>}(\rho_i, \theta_j)|$ は，対称比率規則 $SRR_{<I, J>}(\rho_i, \theta_j)$ に従うタプル数を表す．すなわち，類似度は２つの対称比率規則の両方に従うタプルの，いずれかの対称比率規則に従うタプルに対する割合で表される．この値が閾値以上のとき２つの対称比率規則は同一の対称比率規則集合に統合する．以下ではこの閾値を *minmerge* と表記する．

5. 実験

本実験では提案手法により得られる対称比率規則の妥当性を検討する．妥当性は，与えられたタプルの分布に対しどのような比率規則が得られるかによって測る．そして属性を入れ替えたデータについても同等な比率規則が得られるかを確かめる．また比較対象として，非対称比率規則を抽出した結果も示す．

紙面の都合上最適確信度比率規則のみ示すが，最適サポート比率規則においても本質的に同じ結果である．

5.1　データの概要

本実験では UCI Machine Learning Repository[1] で配布されている Automobile データを使用する．このデータには 398 台の自動車に関する，燃費・重量・馬力など計８項目の数値およびカテゴリデータが記録されている．本実験ではそのうち，排気量と馬力の２属性を対象とする．ただし，いずれの属性値も区間[-0.5, 0.5]を取るよう正規化する．

5.2　実験結果

実験結果として各比率規則を２次元空間に表示する．個々の対称比率規則は濃いグレーの線分，薄いグレーの近傍，および最適領域 $<I, J>$ を表す矩形で表現する．非対称比率規則の場合，最適領域は一方の属性しか与えられないので，もう一方は任意（すなわち区間[-0.5, 0.5]）として矩形を

1 http://www.ics.uci.edu/~mlearn/MLRepository.html

表現する．また同一の比率規則集合に含まれるものはすべて同じ色で示し，異なるものは近傍領域の部分の濃さを変えて示す．

抽出した最適確信度対称比率規則を図２に示す．パラメータは $\varepsilon = 0.045$, $\delta = 0.04$, $minsup = 0.375$, $minconf = 0.8$, $minmerge = 0.3$ として得られた．720 個のパラメータ組（ρ, θ）に対する枝刈りの結果，37 組が残った．このデータは，タプルが密に分布している部分とそうでない部分にそれぞれ異なる線形関係が存在するが，実験結果から対称比率規則はいずれも捉えていることが分かる．

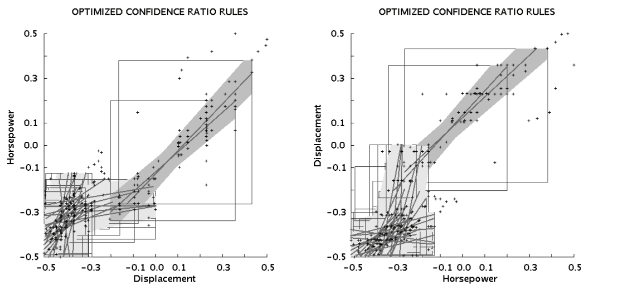

図２　抽出された最適確信度対称比率規則．左右の図で属性を入れ替えた場合を示している

Fig.2 Extracted Optimized Confidence Symmetric Ratio Rules. Left and right figures show a case when the target two attributes are swapped

これに対し最適確信度非対称比率規則の結果が図３である．パラメータは左図が $\varepsilon = 0.045$, $\delta = 0.04$, $minsup = 0.35$, $minconf = 0.75$, $minmerge = 0.3$，右図は $minconf$ のみ 0.7 としたものである．ここで水平方向の属性(左図：排気量，右図：馬力)が最適区間を求める対象となっている．実験結果を見ると，疎な部分の線形関係は捉えられているが，密な部分については垂直方向にタプルが無い部分にまで伸びる比率規則が得られていることがわかる．これは属性を入れ替えても同様であり，非対称比率規則はタプルの分布の疎密に影響を受けやすいといえる．

以上の結果より，対称比率規則は非対称比率規則の欠点を補い，妥当な線形関係を捉えることが出来ると考えられる．

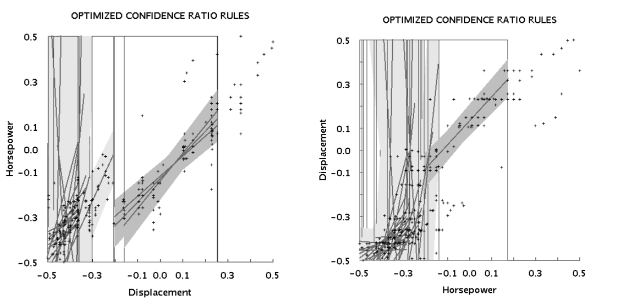

図３　抽出された最適確信度非対称比率規則

Fig.3 Extracted Optimized Confidence Asymmetric Ratio Rules

6. おわりに

本論文では属性を対称に扱う対称比率規則を定義し，その抽出手法を提案した．この対称比率規則に，サポートや確信度など相関ルールマイニングで用いられる概念を与え，サポートあるいは確信度を最大にする最適対称比率規則を抽出する手法を提案した．実データを用いた実験により，提案手法は適切なパラメータで妥当な対称比率規則を抽出できることを示した．

今後の研究として，３属性以上の高次元データに対する適用手法やより効果的な枝刈り手法の検討，また本提案手法でも最悪の場合 $O(RTN^3)$ かかるためより高速なマイニング手法の検討などが挙げられる．

[謝辞]

本研究の一部は，科学研究費補助金特定領域研究(#19024006)による．

[文献]

[1] Han, J. and Kamber, M.: “Data Mining: Concepts and Techniques”, Morgan Kaufmann (2001).

[2] Korn, F., Labrinidis, A., Kotidis, Y. and Faloutsos, C.: “Quantifiable Data Mining Using Ratio Rules”, VLDB Journal, vol. 8, pp. 254-266 (2000).

[3] Hu, C., et al.: “Mining Ratio Rules Via Principal Sparse Non-Negative Matrix Factorization”, Proc. 4th IEEE International Conference on Data Mining, pp. 407-410 (2004).

[4] Hu, C., et al.: “Mining Quantitative Associations in Large Database”, Proc. 7th APWeb Conference, pp. 405-416 (2005).

[5] 濱本雅史，北川博之:”サポートと確信度をもとにした比率規則による線形関係抽出”，情報処理学会論文誌：データベース（TOD32）(2006).

[6] Hastie, T., Tibshirani, R. and Friedman, J.: “The Elements of Statistical Learning”, Springer-Verlag (2001).

[7] Fukuda, T., Morimoto, Y., Morishita S. and Tokuyama, T.: “Mining Optimized Association Rules for Numeric Attributes”, Journal of Computer and System Sciences, vol. 58, number 1, pp. 1-12 (1999).

[8] Duda, R.O. and Hart, P.E.: “Use of the Hough Transformation to Detect Lines and Curves in Pictures”, Communications of the ACM, vol. 15, number 1, pp. 11-15 (1972).

[9] Fukuda, T., Morimoto, Y., Morishita S. and Tokuyama, T.: “Data Mining Using Two-Dimensional Optimized Association Rules: Scheme, Algorithms, and Visualization”, Proc. ACM SIGMOD International Conference on Management of Data, pp. 13-23 (1996).

濱本 雅史　Masafumi HAMAMOTO

筑波大学大学院システム情報工学研究科博士後期課程在学中．知識発見，データマイニングに関する研究に従事．日本データベース学会，情報処理学会，ACM 各学生会員．

北川 博之　Hiroyuki KITAGAWA

筑波大学大学院システム情報工学研究科，計算科学研究センター教授．理学博士(東京大学)．異種情報源統合，データマイニング，文書データベース，情報検索などの研究に従事．情報処理学会，電子情報通信学会 各フェロー．日本データベース学会副会長．日本ソフトウェア科学会，ACM，IEEE CS 各会員．著書に「データベースシステム」(昭晃堂)など．