社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2004－DBS－132　(11)
2004／1／16

# 属性ごとのクラスタリングと関連ある属性の統合に基づくデータの関連付け

松本　唯志[†]　　吉高　淳夫[†]

**概要**　複数の属性を持つデータを参照する際に、個々の属性に注目して類似するデータを探す場合がある。このとき、データの全ての属性について個々に類似するデータを参照することは少なく、属性の代表的な値や稀にしか見られない特殊な値という特徴的な値をとる属性のみに注目して、類似するデータを探すことが多い。そこで本研究では、属性ごとにクラスタリングした結果からデータごとに特徴的な属性を検出し、その属性について類似するデータを関連付ける。また、それらの属性の中で関連が強い属性については統合を行い、統合した属性についてデータの関連付けを行う。このデータの関連付けにより、データの特徴的な属性について類似するデータの参照を容易にする。

# Extracting Relation between Data based on Clustering and Integration of Correlated Attributes

Tadashi Matsumoto[†]　and　Atsuo Yoshitaka[†]

**Abstract**　When a user refers to data which have many attributes, he/she may search similar data of a individual attribute. Then, in many cases, he/she pays attention to the characteristic attributes that take a typical or rare value, and searches data which are similar to referred data. We detect characteristic attributes of individual data from the result of clustering, and extract relation between data of a individual attribute. Furthermore, we integrate correlated attributes, and extract relation between data of the set of integrated attributes. By extracting relation, a user easily refers to similar data of characteristic attributes.

## 1. はじめに

データベースの中で必要とするデータを参照したとき、その参照したデータに類似するデータを探す場合がある。例えば、特定の属性をキーとして検索を行い、必要とするデータを参照したとき、キーとした属性以外に、そのデータの中で特徴的な値をとる属性があり、その属性に注目して類似するデータを探すというような場合が考えられる。このような場合に容易に類似するデータを参照できるようにするためには、データの中で特徴的な値をとる属性を検出し、その属性について類似するデータへの関連付けを行うことが必要となる。

†: 広島大学大学院工学研究科
Graduate School of Engineering, Hiroshima University

ここで、属性値の中での特徴的な値とは、属性の代表的な値と特殊な値を指す。前者は類似するデータが多い値であり、後者は類似するデータが少ない値である。以後、前者をメジャーな値、後者をマイナーな値と呼ぶ。メジャーな値を参照することでそのデータベースでの基本的な情報を得ることができ、マイナーな値を参照することで新しい情報を発見することができる。よって、データが持つ複数の属性の中で、これらのメジャーな値をとる属性とマイナーな値をとる属性（以後、メジャーな属性とマイナーな属性）が重要な属性であるので、これらについてデータの関連付けを行う。

また、参照したデータの中で特徴的な値をとる属性が複数ある場合には、それらの個々の属性について関連するデータを探すという作業を何度か行うことになる。このとき、それらの属性の中で関連が強い属性がある場合、参照したデータに関連するデータとして提示される結果が類似する。類似する結果を度々参照することは冗長であるので、関連が強い属性を１つの集合として統合する必要がある。

本研究では、属性ごとにクラスタリングを行った結果から、データごとにメジャーとマイナーな属性を判断し、その属性について類似するデータを関連付ける。そして、メジャーとマイナーそれぞれについて属性間の関連が強い属性の統合を行う。本研究で行う統合とは、関連が強い属性のみを構成要素とする集合を作ることである。この属性の集合に対して類似するデータの関連付けを行う。これにより、データの中で重要な属性について関連するデータへのリンクを張り、参照を容易にする。

## 2. 関連研究

[1]では、関連ある属性を統合し、統合した属性の重要さを考慮して類似するデータを探すため、主成分分析を用いている。しかし、主成分分析による属性の統合は、全ての属性に対して重みを付けその和をとることで行うため、関連が弱い属性も小さな重みを付けられて統合される。よって、特定の属性に焦点を絞って関連する情報を参照することはできない。本研究では、属性ごとにクラスタリングを行い、関連が強い属性のみを集合としてまとめることで統合し、データの関連付けを行う。

[2], [3]では、複数ある属性の中から重要な属性を選択する手法を提案している。[2]ではクラスタリングを行うアルゴリズムである k 平均法の反復処理の中に、クラスタ内の分散が最小になるような特徴を選択するという処理を加えている。[3]では、ICD アルゴリズム[4]を用いて属性間の関連とデータ間の距離を求め、属性間の直接的な関連と間接的な関連を全ての属性について計算し、他の多くの属性との関連が弱い属性は除き、関連が強い属性のみを選択している。しかしこれらの手法は、データ全体に対して重要な属性を判断しており、個々のデータについてどの属性が重要であるかということを判断していない。よって、データ全体に対して重要であると判断した属性と、参照しているデータの重要な属性が異なる場合があり、このような場合にはデータの重要な属性に注目して関連するデータを参照するということができない。本研究では、データごとにメジャーとマイナーな属性を検出することで、参照したデータに対応した重要な属性の選択を行い、全てのデータについて、データ固有の重要な属性に注目して関連するデータを参照することを可能にする。

## 3. データの関連付け

データの関連付けを行う処理の流れを図 1 に示す。

まず、属性ごとに値の範囲が異なるため、関連付けの前処理としてそれぞれの属性について平均が 0、分散が 1 となるようにデータを標準化する。データベース中の $m$ 個のデータを $d_1, \ldots, d_m$、$n$ 個の属性を $f_1, \ldots, f_n$ とする。そして、データ $d_i$( $i$ = 1, 2, …, $m$)の属性 $f_j$( $j$ = 1, 2,

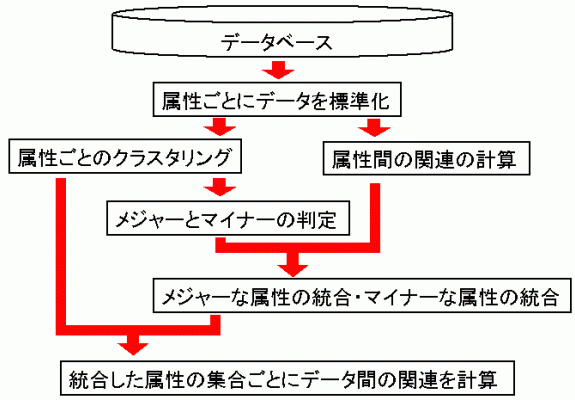


図 1　関連付けの処理の流れ

…, $n$)に対する値を $value(d_i, f_j)$とする。そして、データ $d_i$ の属性 $f_j$ について標準化した値 $value'(d_i, f_j)$を属性 $f_j$ の値の平均 $\mu_{f_j}$ と標準偏差 $\delta_{f_j}$ を用いて、式(1)で計算する。

$$value'(d_i, f_j) = \frac{value(d_i, f_j) - \mu_{f_j}}{\delta_{f_j}} \qquad (1)$$

この標準化したデータに対して以下の 3.1～3.5 の処理を行い、関連付けを行う。

### 3.1　属性ごとのクラスタリング

一般に用いられるクラスタリングの手法の 1 つに k 平均法がある。この手法ではクラスタ数を事前に指定しなければならないので、本研究では適切なクラスタ数を自動的に決定してクラスタリングするため、以下の①～⑥の処理を行う。この処理を全ての属性に対して行う。

① 初期値としてクラスタ数 $k$ を 2 とし、クラスタの重心を平均±標準偏差とする。
② クラスタ数 $k$ として k 平均法によりクラスタリングを行う。
③ 各クラスタの分散を計算し、その値が閾値（0.05）より大きいクラスタを検出する。
④ ③で検出されたクラスタ数を $l$ としたとき、$l$ が 0 なら処理を終了する。
⑤ ③で検出された $l$ 個のクラスタをそれぞれ 2 つのクラスタに分割し、分割前のクラスタの平均±標準偏差をクラスタの重心とする。これにより分割後のクラスタの重心の数 $k'$は $k+l$ となる。
⑥ クラスタ数 $k$ を $k'$とした後、②へ戻る。

### 3.2　属性間の関連の計算

属性間の関連を計算する。属性 $f_s$ と $f_t$ の関連を表す $r(f_s, f_t)$は式(2)によって求める。$\delta'_{f_s f_t}$ は属性 $f_s$ と $f_t$ の共分散であり、$\delta'_{f_s}$ と $\delta'_{f_t}$ はそれぞれ属性 $f_s$, $f_t$ の値の標準偏差である。前処理でデータの標準化を行っているので、$\delta'_{f_s}$ と $\delta'_{f_t}$ はともに 1 となっている。$r(f_s, f_t)$は、$f_s$ と $f_t$ の関係が、傾きが正である比例関係に近いときに 1 に近い値をとり、傾きが負である比例関係であるとき－1 に近い値をとる。$f_s$ と $f_t$ にこのような関係が見られないときには $r(f_s, f_t)$ は 0 に近い値をとる。つまり、$f_s$ と $f_t$ が比例関係に近いとき｜$r(f_s, f_t)$｜が 1 に近い値をとり、そのような関係が無いときには｜$r(f_s, f_t)$｜が 0 に近い値をとる。よって、この｜$r(f_s, f_t)$｜を属性間の関連の強さとする。

$$r(f_s, f_t) = \frac{\delta'_{f_s f_t}}{\delta'_{f_s} \times \delta'_{f_t}} \qquad (2)$$

### 3.3　メジャーとマイナーな属性の判定

個々のデータに注目し、個々の属性値についてのメジャーさを計算する。式(3)によってデータ $d_x$ の属性 $f_j$ の値のメジャー性 $major(d_x, f_j)$ を計算する。ここで、属性 $f_j$ に対するクラスタリングを行った結果生成されたクラスタの中で、所属データ数が最少のクラスタの所属データ数を $num_{min}(f_j)$、最大のクラスタの所属データ数を $num_{max}(f_j)$とする。$num(d_x, f_j)$は属性 $f_j$ において $d_x$ が所属するクラスタの所属データ数とする。

$$major(d_x, f_j) = 2 \times \frac{num(d_x, f_j) - num_{min}(f_j)}{num_{max}(f_j) - num_{min}(f_j)} - 1 \qquad (3)$$

この $major(d_x, f_j)$は、所属データ数が最も多いクラスタにデータ $d_x$ が所属するときに最大

値 1、最も少ないクラスタに所属するときに最小値 - 1 をとる。

全てのデータの個々の属性値についてメジャー性を計算した後、メジャーと判定する閾値の初期値を 1 とし、この閾値を徐々に小さくしていく。そして、メジャー性が閾値以上となる値の数が、値の総数（$m \times n$）の 40%を越えた時点での閾値を $Th_{major}$ とする。マイナーも同様にして、マイナーとなる値の数が、値の総数の 15%を越えた時点での閾値 $Th_{minor}$ とする。

### 3.4 属性の統合

メジャー性と属性間の関連の強さを用いて $d_x$ に対するメジャーな属性の統合、マイナーな属性の統合を行う。個々の属性 $f_j$ に対して以下の(ⅰ)～(ⅴ)の処理を行い、メジャーな属性を統合する。ここで、属性間の関連が強いかどうかを判定するための閾値 $Th_r$ は 0.7 とする。この値は、式(2)の値の 2 乗が一方の変数の分散を他方の変数の分散で推定できる割合を表し[5]、またその割合が約半分となるのが 0.7 であるという理由から決定した。

(ⅰ) データ $d_x$ の属性 $f_j$ の値に対するメジャー性が閾値 $Th_{major}$ 以上のときメジャーな値とし、以下の処理でメジャーな属性の統合を行う。メジャーな値ではない場合は処理を終了する。

(ⅱ) 属性 $f_j$ について関連が強い属性の集合を $Set_{major}$ と表し、$Set_{major}$ の要素を $f_j$ のみとする。

(ⅲ) 属性の中で $Set_{major}$ に含まれない属性の集合を $F$ とする。

(ⅳ) $F$ に含まれる属性 $f_u$ と $Set_{major}$ に含まれる属性 $f_v$ について式(4)を計算する。 そして、$Set_{major}$ に含まれる全ての属性について、式(4)が閾値 $Th_{major} \times Th_r$ 以上となる属性 $f_u$ を全て見つける。

(ⅴ) (ⅳ)で見つけた属性の中で、式(4)の値が最も高くなる属性を $Set_{major}$ に加えて(ⅲ)に戻る。(ⅳ)の条件に該当する属性が存在しない場合は属性 $f_j$ についての集合を $Set_{major}$ と決定して処理を終了する。

$$mr(d_x, f_u, f_v) = major'(d_x, f_u) \times r'(f_u, f_v) \quad (4)$$

$$major'(d_x, f_u) = \begin{cases} major(d_x, f_u), & \text{if (a)} \\ Th_{major} + (major(d_x, f_u) - Th_{major}) \times |Th_{major}|, & \text{otherwise} \end{cases} \quad (5)$$

$$r'(f_u, f_v) = \begin{cases} |r(f_u, f_v)|, & \text{if (b)} \\ Th_r + (|r(f_u, f_v)| - Th_r) \times Th_r, & \text{otherwise} \end{cases} \quad (6)$$

(a)： $|r(f_u, f_v)| \geq Th_r$ (b)： $major(d_x, f_u) \geq Th_{major}$

$major'(d_x, f_u)$ は属性 $f_u$ と $f_v$ の関連の強さが $Th_r$ 未満の際に、$Th_{major}$ との差が小さくなるように補正したメジャー性である。$r'(f_u, f_v)$ は属性 $f_u$ のメジャー性が閾値 $Th_{major}$ 未満の際に、$Th_r$ と属性間の関連の強さの差が小さくなるよう補正した値である。

閾値 $Th_{major}$ 以上かつ $Th_r$ 以上という条件を満たす属性のみを統合した場合には、以下の①もしくは②の条件を満たす属性 $f_u$ が統合されないため、式(4)～(6)により属性の統合の判定を行う。

① $f_u$ と $Set_{major}$ に含まれる属性との関連の強さは $Th_r$ より非常に大きいが、$f_u$ の値のメジャー性が $Th_{major}$ より僅かに小さい

② $f_u$ の値のメジャー性が $Th_{major}$ より非常に大きいが、$f_u$ と $Set_{major}$ に含まれる属性の関連の強さが $Th_r$ より僅かに小さい

マイナーな属性の統合は、閾値の違いとメジャー性の符号の違いがあるため、(ⅰ)～(ⅴ)の処理において以下のように変更することで行う。

- 閾値を $Th_{major}$ の代わりに $Th_{minor}$ とする
- 関連が強い属性の集合「$Set_{major}$」を「$Set_{minor}$」

とする

- （ⅰ)で「閾値 $Th_{major}$ 以上」ではなく「閾値 $Th_{minor}$ 以下」とする
- (ⅳ)で「閾値 $Th_{major}\times Th_r$ 以上」ではなく「閾値 $Th_{minor}\times Th_r$ 以下」とする
- 式(6)において条件(b)を条件(b')とする

$$\text{(b')}:\ major(d_x, f_u) \le Th_{minor}$$

### 3.5 データ間の関連の強さ

統合した属性の集合に対して、注目したデータと他のデータの関連の強さを計算し、関連付けを行う。本節では、メジャーな属性の集合に対してデータ間の関連の強さを計算する場合のみを説明する。マイナーな属性の集合については、メジャーな属性の集合に対する処理における「$Set_{major}$」を「$Set_{minor}$」に変更することで行う。

注目するデータ $d_x$ と他のデータ $d_y$ の属性の集合 $Set_{major}$ に対する関連の強さとして式(7)の $dr(d_x, d_y, Set_{major})$ を計算する。式(8)の $sim(d_x, d_y, f_j)$ は $f_j$ についての $d_x$ と $d_y$ の類似度であり、式(9)は属性 $f_j$ についての $d_x$ と $d_y$ の距離 $dis(d_x, d_y, f_j)$ である。距離 $dis(d_x, d_y, f_j)$ は、属性 $f_j$ についてのクラスタリングの結果、$d_x$ と $d_y$ が所属しているクラスタ間の距離を計算している。ここで、$d_x$ と $d_y$ が所属するクラスタに所属する全てのデータの集合をそれぞれ $S_{d_x}$ と $S_{d_y}$ としている。そして、それらの重心を $g(S_{d_x}, f_j)$、$g(S_{d_y}, f_j)$ とする。クラスタ間の距離は、重心間の距離とそれぞれのクラスタに所属するデータ間の最短距離の積をとることにより計算している。$dis(d_x, d_y, f_j)$ は必ず 0 以上となるので、$sim(d_x, d_y, f_j)$ の最大値は 1、最小値は 0 となる。

$$dr(d_x, d_y, Set_{major}) = \frac{1}{|Set_{major}|} \times \sum_{f_j \in Set_{major}} \left( sim(d_x, d_y, f_j) \sum_{f_{j'} \in Set_{major}} \frac{|mr(d_x, f_j, f_{j'})|}{|Set_{major}|} \right) \quad (7)$$

$$sim(d_x, d_y, f_j) = \frac{1}{1 + dis(d_x, d_y, f_j)} \quad (8)$$

$$dis(d_x, d_y, f_j) = |g(S_{d_x}, f_j) - g(S_{d_y}, f_j)| \times \left( \min_{d_{x'} \in S_{d_x}, d_{y'} \in S_{d_y}} |d_{x'} - d_{y'}| \right) \quad (9)$$

この $dr(d_x, d_y, Set_{major})$ の値が閾値 $Th_{dr}$ 以上であるとき、属性の集合 $Set_{major}$ においてデータ $d_x$ とデータ $d_y$ に関連があると判断する。閾値 $Th_{dr}$ については、メジャーな属性の集合のときは $Th_{major}\times Th_r$ とし、マイナーな属性の集合のときは $|Th_{minor}\times Th_r|$ とする。これらは属性の統合を判定するための式(4)に対する閾値であり、これらの閾値以上の属性のみが統合されているため、式(7)に用いた $|mr(d_x, f_j, f_{j'})|$ の平均が必ずこれらの閾値以上となる。よって、統合した属性の中の多くで $d_x$ と $d_y$ が同一のクラスタにあれば、式(7)の値もこの閾値以上になるという理由から、これらの閾値を用いる。

## 4. 関連の情報の保存

既存のデータベースに対してデータの関連付けを行った後、関連の情報（以後、関連情報）を保存する際に、そのデータベースのデータ構造を変更してしまうと、そのデータベースに関するシステム全てを変更しなければならない。そこで、既存のデータベースを変更せずに関連情報を保存するため、既存のデータベースとは別に関連情報データベースを作成する。

関連情報を関連情報データベースに保存するため 4.1～4.4 のクラスを用いる。これらの 4 つのクラスにより、関連情報をデータごとに、そして属性の集合ごとにまとめる。なお、これ以降ではデータベース中のデータを「データ」と表し、関連情報データベース中のデータを「関連情報データ」と表す。

### 4.1 データポインタクラス

このクラスのインスタンスはデータベース中の 1 つのデータへのリンクを持ち、さらに関連情報クラスのインスタンスへのリンクを属

性の集合が作られた数だけ持つことで、データごとに関連の情報を分類する役割を果たしている。

### 4.2 関連情報クラス

このクラスは、データポインタクラスのインスタンスが指すデータに対して複数作られた属性の集合のうち 1 つについて関連するデータをまとめる役割を持っている。

### 4.3 属性集合クラス

このクラスのインスタンスは属性の集合とその集合がメジャーかマイナーかという情報を保存する。異なるデータであっても属性の統合を行う際に同じ属性の集合ができることがあるため、関連情報クラスのインスタンスに属性の集合の情報を保存するのではなく、この属性集合クラスに属性の集合の情報を保存する形をとっている。

### 4.4 関連データクラス

このクラスのインスタンスは 2 つのデータの関連の強さを保存するためのものである。このクラスのインスタンスは、データポインタクラスと関連情報クラスのインスタンスを通じて辿られる。辿ったインスタンスが指すデータを $d_x$ とし属性の集合を $set_y$ とすると、$set_y$ に対して $d_x$ に関連するデータの中の１つへのリンクとそのデータ間の関連の強さを保存する。

### 4.5 関連情報データの保存例

これらのクラスを用いて関連の情報を保存した例を図 2 に示す。図 2 ではメジャーな属性の集合 $set_1$ についてデータ $d_1$ に関連するデータとして $d_3$ と $d_4$ があり、マイナーな属性の集合 $set_2$ についてデータ $d_1$ に関連するデータとして $d_2$ と $d_4$ がある例を示している。

$dp_1$ は $d_1$ を指しているので、$d_1$ についての関連情報をまとめるインスタンスである。$dp_1$ は $set_1$ や $set_2$ を指す関連情報クラスのインスタンスへのリンクを持っている。$set_1$ を指す関連情報クラスのインスタンスは、$set_1$ について $d_1$ と関連するデータを保存する関連データクラスのインスタンスへのリンクを持っている。これらのリンクの先には、関連の強さ $dr_1$ で $dp_3$ が指す $d_3$ が関連することを示すインスタンスや、関連の強さ $dr_2$ で $dp_4$ が指す $d_4$ が関連することを示すインスタンスがある。

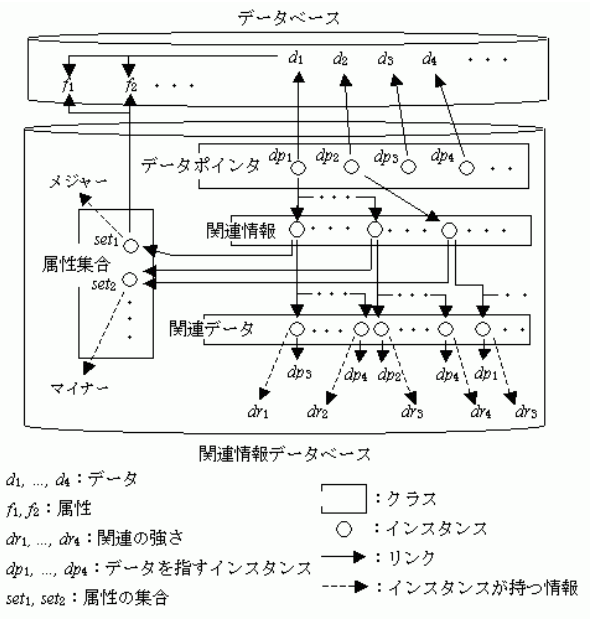


図 2　関連情報データベース

## 5. 実験

データの個々の属性に注目したときに、関連する情報を参照できるかどうかを確認するための実験を行った。実験では、本手法によるデータの関連付けと、重要な属性を選択する一般的な手法である主成分分析による関連付けを行い、両者の比較を行った。

実験に用いたデータは、東京都の各区、市についての構造別着工建築物のデータである。データの属性として、木造の棟数・床面積の合計・工事費予定額や、鉄骨鉄筋コンクリート造の棟数・床面積の合計・工事費予定額などがある。データの例を表 1 に示す。表 1 に示されていない属性としては鉄筋コンクリート造などがあり、属性数は 15 である。また、データ数は 54 である。

表 1　構造別着工建築物のデータの例

| | 総数 | | | 木造 | | | 鉄骨鉄筋コンクリート造 | | |
|---|---|---|---|---|---|---|---|---|---|
| | 棟数 | 床面積 | 工事費予定額 | 棟数 | 床面積 | 工事費予定額 | 棟数 | 床面積 | 工事費予定額 |
| 新宿区 | 834 | 611 747 | 15 635 818 | 376 | 42 597 | 812 001 | 50 | 269 914 | 8 169 321 |
| 千代田区 | 201 | 600 093 | 15 953 271 | 6 | 520 | 9 560 | 31 | 270 954 | 7 473 990 |
| 中央区 | 286 | 494 002 | 10 606 085 | 13 | 712 | 13 300 | 56 | 255 267 | 5 694 486 |
| 港区 | 658 | 2 763 931 | 72 204 909 | 168 | 24 776 | 460 491 | 91 | 1 818 983 | 48 876 529 |
| 文京区 | 623 | 284 521 | 5 604 355 | 309 | 39 000 | 748 267 | 33 | 88 128 | 1 746 560 |
| 台東区 | 484 | 287 261 | 5 894 485 | 93 | 9 722 | 185 443 | 40 | 139 297 | 2 587 370 |
| 墨田区 | 713 | 273 561 | 5 128 703 | 259 | 25 641 | 485 064 | 27 | 87 423 | 1 736 000 |
| 江東区 | 970 | 1 094 901 | 18 992 562 | 332 | 37 124 | 648 096 | 46 | 366 021 | 8 553 353 |
| 品川区 | 1 143 | 461 423 | 9 198 877 | 673 | 75 574 | 1 472 409 | 38 | 105 983 | 2 010 203 |
| 目黒区 | 1 302 | 400 390 | 8 302 018 | 855 | 104 801 | 2 036 238 | 23 | 54 719 | 1 377 950 |
| 大田区 | 2 753 | 1 045 333 | 20 868 729 | 1 431 | 162 650 | 3 054 686 | 34 | 168 268 | 2 937 361 |
| 世田谷区 | 3 763 | 977 488 | 19 881 030 | 2 717 | 343 244 | 6 577 300 | 24 | 49 429 | 1 003 832 |
| 渋谷区 | 899 | 481 884 | 10 559 963 | 421 | 56 002 | 1 081 586 | 35 | 132 000 | 2 705 729 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

### 5.1　実験結果

本手法による関連付けの例と、主成分分析を行った関連付けの結果を示す。本実験では、新宿区のデータに注目して関連付けを行った。

本手法により新宿区のデータの中でメジャーな属性を検出したところ、木造の棟数がメジャーな属性であったため、この木造の棟数について関連するデータの例を示す。本手法では、木造の棟数・床面積の合計・工事費予定額はメジャーな属性として統合されており、この統合された属性の集合について、新宿区と関連があると判定されたデータの結果を表 2 に示す。表 2 より、木造の棟数について新宿区と類似したデータが関連付けられていることがわかる。また、関連付けられたデータは、統合された属性である木造の床面積の合計や工事費予定額についても新宿区と類似した値をとるデータが関連付けられていることがわかる。

次に、主成分分析によって属性の統合を行った際の関連付けの結果を示す。主成分分析を行った結果、属性に重みが付けられて統合されるので、本実験では木造の棟数の重みが最も大きい主成分１つを用いてクラスタリングを行い、新宿区と同一のクラスタに所属するデータを関連付けた。その結果を表 3 に示す。この主成分は、木造の棟数の重みが正の向きに大きく、また総数の棟数の重みが負の向きの大きくなっていた。これより、木造の棟数と総数の棟数が新宿区に類似している江東区や渋谷区が関連付けられている。さらに、木造の棟数と総数の棟数という 2 つの属性値の違いが相殺されてしまうことにより、千代田区や中央区のように木造の棟数と総数の棟数がともに新宿区より小さいというデータも関連付けられている。

### 5.2　結果の比較

主成分分析による属性の統合により特定の属性に注目してデータの関連付けを行った場合では、注目した属性について類似していないデータであっても他の属性の重み付けの影響により、関連付けられた。しかし、本手法では注目した属性について類似したデータのみが関連付けられている。この結果から、個々の属性に注目した関連付けにおいて本手法が有効であると言える。

また、本手法で行った属性の統合により統合された属性についても注目したデータと関連付けられているデータが類似しているため、属性の統合も適切であると考えられる。

## 6.　まとめ

本研究では、データごとに重要な属性であるメジャーとマイナーな属性を検出し、その中でも関連が強い属性は統合を行い、その統合した属性の集合に対してデータの関連付ける手法

表 2　本手法により新宿区と関連付けられたデータ

| | 総数 | | | 木造 | | | 鉄骨鉄筋コンクリート造 | | |
|---|---|---|---|---|---|---|---|---|---|
| | 棟数 | 床面積 | 工事費予定額 | 棟数 | 床面積 | 工事費予定額 | 棟数 | 床面積 | 工事費予定額 |
| 新宿区 | 834 | 611 747 | 15 635 818 | 376 | 42 597 | 812 001 | 50 | 269 914 | 8 169 321 |
| 渋谷区 | 899 | 481 884 | 10 559 963 | 421 | 56 002 | 1 081 586 | 35 | 132 000 | 2 705 729 |
| 豊島区 | 894 | 328 988 | 6 718 372 | 513 | 57 923 | 1 089 947 | 27 | 80 218 | 1 676 246 |
| 江東区 | 970 | 1 094 901 | 18 992 562 | 332 | 37 124 | 648 096 | 46 | 366 021 | 8 553 353 |
| 昭島市 | 581 | 237 035 | 3 589 190 | 412 | 42 289 | 750 166 | 5 | 81 097 | 1 250 800 |
| 国分寺市 | 723 | 127 564 | 2 424 647 | 568 | 57 332 | 1 065 218 | 2 | 18 564 | 320 000 |
| 田無市 | 482 | 102 446 | 1 856 060 | 391 | 40 352 | 755 026 | 2 | 952 | 21 000 |
| 保谷市 | 555 | 115 262 | 2 002 322 | 456 | 48 837 | 855 165 | 1 | 45 | 800 |
| 狛江市 | 424 | 90 262 | 1 451 793 | 342 | 36 215 | 684 864 | 2 | 180 | 4 220 |
| 東大和市 | 594 | 141 086 | 2 191 356 | 479 | 48 648 | 875 986 | 4 | 35 250 | 454 000 |
| 清瀬市 | 540 | 104 304 | 1 471 320 | 392 | 40 700 | 704 324 | 2 | 3 158 | 55 200 |
| 東久留米市 | 685 | 136 320 | 2 224 467 | 572 | 56 248 | 1 012 103 | 5 | 32 822 | 512 500 |
| 多摩市 | 502 | 155 397 | 2 972 832 | 353 | 40 956 | 772 279 | 7 | 28 314 | 574 049 |
| 武蔵野市 | 621 | 170 408 | 3 353 741 | 451 | 53 421 | 1 004 152 | 5 | 31 465 | 718 748 |

：　メジャーな属性

表 3　主成分分析により新宿区と関連付けられたデータ

| | 総数 | | | 木造 | | | 鉄骨鉄筋コンクリート造 | | |
|---|---|---|---|---|---|---|---|---|---|
| | 棟数 | 床面積 | 工事費予定額 | 棟数 | 床面積 | 工事費予定額 | 棟数 | 床面積 | 工事費予定額 |
| 新宿区 | 834 | 611 747 | 15 635 818 | 376 | 42 597 | 812 001 | 50 | 269 914 | 8 169 321 |
| 千代田区 | 201 | 600 093 | 15 953 271 | 6 | 520 | 9 560 | 31 | 270 954 | 7 473 990 |
| 中央区 | 286 | 494 002 | 10 606 085 | 13 | 712 | 13 300 | 56 | 255 267 | 5 694 486 |
| 江東区 | 970 | 1 094 901 | 18 992 562 | 332 | 37 124 | 648 096 | 46 | 366 021 | 8 553 353 |
| 品川区 | 1 143 | 461 423 | 9 198 877 | 673 | 75 574 | 1 472 409 | 38 | 105 983 | 2 010 203 |
| 目黒区 | 1 302 | 400 390 | 8 302 018 | 855 | 104 801 | 2 036 238 | 23 | 54 719 | 1 377 950 |
| 渋谷区 | 899 | 481 884 | 10 559 963 | 421 | 56 002 | 1 081 586 | 35 | 132 000 | 2 705 729 |
| 府中市 | 1 431 | 316 707 | 5 751 718 | 1 027 | 113 906 | 2 132 767 | 8 | 25 371 | 388 400 |

：　重みが大きい属性

を提案した。

今後の課題は、メジャーとマイナーの判定を行う前に、データ自体にメジャーとマイナーな値が存在するかどうかを判定することである。データが一様に分布している場合には、クラスタリングを行った結果、全てのクラスタの所属データ数がほぼ等しくなるため、メジャーやマイナーな値が存在しない。属性がこの状態になっているかどうかの判定と、この状態での関連付けの方法について今後考える必要がある。

**参考文献**


[1] 三治　信一朗, 橋本　周司, “多次元データを把握するための視覚化ツール”, 情報処理学会, コンピュータビジョンとイメージメディア研究報告, No. 127, pp. 167-170 , 2001.

[2] 井上　光平, 浦浜　喜一, “特徴選択を伴うクラスタリングの反復解法”, 電子情報通信学会論文誌, D-Ⅱ, Vol. 83 - D -Ⅱ, No. 4, 2000.

[3] Tomoya Ogawa, et al. , “FEATURE SELECTION FOR EFFECTIVE CALCULATION OF A SIMILARITY MEASURE”, International Conference on Artificial and Computational Intelligence, 2002.

[4] G. Das and H. Mannila. “Context-Based Similarity Measures for Categorical Database”, Proc. of PKDD2000, pp. 200-210, 2000.

[5] 肥田　野直, “統計入門”, 培風館, 1995.