1999 年情報論的学習理論ワークショップ
1999 Workshop on Information-Based Induction Sciences (IBIS'99)
Izu, Japan, August 26-27, 1999.

規則性を持つ部分データを抽出するアルゴリズムの提案

Data decomposition algorithm that extracts subsets of data each of which contains a rule

山川 宏* 岡田 浩之† 渡部 信雄‡
Hiroshi Yamakawa Hiroyuki Okada Nobuo Watanabe

Abstract: To improve the prediction performance we propose new preprocessing method for spreadsheet formatted data. This method decomposes input data into reusable parts, each of which contains a rule. We firstly discuss the Matchable principle, which supporting generalization ability. This principal emphasizes the increasing of matching chance to extract information structure. Next we derives a Matchable criterion suitable for the situation decomposition from the principal. Then we developed a search algorithm for decomposition and speed up it by reducing enormous search space. Simulation demonstrates that the algorithm can decompose mixed situations. This technology is effective for pre-processing of data analysis and pattern recognition.

Keyword: Information structure, Matchability criterion, Situation decomposition, Rule extraction, Prediction performance, Data mining

1 はじめに

本稿では高度な予測能力の実現に必要な情報のセグメンテーション技術とそのための評価基準の提案を行う。図 1 の中央の破線で示すように、直接的に経験を用い予測を行う Memory based reasoning (MeBR) では距離が近いケース (レコード/タプル/イベント) のみを用いる。このために距離が近いケースが存在しない場合に無力である点で予測能力が低い。少ない経験から予測を行うには距離が遠いケースの活用が鍵となる。典型的な Model based reasoning(MoBR) による予測を図 1 の下の破線で示す。まず全てのケースの影響をモデルに蓄積し、モデルを通して予測を行うので、間接的に全てのケースを利用できる。我々が予測能力を向上させるためのアプローチを 図 1 の上側の矢印で示す。まず予めセグメンテーションされていない経験を再利用性の高い小さな部品に分解し、次にその部品を再構成して予測を行う。こうすると未知の入力に対して、過去の部分的な経験を組合わせによる予測を行うことで汎化能力が得られる。本稿ではその要素技術となる経験を分解するための状況分解技術を提案する

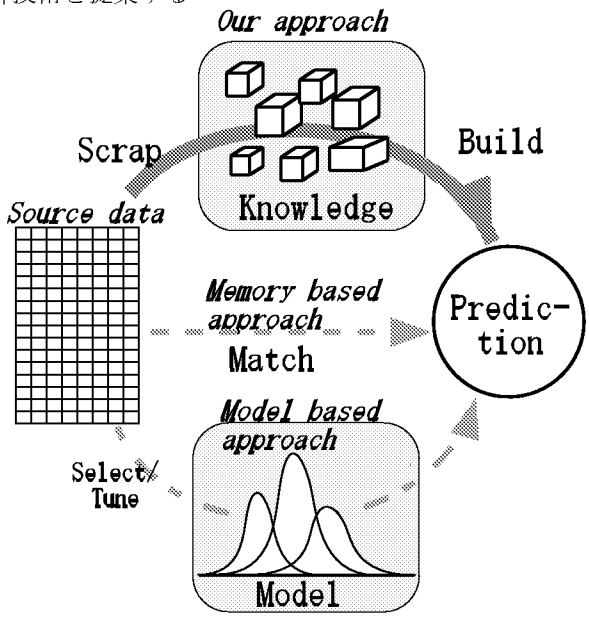

図 1: 予測のための主なアプローチ

*新情報処理開発機構自律学習機能富士通研究室〒 261-8588 千葉市美浜区中瀬 1-9-3 (幕張システムラボラトリ 13F) tel. 043-299-3450
e-mail yamakawa@flab.fujitsu.co.jp
Autonomous Learning FUJITSU Lab. Real World Computing Partnership 9-3, Nakase 1-Chome, Mihama-ku, Chiba City, Chiba 261-8588, Japan
†同上 (same as above)
‡同上 (same as above)

2 状況分解

提案する経験を分解する技術を状況分解と呼ぶ事にする。ここで取扱うデータの形式は、多くの予測手法や統計的手法で利用される標準的なスプレッドシート型データであり、つまり複数の特徴量（属性/フィールド）と複数のケース (レコード/タプル/イベント）のマトリックス構造を持つデータであるとする。

この種のデータに対する分解手法として (1) ケースを選択する技術や (2) 特徴量を選択する技術がパターン認識の分野において数多く見られる。しかし本稿で提案する状況分解は、 (3) ケースと特徴量の両方を同時に選択する点で多くの技術と異なる。状況分解では 図 2 に示すように与えられた全体状況から、後述する Matchability 基準が極大値となる複数の規則性の高い Matchable 状況を抽出する。

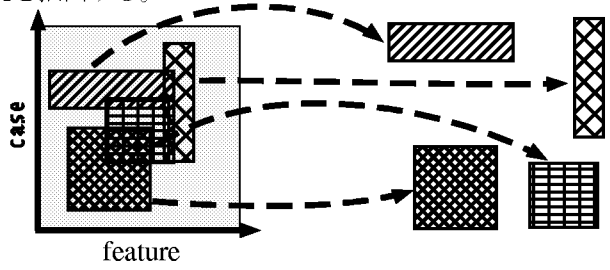

図 2: 状況分解は複数の Matchable 状況を抽出する

提案する状況分解は規則性の高い部分状況を抽出する。一方、規則そのものを抽出する技術は概念獲得や知識発見などと呼ばれる領域で多数存在する。先駆的には特徴量間の関係を導く BACON アルゴリズム [1] が有名である。近年はデータマイニング分野で大量データから規則を抽出する研究が盛んである [2]。特に高須らによる近似的関数従属性の抽出では特徴量とイベントを同時に選択する点で関連が深い [3]。

状況分解では以下で述べる Matchable 原理を利用するので (1) 関連の強い部分情報のまとまりを複数抽出するという性質をもち、上記の概念獲得などの技術もこれに相当する。これに対して分類や判別課題などにおいては (2) 強い関連に基づいてケースや特徴量を分解するので提案する手法とは異なっている。

3 Matchable 原理から Matchability 基準へ

3.1 モデル選択基準と Matchable 原理

標本データから自動的に内部構造を推定 (モデルを選択) する主な目的には予測性の向上、データ圧縮、視覚化の 3 つがある。このような技術を実現する際には必ず何らかのモデル選択基準が用いられるので、まずこれらについて概観する。

MoBR では対象領域における先験的知識を反映するモデルクラスの中から標本データを最もよく説明するユニークなモデルインスタンスが選ばれる。設計されるモデルクラス内の各インスタンスには先験的知識の反映の程度からみてより単純なものを選ぶように評価が与えられる。モデル選択基準に現われる基本的構図は、“データへの無矛盾性” と “モデルの単純さ” のトレードオフである。対象領域の先見知識を利用できない場合には、知識反映因子は単純さのみを表すのでたとえばエントロピーの最大化などが採用されることになる。

最大エントロピー法などでは、“データへの無矛盾性” と “モデルの単純さ” のバランスを調整する任意定数の値に依存して選択されるインスタンスが変化する問題がある。恣意性を除いた基準として期待平均対数尤度から統一的にトレードオフを導く Akaike's Information-theoretic Criterion (AIC) [4] がある。 AIC は自由パラメタ数からモデルの最大対数尤度を引いた基準量を小さくするモデルを選ぶ。最大対数尤度が “データへの無矛盾性” を示し、自由パラメータ数の最小化が “モデルの単純さ” に相当する。

一方、情報理論においてデータ圧縮技術は現実的な課題であり、利用される評価基準は情報量である。データ圧縮に基づく特徴量選択では相互に独立な特徴量が選択される。データ圧縮技術にモデル選択の思想が結びつくことで Minimum Description Length (MDL)[5][6] 原理が導かれた。MDL は「与えられたデータを、モデル自身の記述も含めて最も短く符号化できるような確率モデルが最良のモデルである」と主張する基準で、AIC と同様に任意定数を含まない。符号化されたデータの記述長が小さいことは、モデルがデータの分布を正確に反映したことを示すので、データの記述長の短さが “データへの無矛盾性” となる。また、モデルの記述長の最小化が “モデルの単純さ” を表わす。

以上のようなモデル選択基準は、オッカムの剃刀と呼ばれる構図に基づいており、“データへの無矛盾性” と “モデルの単純さ” の間でのトレードオフに基づきユニークなモデルを選択する。

これに対して Matchable 原理では、「マッチング機会の大きい複数の部分構造を抽出することで予測能力を向上する」。こう考えた理由は、既に経験したデータ内で相互にマッチする部分的な構造は、現在以降に得られる入力ケースに対してもマッチングの機会 (再利用性) が大きくなり易いと推定できるからである。つまり、部分

構造に注目することで一般化能力を得ようとしているのであり、「ある事柄の内容／性質などを明らかにするため、細かな要素に分けていく」分析の考えの見方を少し変えたともいえる。

そして部分構造をに着目することが中心的となるMatchable原理においてはその構造の選択基準においてデータの説明範囲が明示的に含まれることになり、これがオッカムの剃刀型の情報基準との大きな違いになっている。結局Matchable原理においては“モデルの単純さ”と“データの説明領域”のトレードオフがその焦点となる。

ところで計算論的学習論における“モデルの単純さ”への傾向は、単純なモデルほど存在する可能性が大きいという論点で語られることがあるが、Matchable原理の視点ではむしろマッチする機会が増大することによって予測性が増大するという実用的な観点に注目している。つまり利用する構造が複雑化する事でデータ間の相互予測性が減少するのを避けようとしているのである。

3.2 状況分解のためのMatchability基準

Matchable原理を状況分解に適用する場合に利用されるのが、今回提案するMatchability基準で、以下の３つの因子のトレードオフに基づく。“データの説明領域”としては、ケース数と特徴量数が用いられ、“モデルの単純さ”としては部分状況を記述する情報量が用いられる。

基準内に含まれる３因子間のトレードオフは基本的に“モデルの単純さ”と“データの説明領域”間のトレードオフとみなせる。一つには単純なモデルほど多くのケースを説明することが困難であるという一般的傾向であり、もう一つには単純なモデルほど多くの特徴量を説明することが困難であるという一般的傾向である。

なお、状況分解が特徴量の部分集合である点からは、「実世界情報における特定の関連性は、特定の特徴量間の関係に反映する」という先入観に基づいているといえる。逆にいえばそのような部分構造のみに探索範囲を限定しているともいえる(それでも十分膨大な探索空間だが)。

我々が状況分解と呼んでるケースと特徴量の部分集合を選択するという自律的なアルゴリズムを研究する必要性を感じた初期的な経緯は、ミンスキー[7]等が述べているようなマルチエージェントシステムによって知能を実現することを目指していることがある。このシステム内での各エージェントは人間の脳で行われているように時間軸と空間軸の両方に対して部分的な情報を扱っており、なおかつその構造は自己組織的に獲得される。そのために人間のように高度に自律的な学習能力を持つシステムの実現にはこの種の情報構造技術が必須であると考えており、我々は他方で分散知能アーキテクチャーの研究を進めている[8]。

4 定式化

4.1 探索空間

4.1.1 特徴量とケースによる探索空間

Matchable状況の探索は、図3任意の部分状況（つまり任意の特徴量とケースの組み合わせ）からMatchabilityという評価量が極大となる複数の部分状況を選択することである。このため、Matchable状況では特徴量選択およびケース選択の何れの方向についても任意の微少変化に対しても評価の減少を招くことになる。

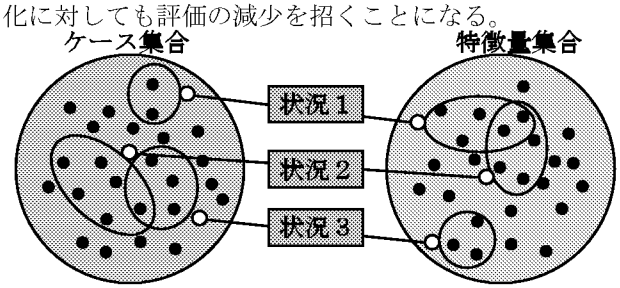

図 3: 特徴量とケースの選択による部分状況の指定

特徴量とケースのNotation N個の特徴量の集合とD個のケースの集合の直積集合の中に値を持つデータを全体状況とする。これは、通常は与えられた問題空間に相当する。図4に示すように、ここでは特徴量とケースのそれぞれを選択ベクトルを利用してその部分集合を指定する。任意の特徴量を選択すベクトルを$\boldsymbol{d}$, ケースを選択するベクトルを$\boldsymbol{n}$とする。これによって指定される任意の部分状況を$J(\boldsymbol{d}, \boldsymbol{n})$とする。

$\boldsymbol{d} = (d_1, d_2, ..., d_i, ..., d_D)$: 特徴量選択ベクトル

$\boldsymbol{n} = (n_1, n_2, ..., n_N)$:ケース選択ベクトル

$J(\boldsymbol{d}, \boldsymbol{n})$: 任意の部分状況

ここで、ベクトル要素$d_i(= t(rue)/f(alse))$,$n_i(= t(rue)/f(alse))$は選択/非選択の二値情報であり、選択された特徴量の数を$d = \sum_{i=1,d_i=true}^{D} 1\ (0 < d < D)$、選択されたケースの数を$n = \sum_{i=1,n_i=true}^{N} 1\ (0 < n < N)$とする。

また、全ての特徴量が選択するベクトルを$\boldsymbol{D}$とし、全てのケースを選択するベクトルを$\boldsymbol{N}$とすれば、全体状況は$J(\boldsymbol{D}, \boldsymbol{N})$となる。

探索空間の広さ 特徴量選択の全空間の広さはベキ集合の数であるから2^D個であり、ケース選択の全空間の広さもベキ集合の数であり2^N個である。そして、部分状

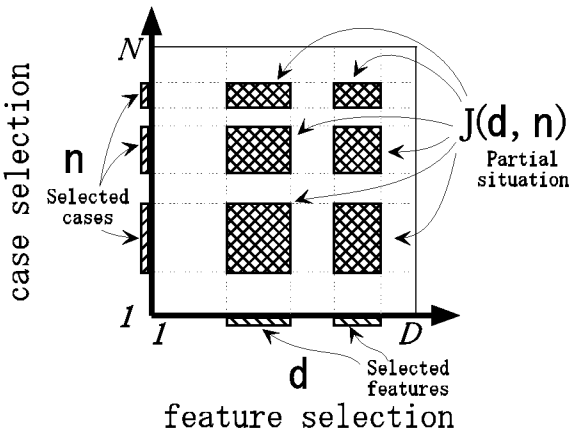

図 4: 特徴量とケースの選択による部分状況の指定

況の数は特徴量のベキ集合とケースのベキ集合の直積集合の数であるから、2^{D+N} 個である。

最近傍　特徴量選択ベクトル $\boldsymbol{d_0}$ の最近傍のベクトル $\boldsymbol{d'}$ は、$\boldsymbol{d_0}$ から任意の一つの特徴量を追加または削除したベクトルである。同様にあるケース選択ベクトル $\boldsymbol{n_0}$ の最近傍のベクトル $\boldsymbol{n'}$ は、$\boldsymbol{n_0}$ から任意の一つのケースを追加または削除したベクトルである。

4.1.2　セグメント選択空間としての部分特徴量空間

今回の定式化では処理の効率化を目的として、ケースを選択する代わりにセグメント選択空間[1]でのセグメント選択を利用する。 セグメント選択空間は特徴量選択ベクトルを指定したあとに各特徴量をセグメントに分離した離散量とみなすことによって得られる空間である。

特徴量毎のセグメント　各特徴量の値は分割された有限個のセグメント内に割り当てられるとすることで名義尺度 (カテゴリー変数) として扱われる。状況分解の内の処理に必要な設定情報は特徴量毎のセグメント数である。そこで、i 番目の特徴量のセグメント数 s_i を列挙したセグメント数ベクトル $\boldsymbol{s}$ を定義する。

$\boldsymbol{s} = (s_1, s_2, ..., s_D)$: セグメント数ベクトル

セグメント選択空間　特徴量選択ベクトル $\boldsymbol{d}$ の決定により指定される部分特徴量量空間では、以下の図 5 に示すように各特徴量毎のセグメントの組合せであるセグメント選択空間が定義できる。このセグメント選択空間における全セグメント数 $S_{\mathbf{d}}$ は、選択された各特徴量のセグメント数の積なので次式となる。

[1]セグメント選択は特徴量選択やケース選択とは明らかに異なる

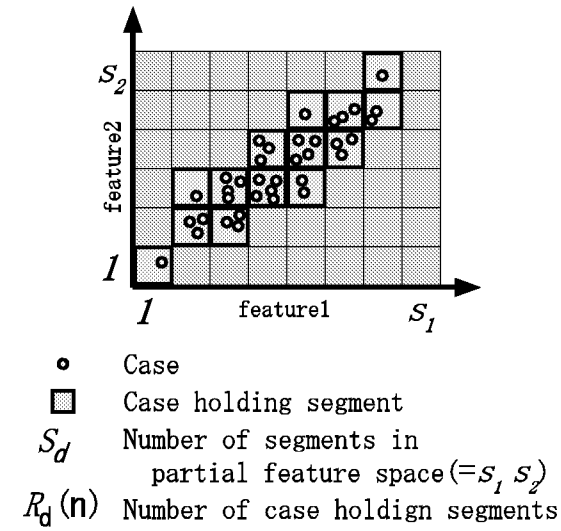

図 5: セグメント選択空間とセグメント内のケース作図

の都合上、２次元特徴量の場合が示してある。

$$S_{\mathbf{d}} = \prod_{i=1, d_i=true}^{D} s_i \tag{1}$$

この $S_{\mathbf{d}}$ の大きさを持つセグメント選択空間からの、任意のセグメントの選択を表わすベクトル $\boldsymbol{r_{\mathbf{d}}}$ を以下のように定義する。

$\boldsymbol{r_{\mathbf{d}}} = (r_{d1}, r_{d2}, ..., r_{dS_{\mathbf{d}}})$: セグメント選択ベクトル

$r_{di} = t(rue)/f(alse)$

なお、セグメント選択ベクトル $\boldsymbol{r_{\mathbf{d}}}$ で選択されたセグメントの数を $r_{\mathbf{d}}$ とする。

$$r_{\mathbf{d}} = \sum_{i=1, r_{d_i}=true}^{S_{\mathbf{d}}} 1 \tag{2}$$

ここで、ベクトル要素 $r_{\mathbf{d}}$i は選択/非選択の二値情報であり、選択されたセグメントの数の範囲は $(0 \leq r_{\mathbf{d}} \leq S_{\mathbf{d}})$ であり、セグメント選択ベクトルを利用した部分状況を $\tilde{J}(\boldsymbol{d}, \boldsymbol{r_{\mathbf{d}}})$ とする。

4.2　Matchability 基準

4.2.1　基本三要素毎の検討

任意のケースと特徴量の選択である部分状況 $J(\boldsymbol{d}, \boldsymbol{n})$ の大きな集合の中から、Matchable 状況を選択するための基準について考える。これまでの議論から選択すべき部分状況の性質として、選択ケース数 n と選択特徴量数 d を増加させると同時に、“モデルの単純さ” のためにセグメント数 $r_{\mathbf{d}}$(式 2 参照) を減少させれば良いことがわかる。ここでは選択徴量数 d を用いる変わりに全セグメント数 $S_{\mathbf{d}}$ を用いるが、これが利用できるのは、部

分特徴量空間における全セグメント数 $S_{\mathbf{d}}$ は、特徴量毎のセグメント数 s_i が 2 以上であることから、選択特徴量数 d と単調増加の関係にあるためである。

4.2.2 規格化された主要三変数

上記の 3 変数の性質を反映し、かつ、意味が解釈できる新たな無次元量を以下に定義する。

1. n/N ： 部分状況に含まれるケース数の比率 → 大きく $(0 < n/N < 1)$
2. $n/r_{\mathbf{d}}$ ： 部分特徴量空間におけるセグメント毎の選択ケース数の平均値 → 大きく $(1 < n/r_{\mathbf{d}})$
3. $r_{\mathbf{d}}/S_{\mathbf{d}}$ ： 部分特徴量空間における選択セグメント数の空間占有率→小さく $(0 < r_{\mathbf{d}}/S_{\mathbf{d}} < 1)$

4.2.3 提案する Matchability 基準

上記の、主要三変数を組み合わせて、特定の部分状況 $J(\boldsymbol{d}, \boldsymbol{n})$ に対する Matchability 基準を以下のように定義する。

$$M(n, r, S_{\mathbf{d}}, N) = C_1 \log \frac{n}{N} + C_2 \log \frac{n}{r_{\mathbf{d}}} - C_3 \log \frac{r_{\mathbf{d}}}{S_{\mathbf{d}}} \quad (3)$$

$$C_1, C_2, C_3 \text{ は正の定数}$$

Matchability 基準は極大値を探すために利用されるので、絶対的な大きさ関しては意味を持たないので、定数の有効な自由度は 2 次元である。

なお、現状の Matchability 基準は上記のような定性的な議論から “モデルの単純さ” と “データ説明領域の” トレードオフを反映するように直感的に導いたものであり、今後は予測性の観点からみた理論的な検討がますます必要であると思われる。そのため C_1, C_2, C_3 の定数の設定についても、正の値にすべきである程度のことしか明確化されておらず、あとはノウハウの域を出ていない。

4.3 極大値探索におけるセグメント空間の利用による高速化

全体状況 $J(\boldsymbol{D}, \boldsymbol{N})$ に含まれる非常に大きな $O(2^{N+D})$ の探索空間から、効率的に Matchability 基準が極大値となる複数の部分状況 $J(\boldsymbol{d}, \boldsymbol{n})$ を取り出す必要がある。一般に全特徴量数 D に対して全ケース数 N が大きいと思われるので、提案するアルゴリズムではケース選択の代わりにセグメント選択を行うことで 2^N のファクターを小さくしている。

このため提案するアルゴリズムの手順を概観すると以下のようになっている。

手順 1: 2^D 通りの特徴量選択ベクトル $\boldsymbol{d}$ の全組合わせについて手順 2 と手順 3 を繰り返す。

手順 2: $\boldsymbol{d}$ によって決定される部分特徴量空間においてセグメント選択方向に関する評価が極大となるセグメント選択ベクトル $\boldsymbol{r}_{\mathbf{d}}$ を探索する。

手順 3: 手順 2 で選ばれたセグメント選択ベクトル $\boldsymbol{r}_{\mathbf{d}}$ が、特徴量選択の変化に対して極大であれば Matchable 状況である。

手順 2 の内部における処理の詳細説明については省略するが、セグメント空間内においてある一部分だけを探索すれば十分で、それ以外には極大値が存在しないことが示せる。このような高速化により、手順 2 の処理量が 2^N から僅か $R_{\mathbf{d}}$ 通りまでに削減することができる。ここで $R_{\mathbf{d}}$ はある特徴量 d を選択した場合のセグメント選択空間において 1 つ以上のケースを保持しているセグメントの数であり、通常 2^N よりも遥かに小さい。このため全体の処理量は手順 1 の 2^D 毎に $R_{\mathbf{d}}$ のオーダーがかかるので、$O(\sum_{\mathbf{d}}^{\mathbf{D}} R_{\mathbf{d}})$ となる。

5 計算機実験と議論

状況分解の効果を示すシミュレーションを行った。例題では図 6 に示すように 3 次元特徴量空間において恣意的に二つの部分状況を組み合わせた全体状況を生成した。各部分状況では、平面上に 11 × 11 個のケースが 0.1 刻みの等間隔に配置される。状況 A は $(x + z = 1)$ 平面であり、状況 B は $(y + z = 1)$ 平面に対応する。

特徴量をセグメントに分割する方法は、各特徴量とも同様で [-0.05, 1.05] の区間を 0.1 刻みで 11 分割した。Matchability 基準の定数は $C_1 = 1.0, C_2 = 0.3, C_3 = 0.7$ とした。状況分解を行った結果を図 6 の右半分に示す。Matchable 状況 1 (MS_1) と 2 (MS_2) はそれぞれ状況 A と B に対応している。しかし、 Matchable 状況 3 (MS_3) はこの二つの部分状況の組み合わせで作られた新たな Matchable 状況であり、$x = y, x + z = 1$ の直線状に分布する。

状況分解による予測能力の向上を説明するために多価関数 $\phi : (x, y) \to z$ を考える。状況分解で得られた部分状況に MeBR を用いれば状況毎に分けて $\phi_{MS_1}(0, 1) = 1.0, \phi_{MS_2}(0, 1) = 0.0$ を出力するが、単純な MeBR では最悪の場合には二つの状況の平均値である $\phi(0, 1) = 0.5$ と出力するかもしれない。また、追加実験によれば状況 $A(x + z = 1)$ の半分が欠けていて $y > 0.5$ の領域にデータが無い場合でも MS_1 と同様な部分状況が抽出される。この場合、 MeBR では不可能な $\phi_{MS_1}(0, 1) = 1.0$ を得

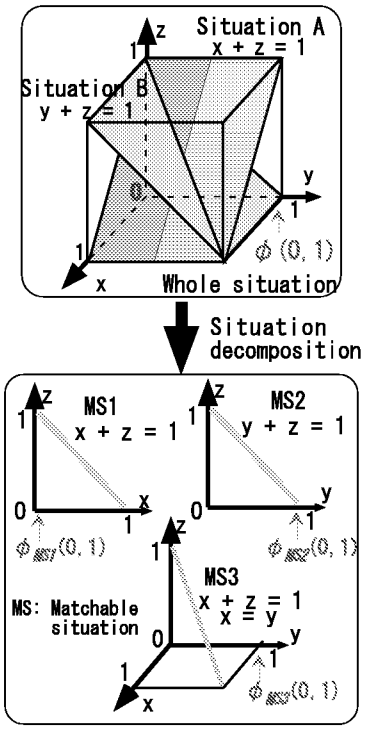

図 6: 計算機実験と結果: 二つの平面上にあるケースが状況分解される

る事ができる。このように状況分離能力や汎化能力において状況分解の効果がある。

6 まとめと今後の課題

Matchability という新しい評価基準を用いてケース（特徴量ベクトル）の集合である全体状況から、複数の規則性の高い Matchable 状況を抽出する手法を提案した。Matchable 状況を用いると、汎化により予測能力を高めることが期待できるので、データ解析、パターン認識、推論、行動決定などの前処理として利用できるだろう。

しかし、現状の Matchability 基準は直感的に導かれたもので、十分に理論的な検討が行われていないので、今後は予測性の観点からみた検討がますます必要であると思われる。また、計算量についても特徴量次元数の指数オーダーとなっているのでこれを改善することが望まれる。さらに、本稿では触れていないが、本技術をより使いやすくするために、抽出した Matchable 状況が過剰な一般化を行ってしまう問題や、欠損値データへの対応などの課題への対応を行っていくつもりである。一方、本技術を既存の予測手法 (たとえば線型回帰、ニューラルネット) などと組み合わせてそれらの予測能力を向上させる試みについては現在進行中である。

参考文献

[1] Langley, P., et al, "Rediscovering chemistry with the BACON system," *Machine Learning: An Artificial Intelligence Approach,* Michalski, Carbonell and Mitchell, pp. 307-329, 1983.

[2] http://www.trl.ibm.co.jp/projects/s7800/ DBmining/index.htm

[3] Akutsu, T. and Takasu, A, "Inferring Approximate Functional Dependencies from Example Data," *Proc. 1993 AAAI Workshop on Knowledge Discovery in Database,* pp. 138-152, 1993.

[4] Akaike, H., "A new look at the statistical model identification," *IEEE, Trans. Automat. Contr.,* vol. AC-19, pp. 716-723, 1974.

[5] Rissanen, J., "Universal coding, information, prediction and estimation," *IEEE Trans. on IT,* vol. IT-30, pp. 629-636, 1984.

[6] 山西健司, 韓太舜, "MDL 入門: 情報理論の立場から," **人工知能学会誌**, vol. 7, No. 3, pp. 427-434, 1992.

[7] ミンスキー, M. 著／安西　祐一郎 訳., "心の社会" **産業図書**, 1990.

[8] Suehiro, T., Takahashi, H. and Yamakawa, H. "Research on Real World Adaptable Autonomous Systems - Development of a Hand-to-Hand Robot," *Proc. 1997 Real World Computing Symposium,* pp. 398-405, 1997.