第3章 データマイニングの基本概念

データベースや情報検索に関連する基本技術であるデータマイニング(data mining)［福田ら 01］［Han ら 01］［Hand ら 01］［Witten ら 99］について説明しよう．データベースの検索も，情報検索も，どちらも条件を与えて，データの集合から条件を満足するデータを探索するという問題に対応している．一方，データマイニングは逆にデータの集合から，その性質を表すさまざまな構造・関係・規則を発見する問題に対応している．このとき発見された構造・関係・規則を，モデルやパターンということがある．その意味で，データマイニングはデータベースにおける知識発見(knowledge discovery in database，KDD)という，より大きな技術の文脈でとらえることができる．一般に知識発見は以下のような複数のステップからなる．

① データ洗浄：データベースからノイズや一貫性を壊すデータを取り除く．
② データ統合：必要に応じて複数のデータソースを統合する．
③ データ選択：データベースから分析対象となるデータを選択的に検索する．
④ データ変換：マイニングに適したデータ構造に変換する．
⑤ データマイニング：パターンの抽出のために知的な方式を用いる．
⑥ パターン評価：ある尺度(関心度)にしたがって本当に興味のあるパターンを同定する．
⑦ 知識表現：マイニングされた知識を効果的にユーザに提示する．

データマイニングはこの知識発見過程における本質的なステップということができる．

さらに，観測されるデータ(結果)をもとに，その現象の本質(原因)を探る学問は，一般に工学における逆問題と称される．その意味でデータマイニングは逆問題の一種であるといわれることがある．さらにデータマイニングは最近注目されている知識管理(ナレッジ・マネジメント，knowledge management)とも関連する．すなわち知識

管理は組織体（企業など）における暗黙知（言葉や規則ではっきりと示されていない知識）を形式知（言葉や規則で明確に説明できる知識）に変えることによって，組織体の構成員の間でそうした知識を共有できるようにすることが目的であり，その意味でデータマイニングを利用することが可能である．

たとえば，現実の店舗であれ，オンライン・ショッピングであれ，顧客は通常ある商品といっしょに別の商品を買うことがよくある．そこで，顧客がよく売れる商品を購入したときに，どんな商品をあわせてよく購入するかという規則性を分析して，その規則性にもとづいて関連する商品の売り場を近づけたり，関連商品を推薦したりすれば，売り上げを伸ばすことが期待できる．このような規則を相関ルール（アソシエーション・ルール，association rule）という．

また，特定の属性に注目して顧客をあらかじめ決められた有益なクラスにクラス分け（分類，classification）して，クラスごとにふさわしい商品のターゲット・マーケティング［石川 02］をしたり，似たような購買行動をとる顧客をグループ化（クラスタリング，clustering）してから特定のグループを絞りターゲット・マーケティングをしたり，顧客の時系列の購買履歴から将来購買する確率の高い商品を予測（時系列予測）してターゲット・マーケティングを行ったりする．このような分類やクラスタリング，時系列予測も，データマイニングの主要な機能の一つである．こうしたデータマイニング技術は，本書で以降に扱う Web，XML，P2P などのマイニングにも応用できる基本的技術である．次章以降で，相関ルール，分類，クラスタリングについて順に説明する．

参考文献

[福田ら 01] 福田剛志，森本康彦，徳山豪：
データマイニング
共立出版，2001.
[Han ら 01] Jiawei Han, and Micheline Kamber：
Data Mining: Concepts and Techniques
Morgan Kaufmann，August 2001.
[Hand ら 01] D. Hand, H. Mannila, and P. Smyth：
Principles of Data Mining
MIT Press, 2001.
[石川 02] 石川 博：
e–ビジネス技術入門教科書–ビジネスモデルと情報技術（IT） IT TEXT
CQ 出版社，2002.
[Witten ら 99] Ian H. Witten, and Eibe Frank：
Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations
Morgan Kaufmann，October 1999.

第4章 相関ルールのマイニング

動機／基本概念／相関ルールの種類／アプリオリ・アルゴリズム／相関ルールの生成／アプリオリ・アルゴリズムの効率化

4.1 動機

ここでは，まず相関ルール［福田ら 01］［Han ら 01］［Hand ら 01］［Witten ら 99］というものが，どのように有効かということから説明したい．さてスーパーマーケットで顧客は，いったいどんな商品を組み合わせてよく買うだろうか．それを分析するのがバスケット分析である．よく買われる商品の組み合わせの分析がいったんできると，以下のような利用法が考えられる．

- よく買われる商品の組み合わせについては，それに含まれる全商品をできるだけ近いところに配置して売り上げを向上させる．
- よく買われる商品の組み合わせは，それに含まれる全商品を最初からセットにして販売する．
- よく買われる商品の組み合わせに現れる各商品を互いに，できるだけ離れた売り場に配置し，その売り場間の移動中に他の商品を買ってもらう機会を増やす．
- よく買われる商品の組み合わせのうち，片方の商品をバーゲンにして買いやすくし，かたや残りの商品の利益率を上げて，トータルで利益があがるようにする．

ここでバスケットの中に入っている商品の集合をトランザクションという．そのとき個々の商品をアイテムという．

一言でいえば相関ルールとは，ある商品を買ったら，別の商品も一緒に買うという事実を規則（ルール）の形で表現したものである．もちろん，後述するように相関ルールの応用は，バスケット分析にとどまらない．Web コンテンツから，よく現れる単語の組み合わせパターンを発見したり，Web のアクセス履歴からよく現れるアクセス・パターンを発見したりと，相関ルールの利用範囲は急速に拡大している．この相関ルールのマイニング（相関分析ともいう）では，商品の組み合わせを計算することが必要に

なるが，一般にその数は多い．すなわち商品の種類を N とすれば，その組み合わせは 2^N となるからである．そこでよく現れる，すなわち頻出するアイテムの組み合わせをいかに効率的に見出すかが，よいマイニング・アルゴリズムかどうかのポイントになる．

4.2 基本概念

ここではまず相関ルールのマイニングに必要となる基本概念を説明しておこう．個々の商品（アイテム，item）を i_k とすると全アイテムの集合は $I = \{i_1, i_2, \cdots, i_n\}$ と表せる．ただしアイテムはショッピングにおける商品にかぎらず，より一般的な概念とする．各トランザクションを T とすると，T はアイテムの集合であり，I に含まれる（$T \subseteq I$）．個々のトランザクションにはトランザクション識別子 TID が関連づけられ，TID によって唯一に識別される．そして今マイニングの対象として考えるデータベースを D とすると，D はトランザクションの集合になる．

以上の概念を用いると相関ルールはつぎのように表現できる．

（定義）相関ルール

$A \Rightarrow B$（A ならば B であると読む）

ここで A，$B \subset I$ かつ $A \cap B = \phi$ である．

つぎに相関ルールに付随する，サポート（support，支持度）とコンフィデンス（confidence，確信度）という概念を定義する．サポート s とはデータベース D においてアイテム集合 A と B をともに含む（そのようなアイテム集合は $A \cup B$ と表せる；$A \cap B$ ではないことに注意）トランザクションの割合である．一方コンフィデンス c は D において A を含むトランザクションに占める，B を含むトランザクションの割合である．ここで確率 P を用いればサポートとコンフィデンスはつぎのようにいいかえられる．

（定義）相関ルールのサポートとコンフィデンス

サポート $(A \Rightarrow B) \equiv P(A \cup B)$

コンフィデンス $(A \Rightarrow B) \equiv P(B|A)$

ここに $P(B|A)$ は条件付き確率を表す．すなわち A が起こったという付帯条件のもとで B が起こる確率である．

＜余談＞相関

相関ルールという用語は，もちろん association rule の訳語である．統計学における相関係数といったら correlation coefficient の訳語である．どちらも相関と訳されるので，混乱を生じるかもしれない．しかしすでに定着した言葉なので，本書でもこれを用いる．

つぎにルールの強さについて考えよう．最小サポート（min_sup）と最小コンフィデンス（min_conf）が与えられたとき，最小サポート以上のサポートを持ち，最小コンフィデンス以上のコンフィデンスを持つルールを強いルールと呼ぶことにする．

アイテムの集合をあらためてアイテムセットと呼ぶことにする．k 個のアイテムからなるアイテムセットをとくに k-アイテムセットと呼ぼう．アイテムセットの出現頻度は，そのアイテムセットを含むようなトランザクションの個数である．出現頻度は，サポート・カウント（support count）とも呼ばれる．

アイテムセットの出現頻度が，最小サポート $\times|D|$（これを最小サポート・カウントという；ただし一般に $|S|$ は集合 S の要素数を表す）以上であれば，そのアイテムセットは最小サポートを満足するという．

最小サポートを満足するアイテムセットのことを頻出アイテムセット（またはラージ・アイテムセット）といい，通常 L_k と記述する（k はアイテムの個数を表す）．アイテムセット A の出現頻度が support_count (A) で表されるとすると前述したサポートとコンフィデンスはあらためて以下のように定義できる．

（定義）相関ルールのサポートとコンフィデンス（再）

サポート $(A \Rightarrow B) \equiv P(A \cup B) = support\_count(A \cup B)/|D| \cdots$ support 式

コンフィデンス $(A \Rightarrow B) \equiv P(B|A) = support\_count(A \cup B)/support\_count(A)$ $\cdots$ confidence 式

たとえば，相関ルールは以下のように表せる．

赤ワイン（買う）$\Rightarrow$ チーズ（買う）［サポート＝ 3 %，コンフィデンス＝ 50 %］

以上で概念の定義はそろった．つぎに相関ルールのマイニングは何をすることなのか，説明する．すなわち相関ルールのマイニングはつぎの二つのステップからなる．

① すべての頻出アイテムを発見する．

② 頻出アイテムセットから強いルールを生成する．

二つのステップでは，第一ステップのほうが計算の手間が大きく，後述するようにマイニング全体の計算量に関する性能改善の努力は，もっぱら第一ステップに注がれることになる．

4.3　相関ルールの種類

いままでアイテム（商品）を対象とした相関ルールについて述べてきた．相関ルールといっても，いくつかの種類がある．一般には相関ルールは以下のような分類基準にもとづいて，複数の分類方法が考えられる．

(1) 値の型による分類

すでに述べたアイテムの有無のような 2 値で表現される属性に関する相関ルールはブーリアン型(2 値型)の相関ルールという．たとえば，

赤ワイン(買う) ⇒ チーズ(買う)

一方，数量的に表される属性間の相関に関するルールを数量型の相関ルールという．数量型の属性は区間に分割して考えるのが普通である．数量型の属性を，区間にマッピングして表現することを離散化するという．たとえば，

25 < 年齢 < 30 ⇒ 白ワイン(買う)

(2) 次元による分類

ルールに現れる属性の種類(それをここでは次元という)によりルールを分類できる．たとえば，

赤ワイン(買う)⇒ チーズ(買う)

は一次元のルールの例であり，

25 < 年齢 < 30 ⇒ 白ワイン(買う)

は多次元(二次元)のルールの例である．

(3) 抽象度による分類

複数のルールの集まり(ルールセットという)を考える．そのときルールに含まれるアイテム間に階層関係が考えられる場合に，それらのルールセットはマルチレベルのルールという．

たとえば以下のようなルールセットは，マルチレベルである．

25 < 年齢 < 45 ⇒ ワイン(買う)

30 < 年齢 < 45 ⇒ 赤ワイン(買う)

そうでない場合をシングルレベルのルールという．

4.4 アプリオリ・アルゴリズム

相関ルールのマイニング・アルゴリズムの基本を理解するために Agrawal らのアプリオリ(Apriori)[Agrawal ら 93] と呼ばれる頻出アイテムセットを求めるアルゴリズムを紹介する．ここではもっとも単純であるブーリアン型で，一次元で，シングルレベルの相関ルールに絞って考えることにする．それ以外の場合については，基本的なアルゴリズムに対する拡張として考えていくことにする．

頻出アイテムセットには，以下に述べるような性質がなりたつ．

(**命題**)性質アプリオリ：頻出アイテムの空ではないすべてのサブセットは，頻出である．