要因分析のためのデータマイニング

5E－1

沼尾雅之　清水周一　木村雅彦
日本アイ・ビー・エム株式会社 東京基礎研究所

1．はじめに

大福帳システム、データベースからの仮説検証など、データベースを積極的にマーケティング戦略に利用していこうとする動きが、流通業を中心として盛んになっている。その中でも、データマイニングは、データベース中から未知の規則を発見できる、仮説生成型アプリケーションとして注目されている。

現在、データマイニングの手法として知られているものを大別すると、以下のようになる。

(1) クラス分類型:パターン認識や学習といわれているもので、クラスのわかっているサンプルデータを訓練例として、クラスを分類するための式、ルール、決定木等を生成するものである。ニューラルネットもここに入れられる。応用としては、顧客の信用調査、不正検出、ポートフォリオマネージメント等があげられる。

(2) クラスタ分割型：属性間の距離などを規準にして、似かよっている属性を持つデータをグループ化するものであり、統計的クラスタリングや整数プログラミングの手法などが知られている。応用としては、顧客のプロトタイピング、バスケットアナリシス等があげあれる。

(3) 演繹データベース検索型：データベースから新しいパターンを導出し、これを数えあげることによって、そのパターンの有効性を検証する。応用としては、関連購買分析があげられる。

(4) 視覚化型：データをわかり易く表示し、対話的にデータの絞り込みなどの操作することによって、データ中の変数間の関係を明確化するものであり、基本的にルールの発見は人に任されている。

これらの手法の中でも、演繹データベース検索型は、扱えるデータのサイズ、柔軟なパターン、および、解の完全性などの点で優れており、大規模データベースに対してもスケーラブルなアルゴリズムが開発されている [1]。この手法は、直接的には流通業において、どの商品とどの商品が一緒に買われたかという関連購買分析に有効であることが知られているが、本稿では、これを一般の要因分析に応用することを示す。POSデータ等の大量の購買履歴と、その時間、場所的な背景データを同時に処理することによって、購買の要因のみならず、要因間の因果関係なども抽出することができるようになる。

2．関連分析

まず、関連購買分析について説明する。対象とするデータベースは、スーパーストアのPOSで得られるようなトランザクションデータ、

$$T = (transaction - id, items)$$

の集合である。ここで、itemsは、商品項目全体の集合

$$I = \{i_1, i_2, ..., i_m\}$$

の部分集合になる。さて、商品項目の集合 X と Y の間の相関関係

$$X \Longrightarrow Y \text{ ただし、}$$
$$X \cap Y = \emptyset$$

を次のように定義する。$X \cup Y$が、データベースのs%のトランザクションに含まれていて、さらにXを含むトランザクションのc%が、Yを含んでいる。ここで、あるトランザクションが、項目集合Xを含むというのは、

$$X \subseteq T.items$$

とする。ここで、s を支持度、c を確信度と呼ぶ。

文献 [1] では、最小支持度及び最小確信度を与えることによって、データベース中から、それらを満たすような全ての相関関係を導出するアルゴリズムが提案されている。これを実際にPOSデータに適用すると、

$$s = 0.01, c = 30.0 :$$
$$\{\text{パン、バター}\} \Longrightarrow \{\text{牛乳}\}$$

というようなパターンのリストが得られる。

Datamining for Causal Analysis.
Masayuki NUMAO, Shuichi SHIMIZU, Masahiko KIMURA.
IBM Research, Tokyo Research Laboratory

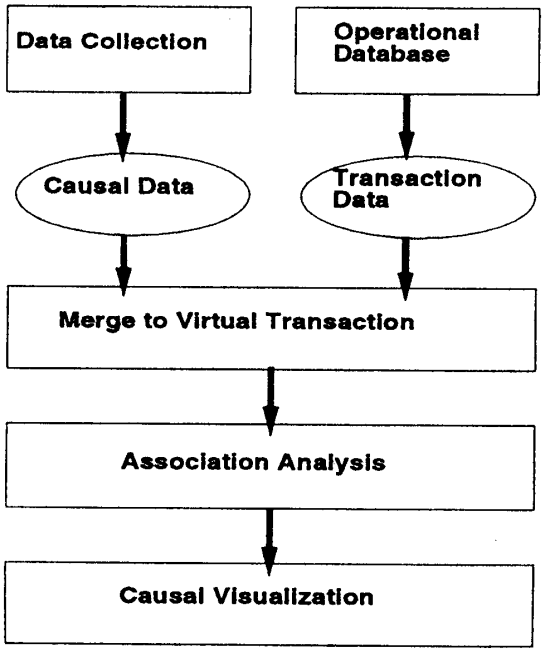

図 1: 関連分析による要因分析

3．演繹データベース検索による要因分析

図1に、前節の関連分析を要因分析に応用する方法を示す。POSデータ単体では、購買日時及び購買品目ぐらいしか入っていないが、これを他のデータベースと組み合わせることによって、さまざまな背景データを得ることができる。例えば、購買日時からは、曜日、時間帯、その日の天気等がわかるし、もし、クレジットカード等での購買が主であるならば、その会員属性データから、顧客の年齢、性別、職業等を調べることもできる。こうして得た属性データを、仮想トランザクション化し、実トランザクションとマージすることによって、属性と商品との間の相関関係、つまり要因分析ができるようになる。

3．1 要因データの収集

要因データとしては、トランザクションデータと関連付けられるものならば、なんでも可能である。そのトランザクションの起こった、時間、場所、当事者の情報はもちろん、2次的なものとしては、時間と場所から天気が導けるし、そのときに起こっていたイベントたとえば、安売りなどのキャンペーンも要因となる。これらを効果的に収集するには、外部データベース、たとえばインターネットなどを利用することが考えられる。

3．2 仮想トランザクションの生成

要因データは、仮想アイテムとして実トランザクションに加えられ、仮想トランザクションになる。つまり、

$$T = (transaction - id, extended - items)$$

ただし、$extended - items$ は、商品項目集合と要因集合の和

$$I \cup C$$

の部分集合になる。

3．3 頻出集合の導出

ここで、文献[1]のアルゴリズムを用いて、頻出 $extended - items$ 集合を求める。これは、データベースの s%の仮想トランザクションに含まれているような $extended - items$ の集合である。このとき、それぞれの集合について、そのデータベース中での頻度、すなわち支持度が同時に得られる。ここで、得られた集合は、要因集合 $X \subset C$ と商品項目集合 $Y \subset I$ から構成されているので、それを分離し、

$$X \Longrightarrow Y$$

を作って、集合全体の支持度及び要因集合だけの支持度から、上記パターンの確信度を求めることができる。

3．4 要因分析結果の視覚化

得られたパターンのリスト中には、右辺および左辺に同じ集合を持つものが、多数あらわれる。したがって、同じものをまとめたり、また、数値属性中には年齢のように、20代、30代に分けたが、結果としては、

$$\{A20\} \Longrightarrow Y$$
$$\{A30\} \Longrightarrow Y$$

が得られた時のように、共通の結果を示す場合もある。視覚化ツールは、このようなパターンをグラフで表示し、また、ノードのマージなどの操作ができるようにし、因果関係が理解できるようにしている。

参考文献

[1] R.Agrawal and R.Srikant, "Fast Algorithms for Mining Association Rules" *Proc.of the 20th VLDB Conference, 1994.*