大規模多品種購買データに対するデータマイニング

3X－1

Data mining on a large-scale and highly diversified purchase data

全　眞嬉*§　藤井　章博†∫　徳山　豪‡§
東北大学大学院情報科学研究科§　宮城大学事業構想学部∫

1. 要旨

本研究ではデータマイニング手法実装のためのプラットフォームとして、データの正規化と視覚化のためのシステムツールの開発し、実データによる実験を行った。実験においては、価格変動の多い食材購買データを対象にし、多変量分析、クラスター分析、結合ルールの既存の分析手法を上記プラットフォーム上で用いる。データとして、10箇所の支店を持っているホテルチェーンのレストランの食材購買データを対象とし、価格変動に対して正しい購入価格変動であったかどうかの判断支援を目的とした。対象データによって分析手法の異なった組み合わせを適用させ、不正購買取引の検出、価格変動の原因分析を行う。

2. データの概要

扱う商品の数は3,000種類以上であり、全施設で取引される購買の件数は1ヶ月で約12万件である。過去3年分のデータが蓄積されており、データの大きさは約600MBである。ホテルレストランの購買データであるため、食材によっては単価が高く、多品種で少量の仕入れを頻繁に行うという傾向がある。又、魚介類、青果類等の価格変動が頻繁な商品が多く取引されているのが特徴である。

購買活動において、価格変動が実際の値上がりであるか、不正による価格変動、つまり異常値であるかどうかの判断は、経理管理上重要である。一般的に魚介類、青果類等の価格変動の要因は数多く、原因と結果の時期に隔たりがあるため、価格変動原因をはっきり特定し、指標として表すのは困難である。従って、不正取引を発見しにくく、データマイニングによる発見支援が有効と考えられる。一方、これらの特徴をデータマイニングの観点から見ると、属性数が多く、属性間の依存関係が複雑になるため、知識ルールの生成や知識抽出は学術的に興味ある問題となっている。

* Jinhee Chun
† Akihiro Fujii
‡ Takeshi Tokuyama
§ Graduate School of Information Sciences Tohoku University
∫ School of Project Design Miyagi University

3. システム

膨大な多品種購買データから知識獲得のために、データマイニングシステムツールの開発を行った。

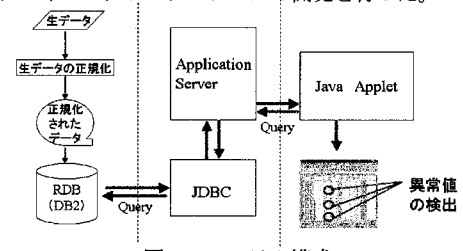

図1．システム構成

システムは3層クライアントサーバシステムのロジックでありJAVA言語で実装した。本システムの構成は生データの入力から、出力までの一連の操作を管理する。データマイニングのためのツールとして、多様なデータマイニング手法をフレキシブルに実装し、複合利用するためのプラットフォームを実現している。

4. 実験

（1）食材の購入時に生じる不正取引の検出、（2）不要な発注の検出、（3）レストラン利用率変動原因分析の3つの目的に関する知識の検出にシステムを用い、購買時の従業員の不正防止、レストラン利用率変動の要因把握という経営戦略上の意思決定支援を想定し、結果としての対象企業の業務改善を目標とする。

実験概要: 多品種のデータに対し実験を行ったが、特に伊勢海老の購買に対する実験結果を報告する。過去2年間の単価を実験対象データとし、単価の変動パターン、変動要因、施設と仕入れ業者との関係を分析する。実験方法は次の分析1、分析2、分析3の順に行う。

分析1: 影響と要因の関係が把握できれば、単価予測が可能になり、異常値の検出も容易になる。そのため、まず多変量分析を利用する。単価変動に影響を与えている主要因を把握するため、施設、仕入れ業者、年、

月を要因とした重回帰分析を行った結果、重相関係数が0.93で図2に示した影響度指数が分かった。伊勢海老の価格変動とその主要因は次の通りである。

①月によって単価が大きく影響され、3月から8月にかけて単価が上昇傾向である。図3は影響度指数の一番大きい月の効果をあらわしている。この図からは8月からは12月までは単価にマイナス影響を与えていることがわかる。

②施設によって仕入れ単価が大きく変動する。この結果は購買をする施設によって仕入れ価格が異なることを意味する。施設によって購買単価の若干の差違は誤差としてあり得るが，2,652円の影響度指数は大きいと言える。この結果に着目し、施設によって単価がどのように変動するか、変動のパターンを見つけるためにクラスター分析を行う。結果は分析2で示す。

③前年度よりは平均単価が1,000円くらい値上がりの変動であることが分かった。

④仕入量の影響度指数は182円でほとんど影響を与えていないことがわかった。

又、過去2年間の実観測データをもとに1999年の単価予測（予測値単価）を作成し、実際観測された1999年の単価（観測値）と比較してみた。観測値と予測値との差から、図2で示した要因以外の他要因が今年の伊勢海老単価に影響を与えた可能性があることが分かった。この残差に注目し、他要因を重回帰分析に加えて影響度を算出する。

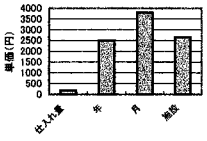

図2．影響度指数

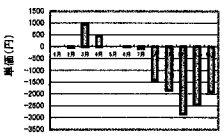

図3．月の効果

分析2: 多変量分析から施設の影響が大きいことから、施設の購買パターンについて分析を行う。類似しているデータをグループ化する為に適しているクラスター分析を行い、視覚的に判りやすい散布図を作成した。対象データは、分析1において伊勢海老の価格変動が大きかった施設の、レストランの宿泊客利用率と、その時の単価である。クラスター分析からは高い単価で購入する施設と、レストラン利用率が低いながら、高い単価で数量も多く仕入行っている施設が分かった。例えば、伊勢海老の価格の変動が多かったA施設において宿泊客数とレストラン利用客数の変化を散布図にしてみると、宿泊数が増加したにもかかわらず、レストラン利用客が減少する時期がある事が判った。この結果から推察される原因として、従業員の発注不正の可能性が考えられる。

分析3: さらにデータ固有の知識を抽出するため、多変量解析やクラスター分析を行って得られた要因属性間の結合ルールを求め、マイニング手法として適用してみた。結果の例としては、『A業者から冷凍イカを週3回以上購入する支店は伊勢海老をA業者から他業者より30％以上高値で購入をしている。（サポート：2％、確信度：73％）』ことが分かった。結合ルールより抽出された知識は確信度とサポートが分かるため意思決定をする時にルールの価値判断の基準と成る。

5. まとめ

本研究対象の食材購買データは、生のデータを人間が検査する場合には異常値の判断が困難であり、食材購買データに対する専門家知識が必要である。本研究において次の3点を主に取り上げ実験を行った。①食材の仕入れの異常取引検出②不要な発注を検出③レストラン利用率の変動原因分析における知識検出の3点において、知識を検出することができた。検出された知識により、購買時の従業員の不正の可能性を抽出することができた。

本研究では、開発したプラットフォームの上で、従来の分析手法を適用し、対象データによって異なった分析手法の組み合わせを用いて知識のルールを生成した。さらに有効なデータマイニングシステムの開発のためには、抽出された知識が意味のある知識であるかどうかの判断基準を定める必要がある。抽出された知識ルールを学習アルゴリズムに移植、適用することにより困難であった知識抽出を可能にし、戦略的意思決定支援処理の能力を高めることを今後の課題とする。

参考文献

[1]P.Cabena,Phadjinian,R Stadler,J Verhees and A Zanasi,『Discovering Data Mining from Concept to Implementation』, IBM Corporation 1998