【論　文】

中小企業のためのデータマイニングシステムに関する研究

董　　彦文

概要：本論文では，ある学生服卸販売会社を具体的事例として，中小企業へのデータマイニング手法の応用を考究する。まず，データマイニング手法を中小企業へ応用する必要性およびこれに際しての課題を述べて，中小企業のためのデータマイニングシステムの一般構成を提案する。次に適用可能な業務分野および適用可能なデータマイニング手法について検討し，データマイニングの適用に関するガイドラインを与える。さらに，顧客信用評価問題を具体例として事例ベース推論という人工知能手法の適用を提案する。

キーワード：データマイニング，人工知能，信用評価，システム設計，事例ベース推論

1. はじめに

急速な経済環境の変化にともない，国内市場の過剰な競争や価格破壊が進み，業績の落ち込みが激しい企業が多いなか，様々な業種で極端に成功している企業がある。このため，従来のビジネス慣行の有効性が低下するとともに，新しいビジネスシステム，組織の必要性が認識されるようになっている。メーカや小売店の顧客へのアプローチも，そうした流れの中で，大きく変わりつつある。従来のマスマーケットを対象にしたマーケティングは，十分にその効力を発揮しなくなっており，顧客の要望を正確に知り，それらに適切に応え続けることで，顧客との強い関係を構築・維持し続けて，利益の最大化を図ることが重要であると認識されてきて，顧客獲得・需要創造から顧客維持へシフトするのが今日のマーケティングの新潮流となっている。顧客知識を活用し，低コストで個別の顧客ニーズに対応することによって，常客との関係を確立し，顧客を満足させる手法，技術が求められている。

また，大容量記憶媒体とデータベース管理システムの低価格化，計算機処理能力の向上，インターネットに代表される情報通信技術の急激な進展につれて，データ収集が容易になったため，高効率・低コストで大規模データベースが多くの企業で構築され，データベースに蓄積されたデータ量が増え続ける反面，これらの莫大なデータから有用な情報と知識を見つけることが非常に困難になり，さらにこれが企業の経営戦略または経営活動に利用されているかというと，大きな疑問が残る。

近年来，蓄積されるデータの有効利用を目的とするデータマイニングが注目を集めている。データマイニングの実体は，特定のツールや技術ではなく，人工知能，数理統計学，データベース技術などの要素技術を使った総合的な問題解決システムということができる。これまでにデータマイニ

ングは，すでにいろいろな分野と様々な企業で適用され，多くの成功事例が伝えられている。

しかしながら，これまでの成功事例の多くは，関連パッケージソフト販売企業または大手企業により紹介され，適用企業の生データが公開できないため各事例の詳細は見えない。また，個別事例に対して高度なデータマイニング手法の開発と提案に関する研究は多いが，如何に多種多様な手法から対象企業と業界に一番適切なものを選ぶべきか，または各種の手法を如何にうまく組み合わせて特定の課題解決に適用すべきであるか，などのように，データマイニング手法の適用方法論に関する研究はまだ少ない。さらに，中小企業への適用と成功事例に関する報告は非常に少ない。そこで本研究では，ある中小規模の学生服卸販売会社を具体的事例として，中小企業へのデータマイニングの応用を考究する。中小企業におけるデータマイニングの必要性と課題を明確にしたうえ，中小企業におけるデータマイニングの適用分野を説明し，中小企業向けのデータマイニングシステムの構成を提案する。さらに，顧客信用評価問題を応用例として，マイニング結果の解釈と活用における人工知能手法の適用について詳しく検討する。本論文の目的は，中小企業へのデータマイニングの適用に関するガイドライン，および中小企業向けのデータマイニングシステムの一般的枠組を与え，中小企業へのデータマイニングの応用に関する研究基礎を構築することである。

2. データマイニング概要

2.1 データマイニングとは

(1) データマイニングの定義

簡単に言えば，データマイニングとは多様で大量の蓄積データから有用な知識・情報を抽出または掘り出すことである。しかし，データマイニングの定義に関しては統一された見解がまだ存在しない。研究者および実務者の視点や立場によって，その主張が大きく異なり，現実にはいろいろなデータマイニングの定義が与えられている。

㈱SASインスティチュートジャパン［1］はデータマイニングの目的と手段の視点から次の定義を与えている：

「データマイニングとは，意味のあるパターンやルールを発見するために，大容量のデータを自動的もしくは半自動的手段で分析し探索することである。」

P. Cabenaら［2］はビジネス分野におけるデータマイニングの役割から，次のようにデータマイニングを定義している：

「データマイニングとは，以前には知られていなかった，有効で，活用できる情報を大規模データベースから抽出し，その情報を用いて重要なビジネス上の決定を行うプロセスのことである。」

また，D. Handら［3］はデータに対する処理の視点から次の定義を与えた：

「データマイニングとは，予知できなかった（因果）関係を発見し，またはデータ所有者にとって理解可能でかつ適用可能な手法を用いてデータを整理するために，観察データの集合を解析することである。」

このほかにも，様々なデータマイニングの定義が与えられている。しかし，これらの見解の違い

はビジネスと他の分野におけるデータマイニングの有用性とは別の問題であり，定義上の異同を避けて，データマイニングの有用性に着目して，データマイニングを議論するのが多い。

(2) データマイニングの手順と KDD (Knowledge Discovery in Database)

データマイニングは非常に大きな労力を要する作業であり，図1に示すようにその一般的手順は次のとおりである [2]。

[**ステップ1 ビジネス目標の決定**] 確実にビジネスに役立ち，データマイニングを成功させるには，まずビジネス上の問題または課題を明確に定義する必要がある。これは，どんなデータマイニング・プロジェクトでも不可欠なステップであり，省くことはできない。

[**ステップ2 データの準備**]

[ステップ2.1 データの選択] ビジネス目標の達成と課題の解決に必要な内部と外部情報を集める。大量のデータがすでにデータベースに蓄積されているため，データの収集はそれほど難しくないが，必要なデータソースを識別し，本当に目的に合ったデータを選択することは意外にも簡単ではない。

[ステップ2.2 データの前処理] データマイニング作業を容易にするため，目標データからノイズや異常値を除去し，データをクリーンなものにする。どれがノイズなのか異常値なのかを判断するには，対象問題に関する知識が必要であり，またマイニング手法の選択にも関連している。

[ステップ2.3 データの変換] これから行う分析，およびデータマイニング・アルゴリズムに必要なデータ形式に合わせて，データを変換する。数値データしか対処できないデータマイニング手法を適用するには，なんらかのルールを用いてカテゴリデータやテキストデータを数値化する必要がある。また，場合によっては元のデータを正規化してはじめて価値のある情報を見つけ出すことがある。

[**ステップ3 データマイニング**] ステップ2.3で変換したデータを用いてマイニングを行い，興味のあるパターンまたは価値のある情報・知識を抽出する。これはデータマイニングプロセスの中核をなすステップであり，適切なデータマイニング・アルゴリズムの組み合わせを選択することを除けば，この処理は高速であり自動化されている。

[**ステップ4 結果の分析**] ステップ3からの出力について解釈および評価を行う。ここでは，支持度や信頼度などの評価指標を計算し，様々な分析方法を用いて定量的に評価するのはよく採用される一般的やり方である。現実問題の解決においては，マイニング結果を理解しやすいグラフな

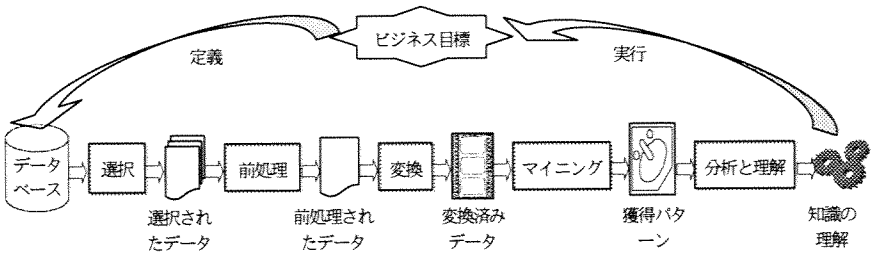

図1. データマイニングの手順と KDD

どの形式で表示する視覚化手法もよく使われ，これは業務担当者の豊富な経験・知識を活かせる大切な手法である。

［**ステップ5　知識の理解**］　ステップ4で得られたビジネス上の知識と情報は，企業のビジネス活動および情報システムに反映させ，経営などにかかわる意思決定に活用する。

データマイニング処理のうち，データの準備がデータマイニング作業量全体の8割を占めるとされている。

図1に示した手順については，データマイニングの少し前の時期に，データベースの知識発見KDD（Knowledge Discovery in Database）という用語が生まれていた。1995年にモントリオールで開催された第1回KDD国際会議では，KDDは「生データからパターンや類似性などの知識を抽出するプロセス全体を意味し，データマイニングはKDDのプロセスの特定のステップ，つまり図1の［マイニング］ステップだけに使われる技術である」と承認された。このため，いまでもKDDのプロセスの知識発見ステップに使われる技術としてデータマイニングを定義する学者が少なくない［4］。最近，データマイニングツールの開発販売ベンダーやこの分野の人気業界誌の関心が高まった結果，データマイニングという用語は，KDDと同じようにデータベースから知識を抽出するプロセス全体を意味するように変えられてしまい，データマイニングとKDDを同義語として扱うのは一般的となり，データマイニングとKDDを区別せず，使用することになってきている。

(3)　データマイニングとビジネス・インテリジェンス

データマイニングは大規模データベースから有用な知識または情報を抽出するシステムであるため，近年来急速に注目されてきたデータウェアハウス/データマートなどの技術・システムとは密接な関連をもつ。

また，IT（情報技術）に基づくビジネスの意思決定を支援するすべてのプロセス，手法，およびツールを表す包括的な用語として，最近ビジネス・インテリジェンスという言葉が使用されてきて，これの方法には，単純なスプレッド・シートから競争に耐えうる大きなインテリジェンス処理にいたるまで広範囲にわたる。データマイニングは，ビジネス・インテリジェンスの重要な新規構成要素となっている。図2に示すように，戦術的および戦略的なビジネス上の意思決定基盤となる可能性およびビジネス活動における利用者の立場に基づいて，種々のビジネス・インテリジェンス・テクノロジーの論理的な位置付けがそれぞれ異なり，ここからデータマイニングと他の関連技術・システムとの相違がわかる［2］。

(4)　データマイニングと統計解析

データマイニングでは，重要な手法として統計解析がよく使われている。これに加えて，統計解析パッケージソフトウェアの販売会社がほとんど独自なデータマイニングツールを開発し，または元の統計解析パッケージにデータマイニング機能を追加し発売しているため，データマイニングを統計解析の一部であるとする考えも存在し，データマイニングを統計解析と混同し区別しない研究者もよく見られる。

しかしながら，統計解析と根本的に異なる点は，その考え方または出発点である。統計解析が「仮説の証明」を分析目的としているのに対し，データマイニングは「問題の特定と原因・構造追求」を分析目的としている。つまり，統計解析は，まず「仮説ありき」からスタートしており，その仮説

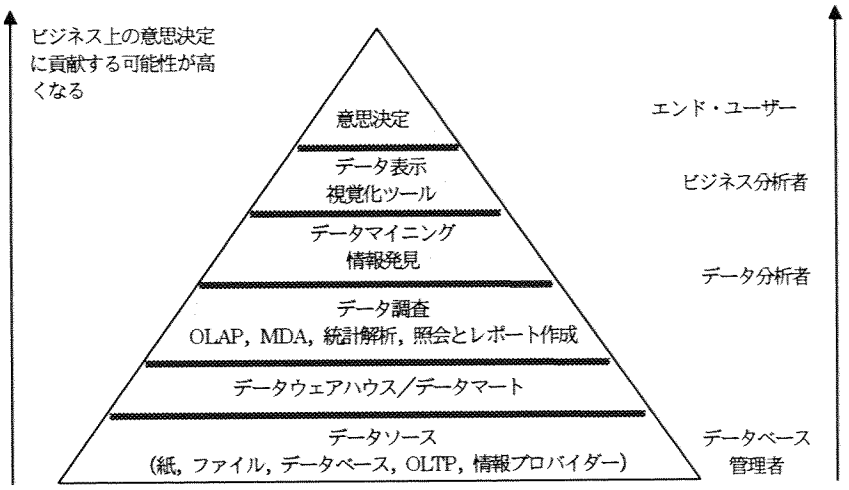

図2. ビジネス・インテリジェンスにおけるデータマイニングの位置づけ

表1. データマイニングと統計解析，多次元データ分析との相違

	多次元データ分析 (MDA, OLAP)	データマイニング	統計解析
主な分析対象データ	履歴データ	履歴データ	実験・調査データ
分析データ量	大	大	小〜中
分析目的	現状把握 (モニタリング)	問題の特定と原因・構造追求 (発見と検証)	仮説検証 (記述と推測)
実行モード	対話型	半自動	対話型
分析プロセスの特徴	人間の判断によりあらかじめ重要だと想定されるパターンを定義する	大量の生データから出発し，そこからパターンを発見，検証していく	ある仮説の下，必要な情報を実験計画法などに基づき収集，分析する
難易度	結果は集計なので非常に簡単。手法に関する特別な知識は不要	分析プロセスが半自動化され簡易に実行。各技術の基本的知識が必要	分析プロセスが対話型であり，各手法の深い知識が必要
最終的結論	より主観的 解釈にばらつき	より客観的 知識の共有化	より客観的 知識の共有化

を検証し証明するためにデータは収集され，分析される。仮説であるモデル式はデータの状態に左右されることなく，統計的に証明できなければ，モデル式は意味のないものとなってしまう。

これに対してデータマイニングは，「データありき」からスタートする。つまり，データは存在するがそこに何が埋もれているかわからない状態から分析作業が始まるのである。そのため，分析前に証明したいモデル式があるはずもなく，仮説に沿ったデータを収集することも，実験を計画することもない。

分析対象とするデータおよび分析プロセスの特徴などの点から，データマイニングと統計解析，多次元データ分析MDA（Multi Dimensional Analysis）およびオンライン解析処理OLAP（On Line Analytical Processing）との相違を表1にまとめる［1］。

2.2 産業界におけるデータマイニングの応用

理論的・技術的にはまだ大きな課題が残っているにもかかわらず，データマイニングに対する期待は大きく，産業界では，すでに統計解析や決定木，ニューラルネットワーク，相関規則などの機械学習・データマイニング技術を使って，実用的問題の解決に取り組んでいる。元田ら［5］のサーベイを参考にして，産業界におけるデータマイニングの応用分野と適用事例を表2に示す。

これまでの公開資料によると，金融分野での適用事例が最も多く見られる。特に米国では，1994年頃から流通業や金融業でデータマイニングの事例が報告され，日本国内でも近年は多くの事例が報告されるようになった。金融分野では，業務の効率と質の改善効果を求めるために，膨大な顧客リストからもっとも有望な顧客候補を見出して，もっとも適切な商品と商品組合せを提示したり，生命保険の潜在的解約候補顧客や効果的なダイレクトメール宛先候補顧客を決定したりする際，

表2. 産業界におけるデータマイニングの適用事例

金融分野	・マーケティング分野 潜在的な住宅ローン申込み顧客の推定 顧客に応じた銀行商品の適切な組合せ（クロスセールス）の設計提示支援 生命保険の潜在的解約候補顧客の発掘 効果的なダイレクトメール宛先候補顧客の発掘
	・業務特化分野 消費者ローン与信審査の半無人化ルールの発掘 顧客に応じたリスク細分型の自動車保険の設計提示支援 証券顧客と営業マンとのトラブル予測 社債格付け推測，クレジット・カードの不正利用パターン推定
流通小売分野	薬局チェーンおよびコンビニ販売データからの優良顧客の発掘
	投入時立ち上がり売れ行きデータに基づく新製品販売予測
	新製品のヒット要因分析，品物の売行き要因分析，牛乳販売量の予測
	消費者購買行動パターンの分析，種々の販促条件下における併売パターンの分析
	商品の在庫管理および発注量の予測
製造分野	ホームページでの顧客意見収集による次世代新製品開発 （カスタマーリレーションマーケティング）
	顧客の製品クレーム情報と製造情報の突合せによる設計・製造現場への品質管理要求発掘
	製造現場の製造条件と製品検査結果の突合せによる製造工程の改善
情報通信分野	テキストマイニングによるアンケート結果・掲示板書き込みおよび電子日報内容の分析
	ホームページ閲覧情報からの個別顧客のプロファイリングと顧客傾向分析
	電話回線網管理のための負荷状況把握や障害診断
	電話網使用需要マーケティングのための通信トラフィックデータ分析
	顧客の通話パターンによる通話回線不正使用検出
	計算機システムへのアクセスログに基づく不正アクセス検出

ニューラルネットワーク，コホーネンネット，クラスタリング，決定木，ラフ集合，重回帰分析など，多様なデータマイニング技術が用いられている。また，与信審査半無人化ルールの適用による消費者ローン無人申込み機の開発や，膨大なクレジット・カード使用記録からの不正利用パターン発掘などの業務特化分野では，データマイニング手法が適用され実績を上げている。流通小売分野では，POS データを用いた優良顧客の発掘や各種商品販売パターンの発掘・分析，インストアでの販売促進用知識の導出，有望顧客の洗出しなどの面で主に小売部門のマーケティングのためデータマイニングが利用されている。金融分野に比べて，流通小売業では，扱う商品や小売条件，顧客行動パターンがはるかに多様であり，ポイントカードなどを導入しても顧客を識別することが非常に困難なため，顧客や購買事例を明確に区別して特徴を発掘することは難しい。

製造分野におけるデータマイニングについては，社内の文書やマニュアル検索などへの適用が多く，品質管理や工程管理への適用事例が報告され，実用化が進められている。手法として，事例ベース検索，テキストマイニング，バスケット分析と決定木などの最新のデータマイニング技術が用いられ，効果を上げている。さらに，主要電機メーカ，家電メーカは，カスタマーリレーションマーケティングへデータマイニングの適用を試みて，ホームページで収集した顧客意見を次世代新製品開発へ生かそうとしているが，まだ試行の段階にある。全体的には，製造業固有のデータマイニング適用事例はあまり多く見あたらない。

情報通信分野では，データマイニング技術がおもにインターネットの顧客マーケティングや電話網管理の分野に用いられている。この分野には豊富な電子化データ蓄積があるので，データマイニングの適用範囲は広い。特にネットワークとコンピュータの不正使用や不正アクセスの検出など，膨大な通信ログから特徴的パターンを発掘する適用は成功を収めている。また，インターネット上の膨大な Web ページから効率よくもっとも必要な情報を見つけ出す検索エンジンの開発に，テキストマイニングなどのデータマイニング技術が活用され成果を上げている。

2.3 データマイニングの応用に際する課題

以上に述べたように，データマイニングはすでに様々な分野に適用され，各分野で用いられているデータマイニング技術も多様である。公表された適用事例から，データマイニングの有効性と有用性が実証された反面，いくつかの問題点が浮かびあがってくる。

(1) データマイニングツールの選択と導入

データマイニングの実用化に伴い，データマイニングツールの商用化も急速に進んでいる。百種類以上のデータマイニングツールと称するソフトウェアパッケージが国内外のベンダーより発売されている [6]。実装されたマイニング手法，利用可能なデータ変換・視覚化機能および得意な業界・業務分野はツールによりそれぞれ異なり，対象となる問題を意識しながらデータマイニングツールを選択しなければならない。また，単体のデータマイニングツールとして典型的な問題に優れた性能を発揮するよりも，現実問題に対するソリューションとなる総合的なシステム性能が強く要求されている。

このため，製造，通信分野では市販ツールを利用するのみならず，自分野の開発者とユーザが有

する技術的な蓄積を活用し，種々の技術をテストしその中から最良のものを選択して，適用対象にカスタマイズ，チューニングしたツールやシステムを自ら構築することが多い。これに対し，金融や流通分野では，事例に適したツールをユーザ自らが開発することは少なく，既製の市販ツールを用いてデータマイニングを実施することが多い。

さらに，データマイニング技術の開発や使用時の体制も業界ごとに異なっている。製造や通信分野では，対象データ収集，技術開発からシステム開発，使用までをすべて自前で行う場合が多いのに対して，金融や流通分野では，コンサルタントや開発企業と組んでシステム開発，あるいはスキーム開発を行い，市販ツールを購入しユーザとしてそのまま利用する場合が多い。

(2)　データ収集のボトルネック

ビジネス現場に蓄積されたデータを利用する際には，データが特定のマイニングを目的として蓄積されたものではないため，データマイニング作業に必要なデータが存在しないまたは不完全なデータしか入っていないことがよくある。このような目的とデータ内容の不整合性は，実際にマイニング分析を実行しないと明らかにできないことも多い。低コストで補足データを収集する手段の確保や，既存データから必要情報を推定する手法などの技術が重要となるが，現状ではまだまだ未開拓である。データマイニングの研究およびその応用ではマイニングの本体技術ばかりが注目されるが，データ収集法の十分な検討・改良，補足処理など，データ収集技術こそが成功の鍵を握っているといっても過言ではない。

(3)　データマイニング実施体制

市販ツールを購入してユーザが使用すれば何とかなるという「市販ツール万能主義」が採られる傾向がよく見られる。効果的なデータマイニングには，豊富な知識と経験に基づいて各事例に適した様々な技術の組合せや設定条件を見つける必要がある。マイニングの目的や対象データの内容は事例ごとに千差万別である。現状の市販ツールは個別技術およびそれらを組み合わせる環境を提供するものであって，事例に即した適切な技術の結合による処理スキームや各種性能評価指標，チューニングパラメータなどの設定までは教えてくれない。データマイニングの実務への適用においては，ユーザが十分な知識や経験を蓄積するための時間や資本を投入するか，コンサルタントや開発会社と密接な連携体制を敷くなどの投資が必要となる。

(4)　分析者スキルと統合・連携との欠如

分析者のスキルがなければ，データを使いやすいように加工する(クレンジング作業)ことも，何の分析ツールを利用すればよいかと判断することもできない。データマイニング技術を提供する研究開発者側にも「研究分野の分断問題」が横たわっている。データマイニング技術は，人工知能，データベースと統計解析などの複数分野の基礎技術に根ざすものであるが，これらの個別分野の研究開発者の連携が必ずしもうまくいっているとはいえず，技術がばらばらに提供されている傾向がある。また，データマイニングの実施スキーム全体を通じた各種技術の組合せにおける整合性や，各種目的やデータ内容に根ざした処理スキームの体系化などに関する研究もまだまだ手付かずの状態である。ユーザの立場に立ち，必要となる補助技術，各種判断指標，体系的マイニングスキームの蓄積などに関する研究が必要である。

3. 事例紹介

3.1 会社概要

本研究では，ある小規模の卸販売会社を対象とする。この会社の取り扱う商品を大きく分類すると，表3に示すとおり24分類がある。主に幼稚園の園児から小・中・高校の生徒・学生用の衣服類と関連スポーツ用品で1,500点があり，色・サイズの区別をつけると，20,000点ほどある。

表4に示すように，得意先は3つの分類に分かれ，それぞれ異なる形態で取引をおこなっている。学生・生徒への直販を学校ごとに1社ずつ，店頭販売も1社の得意先としてまとめると，得意先は800社くらいある。従業員20人で，年商6億である。

表3. 取り扱う商品の分類

分類コード	商品分類	分類コード	商品分類
1	ウェア類	13	水泳用品類
2	タイツ類	15	武道用品類
3	半袖シャツ類	16	林間用品類
4	長袖セーター類	17	オリエンテーション用品類
5	男子ショートパンツ類	20	サンプル
6	女子ショートパンツ類	21	新規取扱商品
7	ランニング類	29	各種メーカー品
8	バッグ・各種袋類	31	野球用品
9	帽子・はちまき類	33	モリリン
10	校章・名札類	35	その他（刺繍など）
11	シューズ類	40	πウオーター
12	水泳ウェア類	65	化粧品

表4. 得意先の分類

分類	得意先の内訳	取引形態
一般	・店頭で販売する商品売上の計上 ・営業担当者を経由し販売する商品売上の計上 ・学校のサークルや一般団体	取引の規模が小さいが，販売ルートにより入金形態は異なり，掛売と現金売りなどがある
小売	学校の生協や学校周辺の小売店	掛売が中心である
学校	学期初めに，制服や運動着などを学生・生徒へ直販し売上金を学校ごとにまとめた得意先	学生・生徒からの注文を受けるルート（直接受注か生協経由）により入金形態は現金と掛売に分かれる

3.2 経営課題とデータマイニング

(1) アソシエーション分析

学生用衣服類は非常に特殊な商品であり，学校の指定を受けていれば一定の販売数量を確保できるメリットがある反面，採用指定が取り消されたらまとまった数量の販売市場が消えるうえ，再び採用指定を受けてもらうには何年もの営業努力が必要となる。また，少子化による需要の絶対減少，リサイクル活動の活発な展開による新規需要の減少，それに服装の指定を中止する公立学校の増加などのため，主要商品の売上減少は避けられない。

これに対応し，学校指定の学生用衣服類のみならず，文房具・スポーツ用品などの非学校指定の学生用品の販売にも力を入れて，これにより売上の減少に歯止めをかけることがますます重要になっている。このため，学校指定商品と非指定商品（以下，一般商品と呼ぶ）との併買動向を学校別，地域別，商品別，販売ルート別に解析し，「ビールとオムツ」のルールを発見することは経営上非常に価値があり，データマイニングの典型的な適用分野でもある。

(2) 優良得意先と優良商品の識別

優秀な営業人材の確保と定着化がすべての企業にとって重要でありながら，数十人規模の中小企業にとっては非常に難しい経営課題である。日々の営業活動体験だけではなく，販売データの解析と分析結果からより正確に得意先・商品の販売動向を把握し，優良得意先と優良商品を確実に識別することは，営業マンのスキルアップと営業活動の生産性向上につながる。判別分析やクラスタ分析などの手法を活用し，蓄積された販売データから得意先・商品の販売動向をいろいろな視点から解析することは，データマイニングの得意分野である。

(3) 得意先の信用評価と不良得意先の識別

長期の不景気のため，得意先の経営不振がますます深刻になり，6億円程度の年商に対して数千万円の滞納または不払いがあり，商品代金の即時回収と不払い防止が重要な課題となっている。このため，日々の商品代金の請求・支払データの解析を通じて，各得意先の支払状況と滞納動向を確実に把握のうえ，各得意先の信用度を評価し，危険な得意先との取引に制限をかけて，不良得意先との取引を中止することは経営上求められ，ここからもデータマイニングシステムを導入する必要性が見られる。

3.3 中小企業の独自課題

上述した経営課題を意識しながら，日常の取引記録と業務システムのデータベースから必要なデータを収集して，関連指標を集計したうえ，実際にいろいろな手法を用いてデータマイニングを行った。その結果，経営上有益な情報が得られ，中小企業においてもデータマイニングが非常に有効であることを実証したと同時に，以下の課題をうまく乗り越えなければ中小企業においてデータマイニングを実用化させることには無理があるということがわかった。

(1) 人材の不足

各種統計解析手法およびその他のデータマイニング手法を理解し，パソコンを熟練に操作して，

データマイニング作業を担当できる人材は中小企業において極めて少ない。このため，この分野のコンサルタントまたは専門家からの支援が必要である。また，データマイニングを日常仕事として定着させて，日常の経営活動または管理業務に活用するには，専門家ではなく一般の業務担当者でも容易に操作できるデータマイニングシステムが必要である。企業の製品・サービスの特徴に基づき，解決すべき経営課題を明確にしてデータマイニングの適用業務分野と適用可能手法を特定したうえ，データマイニング作業をかなりの程度で自動的に行うシステムを構築する必要がある。

(2) 資金の不足

大手企業に比べて，中小企業の情報化が大幅に遅れている。基幹業務情報システムの導入と運営さえ資金と人材の面で大きな負担となっている中小企業が少なくない。IT 投資に対する費用対効果を一層厳しく見るようになった現在，高価なデータマイニング専用パッケージソフトまたはデータマイニングツールを購入するには躊躇する経営者が多い。また投資をしても，本当に活用できるか，と大きな疑問が残る。企業ごとにデータと手法を特定してからデータマイニング機能を基幹業務管理システムに組み入れることが必要であろう。言い替えれば，高度な手法と高機能のパッケージより，わかりやすい簡易手法と使いやすいシンプルなマイニングシステムは中小企業に適している。

(3) データの収集

上述したようにデータ収集のボトルネックはデータマイニングの共通課題である。データベースの膨大なデータからどういうデータを選んで，どういう指標を使ってマイニングを行うかは，経営課題に関する認識と関連業務知識から判断しなければならず，目的指向が大切である。同時にデータベースのどこからデータを集め，どのように関連指標を計算するかを判断するには，企業の情報システムに関する知識が不可欠である。業務と情報システムとの両方に精通する人材がほとんどいない中小企業においては，この問題がより深刻であり，適用分野と手法を限定しても関連データを自動的または半自動的に収集できるシステムを開発し，データ収集作業を支援することが必要である。

(4) マイニング作業支援

時間をかけてデータマイニング作業をしても必ず経営活動に役立てる結果が得られる保証はない。いつデータマイニング作業を行うべきであるかは業務担当者の判断と経営活動のニーズに基づき決めるべきであるが，データの異常などを自動的に監視し，適切なタイミングで担当者に知らせて，解析作業を起こさせる仕組みが必要である。

(5) 結果の解釈

得られた解析結果を適切に解釈し，これを経営活動に反映していくことは，データマイニングの最終目的でありながら一番難しい作業である。現場の関連業務知識だけでなくデータマイニング手法に対する理解も必要である。専門家と現場担当者との共同作業が不可欠である。しかし，データマイニングを日常的に行うには専門家の支援がなくても現場担当者が独自で解析・解釈作業を進める必要があるので，結果解釈の段階には業務担当者を有効に支援するエキスパートシステムが必要であろう。

4. システム構成

4.1 データマイニングシステムの一般構成枠組

データマイニングシステムの構成に関しては，適用対象となる業界と業務課題によりシステム構成がそれぞれ異なるが，大量なデータから有用な情報・知識を掘り出すことはすべてのデータマイニングシステムに共通する基本機能であるため，J. Han ら［4］は，図3に示すような典型的データマイニングシステムの構成を与えている。

4.2 中小企業向けのデータマイニングシステムの枠組

上述した経営課題とデータマイニングの適用課題に関する認識に基づき，中小企業のためのデータマイニングシステムを提案し，その構成を図4に示す。

図4に示したとおり，提案したデータマイニングシステムは，データマイニング作業を2つの部分に分けて行う仕組みとなっている。1つは高度な統計解析手法および人工知能手法を用いて，より複雑なデータ解析とパターン発見を行う部分であり，もう一つは現在ビジネス活動において標準ツールとして使われている Excel などの表計算ソフトに内蔵する分析ツールとグラフ作成・ピボットテーブル機能を活用し，データ解析とパターン発見を日常的に行う部分である。各構成モジュールの機能に関する説明は以下のとおりである。

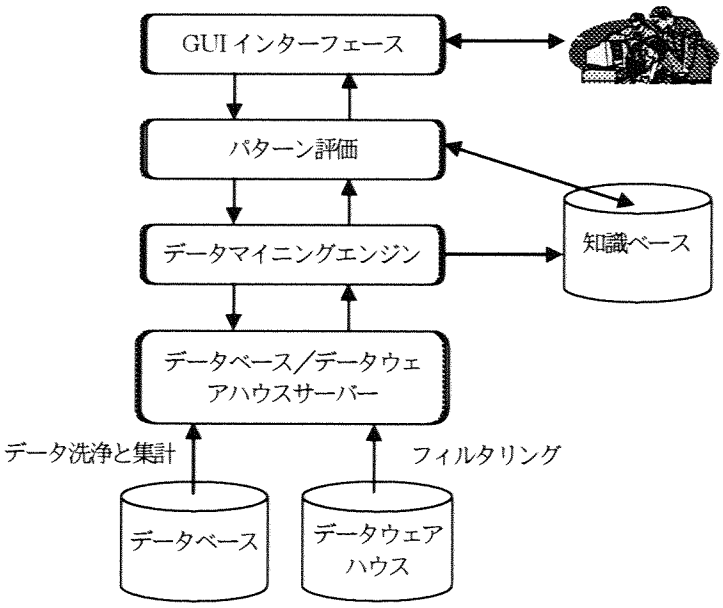

図3.　データマイニングシステムの典型構成

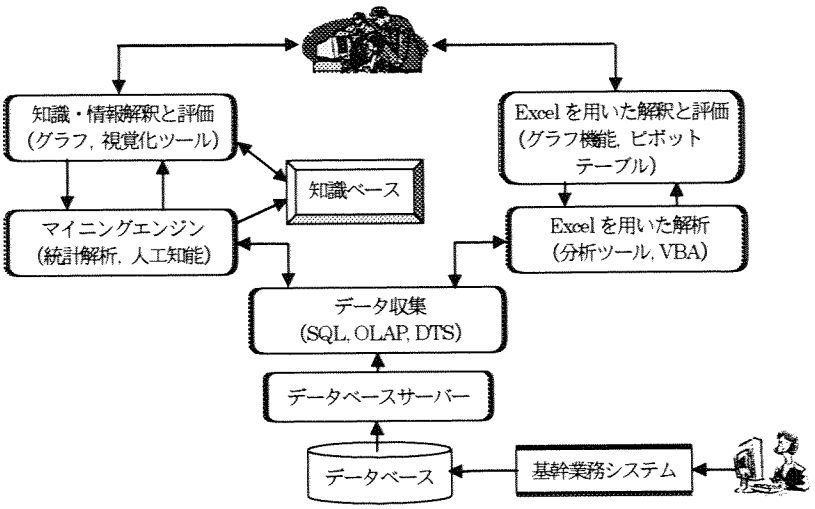

図4. 中小企業のためのデータマイニングシステム

(1) データベースとデータベースサーバー

基幹業務システムの各種データを収集・保管するためのデータベース管理システムとサーバーシステムである。高い効率と信頼性をもち，各種データを収集・保管するだけではなく，データの完全性・整合性を維持するためのエラーチェックとエラー処理機能も働かせなければいけないため，マイクロソフト社の SQL Server または Oracle 社の Oracle データベースなど，専用のデータベース管理システムを導入する必要がある。

(2) データ収集

このモジュールは2つの機能をもつ。1つはデータの洗浄（クレンジング作業）であり，データマイニングを行うために生データを解析しやすい形に加工する。具体的には，分析目的に合わせて大規模データベースから必要なデータ項目だけを抜き出したり，データ採取不足で抜け項目がある場合欠損値を削除・補完したり，不正データや特異的なデータを除去したりする。また，データ形式の変換や正規化などの処理も考えられる。もう1つの機能は，データ収集支援である。データベース管理システムに内蔵するストアドプロシージャなどの SQL プログラミング機能，オンライン解析処理 OLAP 機能およびデータ変換サービス DTS を活用して，データ収集作業をできる限り自動化させることにより，データベース操作・利用などがあまり得意でない現場担当者を有効に支援するのみならず，データ収集作業の効率アップを図れることができる。

(3) マイニングエンジン

決定木，バスケット分析，重回帰分析，相関解析などの統計解析手法，およびニューラルネット，事例ベース推論，機械学習などの人工知能手法を用いて，複雑なデータ解析とデータ処理を通じて，優良顧客の発掘や各種パターンの発掘・分析などを行う。データマイニング目的が予め決められる場合，ニーズに合わせて適用対象にカスタマイズ，チューニングしたツールやシステムを自ら構築することが望ましい。この場合，基幹業務システムと同じ開発ツールを用いて構築すれば，基幹業

務システムと統合することが容易なため，ユーザーインターフェースが統一され，担当者には理解しやすく操作利用が便利である。もちろん，市販のデータマイニングツールを導入する方が開発コストと開発工期の面で有利な場合がある。

(4)　知識・情報解釈と評価

データマイニングエンジンでの解析結果または発見パターンを人間の理解しやすいグラフなどの形式で表示する以外，担当者と対話しながら，マイニング手法と実行パラメーターを指定したり，マイニング結果に関して支持度・信頼度などの評価値を計算したりして，データマイニング作業をより容易にし，知識・情報解釈と評価を有効に支援する。場合によっては，市販のツールを導入する方が開発コストと性能・汎用性などの面で有利であるが，ユーザーインターフェースの統一性から考慮して，このモジュールは基幹業務システムと同じ開発ツールを用いて構築し，基幹業務システムと統合することが望ましい。

(5)　知識ベース

データ収集と情報･知識の発見を有効にガイドし，マイニング結果を正確に評価するための企業･業種独自の領域知識またはマイニング手法とパラメーター選択に関するメタ知識を保存･管理する。自前でデータマイニングシステムを構築する場合，明示的に知識ベースを構築するよりは，関連モジュールのプログラムに有効な知識を組み込んで，暗示的に知識ベースを内蔵することが現実的なやり方である。

(6)　Excel を用いた解析

パソコンの普及に伴い，Excel などの表計算ソフトはビジネス活動においてすでに不可欠な標準ツールとなり，一般の業務担当者でも活用することに熟練している。また，表計算ソフトは絶えずに進化し，組込み関数と分析ツールなどを標準で搭載のうえ，その機能もますます高度化されてきている。このため，Excel などの表計算ソフトに内蔵する組込み関数と分析ツールを使っても，一定レベルのデータマーニング作業ができることはこれまでの実例により実証されている。もちろん，業務担当者がこれまでに積んできた経験とスキルを活用し，Excel などの表計算ソフトを使って，対話的にデータ解析作業を行うことができるが，必要に応じて，VBA（Visual Basic for Applications）やマクロを利用し，繰り返して行われるマイニング作業を自動的に実行することもできる。

(7)　Excel を用いた解釈と評価

Excel などの表計算ソフトのグラフ作成機能を活用し，データとデータ解析結果を直感的に理解しやすいグラフ形式で表示したり，ピボットテーブル機能を活用し，データベースのデータをいろいろな視点から集計・解析したりして，有用な情報と知識・パターンを発見する。また，シナリオマネージャなどのツールを利用して，発見された情報・パターンに対してシミュレーションと評価を行うことができる。

典型的なデータマイニングシステムの構成案に比べて，提案したシステム構成案は，以下の点を強調するところに独自な特徴をもつ。

(1)　高度な専用ツールではなく，Excel などの表計算ソフトを使って，データ分析とデータマイニング作業を日常的に行うことを考慮して，データ収集の自動化と作業支援に重点を置く。これにより中小企業現場において積んできた業務管理ノウハウとパソコン利用経験が十分に活かさ

れ，投資の少なく効果的なデータマイニングシステムを構築することができる。

(2) データマイニング手法の高度化と汎用性を追求せず，中小企業の経営課題とデータマイニングの目的を明確にしたうえ，目的にあわせて関連知識を整理し，データマイニング作業をできる限り自動化させるよう対応モジュールを開発することを強調する。対象範囲と機能の限定を通じて，よりシンプルなシステムを構築することができる。これにより，目的が明確でだれでも簡単に利用できるシステムが得られる。

(3) 基幹業務システムとの連携・統合を強調する。いつ，どの種類，どの範囲のデータに対してデータマイニング作業を行うべきであるかは業務担当者の判断で決めるだけでなく，データの異常などを自動的に監視し，適切なタイミングで担当者に知らせて，そして関連データを自動的に収集し，より詳しく作業指示を与えたうえ，マイニング作業を起こさせる仕組みを基幹業務システムに組み込む必要がある。また，データの洗浄（クレンジング）作業を簡単に行うには，基幹業務システムのデータ収集と保存関連の機能またはモジュールに適切な工夫が必要である。

4.3 データマイニング手法の適用性

P. Cabena ら [2] はビジネス分野におけるデータマイニングの機能・役割を次の 4 つに分類する。

- 予測：いままでの観測データから将来を予測する。商品の売行き傾向・販売量の予測などのようにビジネス分野における予測機能の必要性が高い。
- セグメンテーション：対象特性の類似性と相違性に基づき，顧客や商品などをいくつかのセグメントまたはクラスに分類する。マーケティング分野ではセグメンテーションが不可欠な機能である。
- 関連性分析：異なる物事同士の関連を見つけ，かつ定量的に相関関係を評価する。
- 外れ値の検出：以前に知られていた予想または標準から大幅に逸脱する異常値を見つける機能である。異常値から本当に役立てる情報が発見できることは大いに期待されて，外れ値の検出機能はますます認識され，その重要性が高くなっている。

統計解析，人工知能，データベースとその他の分野から膨大な数のデータマイニング手法が提案されて，これらの手法の一部を対象事例会社のデータマイニングに適用したところ，データマイニングの目的とデータの種類などによりデータマイニング手法の有効性が大きく変わって，場合によっては有効なマイニングがまったくできないことがわかった。この事例会社のデータマイニング経験に基づき，主なデータマイニング手法とその特徴，およびこれらの手法と上述したデータマイニングの 4 機能との適合性を考究し，その結果を表 5 に示す。ただし，適合性は，次の 4 段階で評価される：

- ◎：指定手法が該当機能の実現に非常に有効である。
- ○：大部分の問題に対して指定手法が該当機能の実現に有効である。
- ◇：一部分の問題に対して指定手法が該当機能の実現に有効である。
- △：指定手法が該当機能の実現に必ずしも有効ではないが，問題によっては適用可能である。

表5. データマイニング手法とその適合性

分類	データマイニング手法		4機能との適合性			
	手法	特徴（長所，短所）	予測	セグメンテーション	関連性分析	外れ値の検出
統計解析	決定木	・分析対象をクラス分類し，IF-THEN ルールで表す ・カテゴリーデータ等の非数値データを高速に処理できる ・結果がわかりやすい ・連続データのグループ化が難点	◎	◎	○	◇
	相関分析	・異なる商品の併売傾向を解析するバスケット分析 ・商品の売行き変動を解析する時系列パターン抽出 ・異なる商品の売行，顧客購買傾向を解析する類似時系列パターン抽出	○	△	◎	△
	主成分分析	・固有値と寄与率に基づき多数の要素を含むデータを集約できる ・結果の解釈が主観的	○	◎	○	△
	クラスタ分析	・商品や顧客をいくつかのクラスに高速に分類できる ・多数のアリゴリズムが利用可能 ・結果の意味づけや解釈が難しい	△	◎	△	○
	ビジュアリゼーション（可視化）	・人間の知識と経験を活かして適用範囲が広い ・複雑なパターンをわかりやすく表示できる ・定量的な分析が難しい	○	○	○	○
人工知能	ニューラルネットワーク（BP）	・バックプロパゲーション法を用いて，複雑なルールやパターンの学習に適応し，各種の問題にも高い適合性をもつ ・教師信号を与えて学習する必要がある ・精確な結果が得られながら因果関係の解釈が難しい	○	○	○	○
	コホーネンネットワーク（SOM）	・自己組織マップとも呼ばれ，教師信号なしで効率よく分類問題を解決できる ・多次元データを圧縮し，低次元マップによる可視化が便利である	◇	◎	◇	○
	ルールベース推論	・IF-THEN ルールで情報・知識を表現しあらゆる問題を効率よく解決できる ・ルールの獲得は専門家のノウハウに頼る。自動獲得は難しい	○	○	○	○
	事例ベース推論	・事例同士の類似性に基づき結論を見つけて，あらゆる問題を効率よく解決できる ・ルールの獲得が困難な場合でも適用できる ・結論の正当性評価が難しい	○	○	○	○
	テキストマイニング	・文書から有用な情報・知識の発見に有効である ・Web 書込みの管理と日報の解読などの分野で大いに期待されている ・現実にはまだ多数の課題が残っている	◇	○	○	○

しかしながら，データマイニング手法の有効性はデータマイニングの目的と要求，問題の性質とデータの種類，および手法の使い方などにより強く影響され，表5の適合性評価はあくまでも典型的な問題と通常の使い方に限定した場合に得られた評価であり，他の場合には参考にならないことがある。

また，市販のデータマイニングツールには，上述したすべてのデータマイニング手法を搭載するものが現時点でまだ存在しない。搭載した手法と詳細アルゴリズムはツールによって異なり，購入のとき詳しく調べる必要がある。

5. 信用評価問題への人工知能手法の適用

5.1 信用評価問題

3.2 節で説明したとおり，商品代金の不払と滞納を防止するため，日々の商品代金の請求・支払データの解析を通じて，各得意先の支払状況と滞納動向を確実に把握のうえ，各得意先の信用度を評価する必要がある。

現時点で入手可能なデータを考慮して，以下の指標を用いて，得意先の信用状況を解析し評価する。

- 得意先分類：一般，小売と学校をそれぞれ 0, 1, 2 と数値化する。
- 滞納額：評価期間中各月滞納金額の平均値
- 最大滞納日数：評価期間中代金滞納日数の最大値
- 滞納回数：評価期間中代金滞納が発生した回数
- 売上：評価期間中売上金額の合計値
- 取引回数：評価期間中売上が発生した月数
- 滞納割合：評価期間中売上金額の合計値に対する滞納平均額の割合

また，各得意先の信用度は下記の図 5 に示すように 0 から 1 までの数値を用いて 6 段階に分けて評価する。

5.2 事例ベース推論を用いた信用評価

以上の信用度評価問題に対して，最近 1 年間のデータを基幹業務システムのデータベースから収集し，さらに図 5 の評価基準に基づき経理担当者に各得意先の信用度を与えてもらった。これらのデータから有効なパターンまたは規則性を発見しこれからの信用度評価と予測に利用することをデータマイニング目的として，判別分析やクラスタ分析などの手法を適用して，いろいろと調べたところ，直接にこれからの評価と予測に利用できそうな結果は得られなかった。これは，得意先の種類が異なり，それぞれ複雑な行動パターンで取引を行い，一意的な規則性または共通性の高い行動パターンが存在しないためであると考えられる。

そこで，一意的な評価（予測）より，複数の評価（予測）値に加えてそれらの支持度と信頼度を与える手法を求めて，事例ベース推論手法を取り入れることにした。

事例ベース推論とは，現在の問題と過去に解いた問題との類推を用いて問題解決を行う推論方式

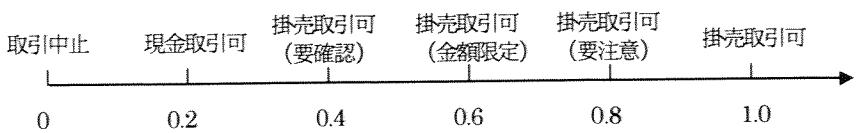

図 5. 得意先の信用度

表6. 得意先グループと特性指標を表す変数と記号

グループ	特性指標							信用度
	分類	滞納額	滞納日数	滞納回数	売上	取引回数	滞納割合	
G_1	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	$y_1=0.0$
G_2	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	x_{27}	$y_2=0.2$
G_3	x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	x_{36}	x_{37}	$y_3=0.4$
G_4	x_{41}	x_{42}	x_{43}	x_{44}	x_{45}	x_{46}	x_{47}	$y_4=0.6$
G_5	x_{51}	x_{52}	x_{53}	x_{54}	x_{55}	x_{56}	x_{57}	$y_5=0.8$
G_6	x_{61}	x_{62}	x_{63}	x_{64}	x_{65}	x_{66}	x_{67}	$y_6=1.0$
評価対象	x_{71}	x_{72}	x_{73}	x_{74}	x_{75}	x_{76}	x_{77}	$y_7=?$

である。事例ベース推論では，過去に解決した問題を問題解決事例として保存し，問題解決事例を現在の問題に利用する［7］。本研究では，次のとおりに事例ベース推論を用いて，得意先の信用度を評価する。

(1)　事例と新規評価対象の記述

表6に示す変数と記号を用いて，過去の事例と新規評価対象を表す。

表6から事例としての各得意先はその信用度評価値に基づき，それぞれG_1～G_6との6つのグループに分けられ，各グループの得意先の特性指標と信用度はそれぞれx_{11}～x_{67}，y_1～y_6とする。また，新しい評価対象とする得意先の特性指標と信用度はそれぞれx_{71}～x_{77}，y_7とする。

(2)　評価対象と事例グループの距離の評価

新規評価対象とする得意先の特性指標x_{71}～x_{77}が与えられると，次式(1)を用いて，この得意先と事例との距離を計算する。

$$d_i=\sqrt{\sum_{k=1}^{7}(x_{ik}-x_{7k})^2} \quad , \quad i=1,2,\cdots,6 \tag{1}$$

ただし，各事例グループに2つ以上の得意先がある場合，それぞれの事例得意先に対して，式(1)の距離を求めてから，さらにグループごとに距離の平均値を計算し，この平均値をd_iとする。

(3)　類似度の評価

新規評価対象と各事例グループとの類似度をs_i，$i=1,2,\cdots,6$として，s_iは次式(2)を用いて評価する。

$$s_i=1-\frac{d_i}{\sum_{k=1}^{6}d_k} \quad , \quad i=1,2,\cdots,6 \tag{2}$$

式(1)の距離と式(2)の類似度の定義により，類似度の値が1に近いほど，新規評価対象と事例との類似度が高い。これらの類似度は対応する信用度予測値への支持度と見なすことができる。

(4)　信用度の評価

以上の計算結果により，新規評価対象の信用度予測値は次のとおりに決められる。

信用度予測値y_T	0	0.2	0.4	0.6	0.8	1.0	平均値 y
支持度	s_1	s_2	s_3	s_4	s_5	s_6	—

ここでの信用度予測値は2つの形式で与えられる。

・複数の予測値とそれらの支持度：新規評価対象が各事例グループにどれだけ近いかを，支持度から判断できる。

・平均値：1つの値で総合的に信用度を評価できる。

ただし，信用度予測値の平均値は次式（3）を用いて計算する。

$$y=\sum_{k=1}^{6} s_k y_k / \sum_{k=1}^{6} s_k \tag{3}$$

具体的事例と計算結果およびこの手法の有効性については別報の論文に報告させていただきたい。

6. おわりに

以上では，ある学生服卸販売会社を具体的な事例として，中小企業へのデータマイニング手法の応用を考究した。得られた成果を要約すると，次のとおりである。

・統計解析などの手法と比較しながら，データマイニングの定義・位置づけを議論したうえ，データマイニングの応用と課題について検討した。

・対象とする中小学生服卸販売会社の事例を踏まえて，中小企業の経営課題とデータマイニングの必要性を明確にしたうえ，中小企業におけるデータマイニングの適用分野と課題を考究した。

・高度な統計解析手法および人工知能手法を用いたデータマイニングと，Excelなどの表計算ソフトを活用し日常的に行うデータマイニングを区別し，中小企業向けのデータマイニングシステムの構成を提案した。

・対象事例のデータマイニング作業経験に基づいて，主なデータマイニング手法とデータマイニングの4機能（予測，セグメンテーション，関連性分析，外れ値の検出）との適合性について検討した。

・事例ベース推論手法を用いて得意先の信用評価を行うアプローチを提案した。

いろいろな制限のため，具体的なデータと結果が公表できないが，本論文の議論と提案は他業種の中小企業にも適用できると考えている。本論文の提案を具体化にして，実際に対象企業に適用し，確実な経済的効果を得るには，まだ相当の時間と労力が必要であり，これからの研究テーマとしたい。

参考文献

1. SASインスティチュートジャパン：データマイニングがマーケティングを変える！：経験とカンを科学する最新手法，PHP研究所，2001.

2. P. Cabena, etc 原著, 河村佳洋・福田鋼志監訳: データマイニング活用ガイド, 概念から実践まで, 星雲社, 2000年9月.
3. D. Hand, H. Mannila, P. Smyth: Principles of Data Mining, The MIT Press, 2001.
4. J. Han, M. Kamber: Data Mining: Concepts and Techniques, Morgan Kaufmann, 2000.
5. 元田　浩, 鷲尾　隆: "データマイニング展望", システム/制御/情報, Vol. 46, No. 4, pp. 169-176, 2002年4月.
6. 河野浩之: "データマイニングツール", システム/制御/情報, Vol. 46, No. 4, pp. 209-214, 2002年4月.
7. J. Kolodner: Case-Based Reasoning, Morgan Kaufmann Publishers Inc., 1993.