5G-08

産業分野のITソリューション

－ データマイニングフレームワークによるデータ活用 －

松本一教，早瀬健夫，村田由香里，田原歩，守安隆
E-mail: kazunori@sitc.toshiba.co.jp
株式会社 東芝
情報社会システム社 SI技術開発センター

概要

産業分野での制御系システムと情報系システムの容易な連携を実現するソフトウェアフレームワークを開発中である.この連携により，監視制御データを実りある知識の源として扱うことが可能となる．本稿では，フレームワーク上に構築されるデータマイニング機能の実現について概要を紹介する.

1. はじめに

遠隔のデータを一括して中央に収集し，管理したり，遠隔の設備を中央から制御するというシステムが産業上多くの分野で使われている．そのためのシステム構築コスト削減や，要求仕様変更に迅速に対応できる手段の提供は重要な課題である．筆者らが開発中のソフトウェアフレームワーク[1,2,3]は，このような産業上の遠隔制御監視を対象とするもので，再利用性や仕様変更容易性などソフトウェア工学的要求に答えられるよう設計されている.

産業分野での監視制御系システムと情報系システムとの連携の重要性は既に指摘されており，SCM(Supply Chain Management)，ERP(Entrerprise Resource Planning)システム等との接続も試みられている．最近は産業活動における知識利用の重要性が益々高まっているため,そうしたシステムとの結合だけでなく，生データから知識を抽出する技術であるデータマイニングとの融合が期待されるようになっている.

2. システム構成

本フレームワークには幾つかの主要なサブシステムから構成されている．データマイニングの立場からは，データを収集してデータベース作成を担当するサブシステムと，データベースに対してデータマイニングを適用するサブシステムの 2 種類が重要である．データベース作成サブシステムでは，監視制御対象の一部に組込まれ低レベルのデータを収集するシステムと，それら複数個を管轄下においてある程度大域的なデータ収集を行なうシステムとが階層的に組合され構成されている.産業システムおよび企業の全体がインターネットで結ばれているという状況では，分散オブジェクト技術がこの実現のための核となる.その詳細について機を改めて報告予定であるが，適用事例の一部は[2]に示してある.

3．データマイニング

抽出する知識の種類に応じて，ニューロ学習に基く方法,決定木学習に基く方法,など様々なデータマイニング方法が開発されており，

A Solution of Industrial Information Systems: a Data Mining Framework
Kazunori MATSUMOTO, Takeo HAYASE, Yukari MURATA, Ayumu TAHARA, and Takashi MORIYASU
Information and Industrial Systems & Service Company, TOSHIBA Corp.
3-22 Katamachi, Fuchu, Tokyo 183-8512, Japan.

ツールとして市販されているものも多い．開発中のフレームワークでは，これら各方式をサポートする予定であるが，現状では相関ルールマイニングに関する実装を主として進めている．本稿では，筆者らの開発した相関ルールマイニング方式について特に説明する．

相関ルールマイニングでは，ルールの支持度*(support)*や信頼度*(confidence)*の与え方により，極めてまれにしか発生しないデータをも無視せずにルール探索の対象とすることができる．このため，低頻度でしか発生しない異常事態に対する知識抽出も十分に行える．ISA 階層で与えられるデータの概念階層を利用するという拡張による強力化もされてている[4]．しかしその代償として，データベース全体を繰返し走査する必要が生じ，性能上の問題の原因となっていた．筆者らの開発した方式では，相関ルールのデータマイニングを情報検索の特殊ケースと見なす立場を採用している．すなわち，データベースに対して，事前にシグネチャと呼ばれる一種のインデクスに相当するものを計算し，シグネチャファイルに格納するようにしている．いったんシグネチャファイルが作成されれば，相関ルールのデータマイニングはほとんどシグネチャファイルだけを参照して行うことができる．しかも参照のほとんどはビット演算だけを用いて高速に実現できるというメリットを持つ．しかし，高速なデータベースアクセスが可能となる代償として，シグネチャファイルという記憶領域のオーバーヘッドが必然的に発生することになる．

今，データマイニングの対象となるデータベース中に存在するレコード数を N とし，各レコード中でのデータ項目数を r とする．抽出される相関ルールの支持度を $0<\sigma<1$ とする．このとき，1 レコードに対する理論的に最適なシグネチャは，

$$b \approx \left(\frac{1}{\ln 2}\right)\left(\frac{r}{m}\right)\log_2\left(\frac{1}{\alpha\delta N}\right)$$

ビットの記憶領域を占めることが示されている[4]．ここに，$0<\alpha<1$ はシグネチャを用いた本方式により必然的に発生する一種のノイズ率であり，知識抽出の純度に影響する．即ち，ノイズを少なくした知識抽出を行うためには，より多くの記憶領域が要求されることになる．また，m は与えられる概念階層の平均深さである．結局，シグネチャファイル全体のサイズは，bN ビットということになる．実際には，レコード数 $N=10^6$, $\sigma=10^{-2}$, $\alpha=10^{-1}$, $r=30$, $m=2$ のときに，b の値は約 650(ビット)となり，bN は約 80 メガバイトということになる．この程度の領域オーバヘッドで，極めて高速なデータマイニングが可能となる．

おわりに

データマイニングアルゴリズムだけではなく，データ収集や加工の段階からの一括してサポートが必要である．本フレームワークでは，産業システムからのデータ獲得段階まで溯ったサポート実現する．実際の適用評価を含め，より詳細な内容については機を改めて報告する．

参考文献

[1] 早瀬健夫 ほか，産業分野の IT ソリューション，情処第 60 回大会 5G-09

[2] 村田由香里 ほか，産業分野の IT ソリューション，情処第 60 回大会 6S-09

[3] 田原歩 ほか，産業分野の IT ソリューション，情処第 60 回大会 3G-03

[4] K.Matsumoto，Data Mining of Generalized Association Rules Based on a Method of Partial-Match Retrieval，DS99，1999.