プロセスデータに対するデータマイニング手法の検討

河野浩之†
† 京都大学大学院情報学研究科

1 はじめに

1990 年以後，データベースシステムやデータウェアハウスに記録されるトランザクションデータに対する高度処理を目指し，データマイニング研究が行われてきた [7]．現在に至るまで，様々な情報システムから生成される実データに対して，アルゴリズムの実装と，その性能改善が試みられ，分析・解析・発見を行う各種手法の統合環境を提供するデータマイニングツールの開発が活発に進められている．

我々は，データベースから知識獲得を行う KDD (Knowledge Discovery in Databases) の研究領域に興味をもち，通信ネットワーク上のパケットモニタリングで得られるデータに基づく知識獲得の研究 [8] を，1992 年に開始した．また，POS を始めとする企業情報システムに広く用いられている関係データベースとオブジェクト指向データベースに対して，新しい発見アルゴリズムを提案した [14]．

1995 年以後，半構造データである Web データを対象に，相関ルール [17] の良好な性質に着目した研究を行った [13]．他のテキストデータとして，雑誌記事索引データや INSPEC データベース，さらに，ピアツーピアネットワーク上のデータなど [1]，情報フィルタリング技術を中心に研究を進めている．また，PHS や GPS などを利用した位置データ取得を用いて，交通計画における基礎データの変質を視野に入れた時空間データとしての高次元データ処理技術を，2000 年 [9] 以後進めている．しかしながら，これら対象データのいずれもが，企業情報システムで生成される社会活動に伴う実データである．

そこで，本稿では，我々にとって，新たな領域の実データである製鉄プラントにおいて生成されるデータを題材に，プラントなどのエンジニアリング・プロセスにおいて生成される大量データに対するデータマイニングの可能性を検討する．

以下，2 章では，エンジニアリング・プロセスから生じるデータ量を扱うための基本的な技術基盤が形成されているかを，データマイニング研究の動向を踏まえながら述べる．次に，3 章では，データマイニングツールに実装されているアルゴリズムの利用状況を考える．4 章では，製鉄プラントで得られた実データに基づいたデータマイニング技術の可能性を探る．そして，今後議論を深めるべき研究アプローチを 5 章で議論する．

2 データマイニング研究の動向

産業界で利用されているデータマイニング技術の状況を，データマイニング研究の動向に基づきながら簡単に取り上げる．例えば，KDD[15] や PAKDD[4] におけるインダストリアルトラックにおける研究は，マーケティングデータを扱う事例が殆どである．他には，Web トラフィックやネットワーク監視を扱うものが目立つ．また，時系列データ解析処理の研究では，自然現象の観測データや医療データなどを扱うことが多い．

また，kdnuggets * においても，ASP，銀行，バイオ，CRM などにおけるソリューションとしてのデータマイニングが多数紹介されている．関連研究である情報視覚化 (Information Visualization) では，対象データの大半が抽象度の高いテキスト情報であり，通信ネットワーク，天体観測，医用画像などから得られるデータは少ない [2]．

一方，現在，ペタバイト (PB, (10^{15})) 規模のデータに対するデータマイニング技術の可能性が検討されている [16]．この点からは，データ蓄積に起因する制約は極めて少なくなりつつことが分かる．従って，これまでデータ蓄積を行うことがなかった応用領域においても，データを長期的に蓄積することで，新たな可能性をもつ手法導入の機会が生じている．すなわち，公的機関から提供されており，データマイニング研究に良く利用されている天文データ／気象データ／地震データなどに限らず，民間企業で構築されるプラントにおいて生成される大量データが新たな対象として加わると考えられる．

実際，通信ネットワークと大規模ハードディスクの急速な普及により，従来から大量データ処理が行われていたプラントシステムに関係するデータは，より大規模化を遂げつつある．そして，建設・運転，生産管理・在庫管理・出荷管理などに関する様々な処理基盤となる PIMS(Plant Information Management System)[12] †と呼ばれる情報システムの導入により，上流から下流

Potential of Data Mining for Engineering Process Data
† Hiroyuki KAWANO (kawano@i.kyoto-u.ac.jp)
Department of Systems Science, Graduate School of Informatics, Kyoto University

* http://www.kdnuggets.com/solutions/index.html
† http://www.yokogawa.co.jp/EXASOFT/01_Top_Page/index.htm など．

に至る広範な情報統合が進みつつある．よって，これらのデータに対して，データマイニング技術適用の可能性を検討することは興味深い [11].

3 データマイニングツール [10]

プラントに関わる様々なプロセスデータの情報統合を広範に進めながら，データマイニングを容易に行うには，優れたアルゴリズムの組合せ，もしくは，適切なツール群が必要である．なお，主要タスクである「分類，クラスタリング，関連・相関性，系列，特異値・不正発見，レポート作成・要約，視覚化，統計，OLAP・次元解析」などの手法は，幅広く実装されている．

また，商用データマイニングツールのモジュールに，多くのアルゴリズムが採用され，Suites として提供されている．例えば，決定木を用いた分類 (classification)，統計解析ソフトウェア (statistical analysis software)，視覚化ソフトウェア (visualization software) などである．

ただし，データベース関係のデータマイニング研究に大きな影響を与えた相関ルールに限定しても，精度を決めるパラメータ調整手法，一般化操作に対する概念木の利用，データ更新に対応した漸増的な相関ルール更新など，対象となるデータサイズ等により，その性能は大きく異なる．なお，この種の性質は，否定を伴う相関ルール (negative association rule) や，ルールに基づく因果関係 (causality) などを含め，文献 [17] で詳しく検討されている．

したがって，プロセスデータの特性に合致するタスクを定め，そして，同一タスクに属するデータマイニングツール群の中から，問題解決に適した手法選択が行えるか否かが，データマイニングツール利用の課題となる．

しかしながら，kdnuggets において紹介されている適用分野例を見る範囲では，アプリケーションサービスプロバイダー，銀行・与信業務，バイオ情報学，電子商取引，マーケティング，ダイレクトメール解析，ヘルスケア，CRM (Customer Relationship Management) や SCM (Supply Chain Management) などが対象となっており，ここからは，プロセスデータ処理にツールを直接利用するには課題が残されていると考えられる．もっとも，PIMS に蓄積される業務データに関しては，意思決定支援 (DSS, Decision Support System) に関係するデータマイニングツール利用の可能性がある．

4 プロセスデータへの適用可能性

製鉄の精製プロセスにおいては，脱硫・脱リン・脱炭などの効果的な処理が，高品質鋼を低コストで生産するために要求される．そのため，各工程に関係する主要な属性に対して，長期にわたる時系列データとして蓄積されている．

そこで，その種のプラントシステムから得られたデータを題材に，プロセスデータに対して，前章で述べた様々なタスクが，どのように利用できるかを，典型的なマイニング手法を用いながら検討を進めた．

具体的なデータとして，8ヶ月間の間記録された，製鉄プラントに関わるデータを用いた．また，当該データの属性から 147 を選択し，該当期間中から 12,881 データを採用した．なお，一部データには欠落が存在した．

データの属性は,「受銑日時，各作業工程の開始時間と終了時間」，さらに,「P, S, Si, C, Mn, CaC_2, SiO_2 などの処理前推定量」,「脱燐・脱硫処理に要した時間と各実績値」などが含まれる．その多くの属性は，数値属性であるが，一部，文字属性も利用されており，例えば，「処理」属性は「A, B, C」の 3 クラスから成る．

次に，時系列データに対して，はずれ値検出を幾通りか実行し，除滓実績 X に対して，稀に値 60 から 120 程度を取るケースが存在することなどを求めた（ファイルサイズの制約などのため，グラフは割愛した）．

続けて，6 属性から成る除滓実績を対象として，C4.5 を適用し幾つかの決定木の生成を行った（図 1 に，その一部を示す）．非常に深い木が生成されることが大部分であったが，比較的コンパクトに表現された要約された決定木 (Simplified Decision Tree) から，除滓実績 X 値 ≤ 3 の条件をもつルールが求まった．なお，計算には，Sun Cobalt LX50（Intel Pentium III 1.5GHz 512MB SDRAM）Linux を用いており，いずれも 1 秒程度の実行時間を要する程度であった．

除滓実績 X	A	B	C
0	501	2,152	1,338
非ゼロ	594	748	7,548

表 1: 除滓実績 X 値に対する処理記録

さらに，データ中には，処理 A が 1,095, 処理 B が 2,900, 処理 C が 8,886 レコード存在している．そして，除滓実績 X が値 0 を取るレコードが 3,991 件ある．また，除滓実績 X が 0 の時に，処理 A となるレコードは 2,152 件記録されていた（表 1）．なお，このような手順に従いながら，得られた多数のルール評価を繰り返し，併せて専門家の判断も得ながら，ルールの精錬を進めることになる．

以上の過程では，時系列解析，はずれ値検出，決定木，分類などが必要であるが，その他にも，ニューラルネットを初めとする非線形モデルなどを柔軟に組合せる

```
C4.5 [release 8] decision tree generator
Wed Jan 15 12:19:45 2003
---------------------------------------

    Options:
File stem <process-5>

Read 12881 cases (2 attributes) from process-5.data

Decision Tree:

X <= 3 : A (3993.0/1840.0)
X > 3 :
|   X <= 33 :
|   |   X > 30 : C (381.0/102.0)
|   |   X <= 30 :
|   |   |   X <= 24 :
|   |   |   |   Y > 100 : C (3856.0/351.0)
|   |   |   |   Y <= 100 :
|   |   |   |   |   Y > 83 : C (2420.0/222.0)
|   |   |   |   |   Y <= 83 :
|   |   |   |   |   |   X <= 15 : C (112.0/19.0)
|   |   |   |   |   |   X > 15 :
|   |   |   |   |   |   |   Y <= 78 : A (3.0/1.0)
|   |   |   |   |   |   |   Y > 78 : C (32.0/3.0)
|   |   |   X > 24 :
|   |   |   |   Y > 92 : C (1167.0/139.0)
|   |   |   |   Y <= 92 :
|   |   |   |   |   X <= 26 : C (80.0/10.0)
|   |   |   |   |   X > 26 :
|   |   |   |   |   |   X <= 29 : C (90.0/13.0)
|   |   |   |   |   |   X > 29 :
|   |   |   |   |   |   |   Y <= 91 : C (23.0/5.0)
|   |   |   |   |   |   |   Y > 91 : A (4.0/2.0)
（途中省略）


Simplified Decision Tree:

X <= 3 : A (3993.0/1862.3)
X > 3 :
|   X <= 33 : C (8168.0/889.0)
|   X > 33 :
|   |   X <= 35 : C (160.0/80.9)
|   |   X > 35 : A (560.0/247.6)


Tree saved


Evaluation on training data (12881 items):

 Before Pruning           After Pruning
----------------   ---------------------------
Size      Errors   Size      Errors   Estimate

  33  3018(23.4%)      7  3024(23.5%)    (23.9%)   <<
```

図 1: 除滓実績値に対する決定木

ことが有効である．なお，この種のツール群を利用できる環境は整ってきており，既存の枠内で収集したプロセスデータに対して，データマイニングツールを利用した処理は，容易になっていると判断できる．

5 プロセスデータ処理における課題

前章では，サンプルデータ規模のプロセスデータに対して，ルールを求める環境が整っていることを示した．しかしながら,「よく知られたルール」が求められることも分かっており，今後，この種のプロセスデータから優れた知識を求めるには，どのような研究を進めるべきかを検討することが重要である．そこで，本稿では，以下の 3 点に注目して議論を進める．

1. 長期間にわたる高密度データ蓄積
2. プロセスデータ処理のためのデータ構造とアルゴリズム
3. PIMS に蓄積される上流・下流工程に関わる属性の利用

まず，項目 1) は，サンプルデータ収集で用いられていたように，何らかのイベントをトリガーとしたデータ蓄積が一般に行なわれていることから，連続的なシステム挙動を把握できない疎データとなっている．もちろん，同一属性を高密度にデータ収集を行っただけでは，データクリーニングによる除去対象になる．そこで，プラントの全工程で測定されるデータを対象として蓄積するプラント・データウェアハウスの構築が求められる．実際，大容量記録媒体の低価格化が進んでいることから，合理的なコストで，詳細なプロセスデータ蓄積が十分できると期待される．

次に，項目 2) に関しては，マルチメディアデータや時空間データを対象としたデータマイニングアプリケーションの研究成果が，効率的な処理を行うデータ構造として役立つ．なお，前章で扱ったデータに対する時系列データ解析や決定木生成では，連続値をもつ全属性の組合せ数が大きな計算コストとして必要であった．もっとも，プラントには，物理的システムの制約が多数存在することから，多くの探索候補がカット可能である．

ところで，プロセスデータに対する知的処理の研究は，様々な角度から行われてきた．データマイニングの源流の一つである人工知能領域において，定性推論の研究が，1960 年代以後繰り返し試みられている [6]．よって，全システムから得られる部分データに対してモデル記述を行い，部分モデルを積み重ねることで系全体の挙動を求める手法を，うまく取り入れることにより，探索候補の効率的なカットが実現できると考えている．

また，統計的手法を含め，既存手法の多くは，部分データから全システムの特性を推定するものとして活用されてきた．しかし，大量データを蓄積し操作することが可能になるにつれ，システムのもつ特徴を変えることなく，測定データの状態で利用することが可能になる．よって，微分方程式系などに抽象化して操作するのではなく，複雑な挙動を示すデータそのものによるシステム操作モデルを，経済的に合理的なコストで実行することも可能であろう．

最後に，項目 3) であるが，高付加価値・高品質な製品の多品種少量生産を低コストで実現するために，今後，業務システムにおけるデータと，プラント側のプロセスデータとを密接に関連付けることが要求されると予想される．すなわち，プラントにおける制約が増加するにつれ，業務系で生じた要求を最大限満足するために，プロセス制御に内在する制約／ルール／知識を，業務システムへと適確にフィードバックする技術開発が必要である．

以上，本稿では，これまでのデータマイニング研究において，あまり取り上げられていないプラントにおけるプロセスデータへの適用を検討した．既に，ツール群は充実していることから，手法開発に関わる技術的課題は比較的少ないと考えられる．むしろ，業務システムを含めたデータを取り入れることで，プラント運用知識のフィードバック領域を広げる視点が必要になると予想される．

謝辞

本稿の一部は，文部省科学研究費の研究成果による．

参考文献

[1] M. Nakatsuji, M. Kawahara, H. Kawano, “Advanced index refinement by classifiers and distillers in P2P resouce discovery,” International Conference on Intelligent Agents, Web Technologies and Internet Commerce (IAWTIC2003), Feb. 2003. (to appear)

[2] S. K. Card, J. D. Mackinlay and B. Shneiderman (Eds.), “Readnings in Information Visualization Using Vision to Think,” Morgan Kaufmann, 1999.

[3] S. Chakrabarti, “Mining the Web: Analysis of Hypertext and Semi Structured Data,” Morgan Kaufmann Publishers, 2002.

[4] M.-S. Chen, P. S. Yu and B. Liu (Eds.), “Advances in Knowledge Discovery and Data Mining,” Springer, 2001.

[5] T. Elomma, H. Mannila and H. Toivonen (Eds.), “Principles of Data Mining and Knowledge Discovery,” LNAI2431, Springer, 2002.

[6] 淵一博（監修）, “定性推論,” 知識情報処理シリーズ（別巻１）, 共立出版社, 1989.

[7] J. Han and M. Kamber, “Data Mining : Concepts and Techniques,” Morgan Kaufmann, 2000.

[8] H. KAWANO, S. NISHIO and T. HASEGAWA, “Knowledge Acquisition in Communication Networks,” IEEE Region 10 Conference, Tencon 92, pp.881–885, Australia, Nov. 1992.

[9] 河野浩之, “位置情報活用に関する基礎的考察〜経路データ活用とインデックス〜,” 情報処理学会研究報告, 2000-DBS-122, pp.291–298, 2000.

[10] 河野浩之, “データマイニングツール,” システム/制御/情報「データマイニング特集号」, Vol. 46, No.4, pp209-214, 2002.

[11] 倉橋節也, 寺野隆雄, “学習分類子システムを用いたプロセス時系列からのデータマイニング,” 情報処理学会知能と複雑系研究会，No.128-1, 2002.

[12] 真子秀樹, 中野信, 花島勝, “プロセス情報システムによる装置産業のための製造実行（MES）システム,” 日立評論, Vol.83, No.6, 2001.
(http://www.hitachi.co.jp/Div/omika/hyoron_06/index.html)

[13] 西村 英樹, 伊藤 耕一郎, 河野 浩之, 長谷川 利治, “重み付き相関ルール導出アルゴリズムによる WWW データ資源の発見,” 第７回データ工学ワークショップ (DEWS’96), pp.79–84, 1996.

[14] S. NISHIO, H. KAWANO and J. HAN, “Knowledge Discovery in Object-Oriented Databases: The First Step,” *Proc. of AAAI-93 Workshop on Knowledge Discovery in Databases*, pp.186–198, July 1993.

[15] F. Provost and R. Srikant (Eds.), “Proc. of the Seventh ACM SIGKDD,” ACM, 2001.

[16] A. S. Szalay, J. Gray and J. vandenBerg, “Petabyte scale datamining: Dream or reality?,” Technical Report MSR-TR-2002-84, 2002.

[17] C. Zhang and S. Zhang, “Association Rule Mining, Models and Algorithms,” Springer, 2002.