データマイニングの数理モデル構築と確率感度解析

（課題番号　１１６８０４３５）

平成１１年度～平成１２年度科学研究費補助金（基盤研究（C）（２））研究成果報告書

平成１３年３月

研究代表者　香田　正人
（筑波大学社会工学系教授）

はしがき

ＩＴ技術の発展により，１テラバイトを超えるような巨大データウェアハウスが構築されることも珍しく無いが，システムの複雑化にともなうデータ検索やデータ管理上の問題が発生してデータ検索応答時間の著しい低下や，昨今ビジネスインテリジェンスとしてニーズの高い知識発見型の高度な検索要求に対応できない事態も散見される．

データマイニングは，ＡＩ（人工知能）や統計学，オペレーションズ・リサーチ（ＯＲ），データベースに関する各研究分野の先端技術を総合的に活用することで，従来はデータの巨大な山中深く埋もれて死蔵されていた未知で有用な情報（ルールや仮説，パターン等）を，あたかも金鉱脈を探掘するかのように検索・発見してゴールドラッシュをもたらす革新技術であるとの認識やアナロジーから，ＣＲＭ（顧客関係管理）などのビジネス分野を中心に応用されている．データマイニングは，仮説検証型の従来のデータ解析手法と比較して種々の優れた特徴を有しているが，その代表的なものとして，機械学習による知識発見機能がある．

本研究では，ニューラルネットワークの持つ学習機能に着目し，確率感度解析理論の適用による新しい確率的学習アルゴリズムの導出を行ない，それを用いたデータマイニング・モデルの構築について考察した．また，ブートストラップ法や最小記述長（ＭＤＬ）にもとづく情報量基準の適用についても考察を行ない，これらの技術の適用によりデータマイニングについての高度な数理モデルを構築可能であることが示された．こうした研究と合わせて，データマイニングの製造技術への応用についても一連の考察を行なった．

今後は，本研究にもとづいて，さらに高度なインテリジェンスを有するデータマイニング・モデルの構築を行ない，その知識発見性能の向上を図る予定である．

Keywords: Data Mining, Neural Network, Stochastic Sensitivity Analysis, Bootstrap Method, Minimum Description Length (MDL)

データマイニングの数理モデル構築と確率感度解析

科学研究費補助金（基盤研究（C）（2））

研究組織

研究代表者：香田正人（筑波大学社会工学系教授）
研究分担者：吉田武稔（北陸先端科学技術大学院大学 知識科学研究科助教授）
研究分担者：鈴木秀男（筑波大学社会工学系講師）

研究経費

平成１１年度	２，５００千円
平成１２年度	１，１００千円
計	３，６００千円

研究発表

（1）学会誌等

1. G. Dupret and M. Koda, "Bootstrap Training for Neural Network Learning," 「確率数値解析における諸問題 Ⅳ」京都大学数理解析研究所講究録 **1127**, pp. 27-35, 平成 12 年 1 月.
2. M. Koda and H. Okano, "A New Stochastic Learning Algorithm for Neural Networks," Journal of the Operations Research Society of Japan, Vol. 43, No. 4, pp. 469-485, December 2000:
3. G. Dupret and M. Koda, "Bootstrap Re-sampling for Unbalanced Data in Supervised Learning," European Journal of Operational Research, in press.
4. T. Yoshida and H. Touzaki, "A Study on Association among Dispatching Rules in Manufacturing Scheduling Problems," Proc. 7th IEEE Int. Conf. Emerging Technologies and Factory Automation, pp. 1355-1360, October 1999.
5. H. Suzuki, "Recognition of Unnatural Patterns in Manufacturing Processes Using the Minimum Description Length Criterion," Communications in Statistics: Simulation and Computation, Vol. 29, No. 2, pp. 583-601, December 2000.

（2）口頭発表

1. M. Koda, "Bootstrap Training for Neural Networks," Watson KDD PIC Seminar Series, IBM T.J.Watson Research Center, New York, August 13, 1999.

2. M. Koda, "Status of Industrial Applications of Data Mining Technology in Japan," Management Technology Workshop, IBM GBIS Technology Center, Dallas, August 17, 1999.
3. G. Dupret and M. Koda, "Bootstrapping for Neural Network Learning," 日本オペレーションズ・リサーチ学会 1999 年度秋季研究発表会, 東京, 平成 11 年 9 月 21 日.
4. 香田正人, "知識発見科学としてのデータマイニング（チュートリアル）" 応用統計学会チュートリアルセミナー「データマイニング」, 東京, 平成 11 年 10 月 26 日.
5. M. Koda, "Sensitivity Analysis in Data Mining," Department Seminar, Dept. of Experimental Science and Technology, Rey Juan Carlos University, Madrid, November 18, 1999.
6. M. Koda, "Stochastic Neural Network Formulation using SDE," Statistics and Operations Research Seminar Series, Pompeu Fabra University, Barcelona, November 22, 1999.
7. 香田正人, "データマイニングの産業応用の現状について（基調講演）" 「高度情報機器を用いた新たな実態調査手法の開発に関する研究」, (財) 計量計画研究所, 東京, 平成 12 年 5 月 8 日.
8. G. Dupret and M. Koda, "Bootstrap Re-sampling and Cross-Validation for Neural Network Learning," 応用統計学会データマイニング研究部会, 東京, 平成 12 年 6 月 12 日.
9. M. Koda, "An Overview of Industrial Applications of Data Mining and Knowledge Discovery in Japan," Dept. Industrial Systems Eng. and OR Society of Singapore Joint Seminar, The National University of Singapore, Singapore, July 4, 2000.
10. G. Dupret and M. Koda, "Bootstrap Training for Neural Network Learning," 5th Conference of APORS, Singapore, July 7, 2000.
11. 岡野裕之, 香田正人, "確率雑音反応法による連続系での最適化" 日本オペレーションズ・リサーチ学会 2000 年度秋季研究発表会, 東京, 平成 11 年 9 月 27 日.
12. 山部浩司, 八巻智, 山本良次, 香田正人, "決定木を用いた複合学習モデルについて" 日本オペレーションズ・リサーチ学会 2000 年度秋季研究発表会, 東京, 平成 12 年 9 月 28 日.
13. M. Koda, "An Overview of Industrial Applications of Data Mining and CRM in Japan," Adaptive Friday Program, Adaptive Information Systems and Modeling in Economics and Management Science, Vienna University of Economics and Business Administration, Vienna, November 8, 2000.
14. 岡野裕之, 香田正人, "白色雑音を用いる勾配法とそのＴＳＰへの応用" 情報処理学会第 77 回アルゴリズム研究会, 横浜, 平成 13 年 3 月 12 日.
15. 香田正人, "発見科学やデータマイニング技術の産業応用について" 日本オペレーションズ・リサーチ学会第 206 回新宿ＯＲ研究会, 東京, 平成 13 年 3 月 21 日.

研 究 成 果

目　　次

研究成果の概要

本研究の目的は，データマイニングに関して既に得られている数理モデルの特長を拡張し，従来のデータベース技術や統計的手法の利点を活用して，新しいデータマイニング・モデルの構築と解析を行い，さらにシミュレーションやテストデータにより評価・検証して，モデルの特性を明らかにすることであった．この目的に沿って研究を行なった結果，下記の結果を得た．

（1）確率感度解析による新しい確率的学習アルゴリズムの導出

データからの知識発見のために，ニューラルネットワークなどの機械学習手法を用いたデータマイニング・モデルにおいて，データの不確定性やノイズ，予測不可能な変動などがモデルの推論結果に与える影響の度合いを確率感度解析によって定量的に評価し，それにもとづいたニューラルネットワークの新しい確率的学習アルゴリズムの定式化に成功した．

（2）ブートストラップ手法の最適リサンプリングへの応用

データマイニングの実際の応用においては，学習用データが統計的に偏ったり不充分な量の教師データしか得られないケースによく遭遇するが，統計的に大きく偏ったデータからの知識発見のために，復元を許したリサンプリング手法であるブートストラップ法の適用を提案し，シミュレーションによりその有効性を検証した．さらに，学習データの複合的組み合わせを繰り返して交叉検証(Cross-Validation)を行う場合にも，本提案手法を拡張してデータマイニング・モデルの汎化能力を改善できることを明らかにした．

（3）データマイニングの製造スケジューリングや工程管理への応用

データマイニングの新しい応用分野として，生産データ解析における相関（Association）分析やＭＤＬ（最小記述長）などの情報量基準の適用について考察を行ない，最適スケジューリングや工程管理手法を提案した．本研究で提案した手法の有効性を数値実験により定量的に示した．