JS

データマイニングーその発展と今後の展望ー

（阪大産研）鷲尾　隆*

1.　データマイニングの発展

実用化されるデータマイニング

データマイニング技術は９０年代前半に基礎研究が本格化したが、ここ５年程の間に産業界で広く実用化されている。主に金融や流通などでマーケティング調査への適用が主流であったが、最近では製造業や通信サービス業における品質管理や顧客管理などへの適用も進んでいる。このような急速な普及の背景として、データデジタル化やデータベース化による豊富な対象データ蓄積及び既存の統計的手法と新しいデータマイニング手法の融合、統合化マイニングツールの市販などが挙げられる。 これに対して、化学や薬品、医療分野へのデータマイニングの適用はようやく始まったばかりである。生命科学分野でもゲノム解析などで早くから適用が試みられているが、まだ顕著な成果には結びついていない。これは分野の特性がマーケティングや管理とは大きく異なるためである。１つは化学構造式やたんぱく質構造、ゲノム配列、治療履歴の不定期時系列データなど、取り扱いが困難な構造や性質を持つデータが多いという技術的困難のためである。もう１つは、データマイニング解析に加えて分野特有の極めて高度な専門知識が要求され、解析と分野専門知識双方に精通した人材が少ないというマネジメント的困難のためである。

ここでは、はじめにデータマイニングの産業界適用の現状について概観する。次に既存データマイニング技術の医療分野への適用例を概説する。更に化学構造などの構造化データに関する先端的技術の化学分野への適用例について概説する。これらを通じて分野専門家とデータマイニング技術者の協力の重要性が明らかとなる。最後にデータマイニングの今後を展望する。

産業界の適用現状

表１に各界における代表的適用事例を掲げる。公開適用事例として、もっとも多く発表されているのは金融分野である。米国では、1994 年頃から報告されているが、日本でも近年は多くの事例が報告されている。この分野はリレーショナルデータベースを中心とする表形式データが多く、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、分類決定木、ラフ集合、重回帰分析、

表１　データマイニングの代表事例

金融分野
・マーケティング分野 潜在的住宅ローン申込み顧客の推定，顧客に応じた銀行商品設計提示支援，生命保険潜在的解約候補顧客発掘，ダイレクトメール宛先候補顧客の発掘 ・業務特化分野 ローン審査ルール発掘，リスク細分型保険設計提示支援，社債格付け推測，クレジット・カード不正利用推定
流通・小売分野
薬局販売データからの優良顧客発掘，立上がり売行きに基づく新製品販売予測，新製品ヒット要因分析，売行き要因分析，牛乳販売量予測，消費者購買行動分析，販促下の併売パターン分析
製造分野
顧客意見収集による次世代新製品開発，設計・製造現場への品質管理要求発掘，製造条件と検査結果の対比による工程改善
通信分野
顧客プロファイリングと顧客傾向分析，電話回線負荷状況把握や障害診断，通信需要把握のためのトラフィック分析，通話回線不正使用検出，アクセスログに基づく不正アクセス検出
化学・製薬・生物・医療分野
分子構造と生理活性の相関解析，遺伝子発現と生理学的効果の相関解析，科学的根拠に基づく医療の知識獲得

記憶ベース推論など、多様なデータマイニング技術が用いられている。

流通分野では、小売部門のマーケティングのためのデータマイニング適用が主流であり、分類決定木、バスケット分析、重回帰分析、相関解析などの技術が用いられている。ただし、金融分野に比べて扱う商品や小売条件、顧客行動パターンが遥かに多様であり、顧客や購買事例を適切に類別して特徴を発掘することに難しさがある。

製造業固有のデータマイニング適用は、これまであまり多くの事例が見あたらなかったが、最近になってサプライチェーンマネジメントの広まりから、原料調達から工場生産、顧客へのデリバリーまで一貫した記録情報が蓄積され、これから顧客使用段階での不具合発生まで視野を広げた、製品品質や製造工程、流通工程の管理への適用が広まり出している。この分野では，事例ベース検索やテキストマイニング、バスケット分析、分類決定木などの技術が用いられ、効果を上げている。

通信分野では，主にインターネット網や電話網

*E-mail washio@ar.sanken.osaka-u.ac.jp

管理の分野にデータマイニング技術が用いられている。使用技術は分類決定木、バスケット分析、ベイジアンネット、ニューラルネット、テキストマイニング、各種統計的手法など多岐に亘る。通信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可能性は高い。

使用技術と実施体制

産業界で用いられているデータマイニング技術は多様である。必ずしも先端技術のみが用いられているわけではなく、伝統的統計手法も多用されている。採用される技術は、各目的やニーズ、データ仕様のみならず、データマイニングツールの開発者やユーザーによっても左右される。特にわが国では製造、通信分野の開発者やユーザーは技術的蓄積を有し、市販ツールを利用するだけでなく、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構築することが多い。これは効果的なデータマイニングの実現には、単独技術だけではなくデータの前処理や結果の後処理も含め、各事例に適した先端、既存技術の組み合わせや条件設定を必要とすることが多いためである。

このような特徴は、技術系の研究者や技術者を多く抱え、独自のデータ表現や処理方法が研究開発されている化学、薬品、生物、医療分野についても当てはまると考えられる。またはじめに述べたように、取り扱い対象データ構造が特殊で先端の解析技術が要求されることが多く、分析に深い分野専門知識が要求されることから、分野専門家とデータマイニング解析者間の緊密な協力の下で、研究開発的体制を敷くことが重要となる。

2. 既存技術とその適用例

科学的根拠に基づく医療（ＥＢＭ）への適用

データイニングで用いられる技術は大きく分けても数十種類以上あるが、その中で部分的傾向条件に基づき事例が属するカテゴリを推定する手法を分類(Classification)手法という。この手法は疾患の推定規則や薬品の作用内容・程度の推定規則の発掘などに用いることが可能である。そのなかでも代表的な決定木分類学習法 C4.5[1]を、科学的根拠に基づく医療(EBM)のための基礎知識収集に適用した例を述べる。これは国立循環器センターの医師チームと筆者等の共同分析である[2],[3]。この事例の成功は、現場医師等の強い問題意識や目的設定、データ収集整備といったマネジメント体制と、データマイニング解析者による技術協力よってもたらされたものである。

決定木分類学習法 C4.5 は図１に示すように、元

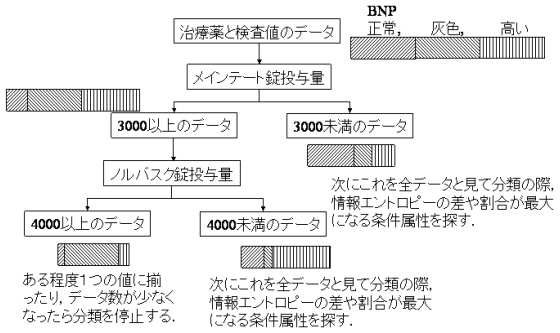

図１　決定木分類学習法の概要

データの各事例（治療投薬パターン）のカテゴリ（治療効果）の平均的分布に比べて、より特定のカテゴリに分布が集中する事例が満たす条件項目をデータから見つけ出すことを再帰的に繰り返し、部分的傾向により一定基準以上分布が偏る条件の組み合わせを推定規則として導出する。

国立循環器センターにおける入院心疾患患者の退院時投薬・治療歴データ２７４件につき、この手法を適用した。説明条件項目として薬効を規格化した 52 グループの治療時投薬品群と量を、目的カテゴリとして BNP, LAD, %FS, LVDd などの心機能検査指標の各'低い', '正常'、'灰色'、'高い'、'異常に高い'などの分類を取った。得られた推定規則から医師が読み取った知見の一部を表２に示す。各指標について、○が効果あり、△が中立、×が悪影響あり、-が不明を表す。既に医師が知っている多くの知識と共に、知られていなかった知識が仮説として得られた。特に糖尿病治療薬と消化性潰瘍剤は、その後、動物実験などでも効果が検証され、新薬の開発に結びつきつつある。

表２(a)　発見された知識：既知の知識

	BNP	%FS	LVDd
β-blocker	○	-	○
Ca blocker	△	○	-
ACEI	△	○	-
ARB	-	○	-
坑アルドステロン	×	×	-
抗血小板	○	-	×
スタチン	△	-	-
強心剤	×	×	-

表２(b)　発見された知識：新しい知識

	BNP	%FS	LVDd
糖尿病治療薬	-	○	-
消化性潰瘍剤	○	-	-
潰瘍治療薬	○	-	-
制酸剤	×	-	-

3. 先端技術とその適用例

構造活性相関解析とデータマイニング技術

構造活性相関解析では、複雑な化学構造式と化学的活性との関係分析が行われる。代表的な手法はQSARである。分子構造データに前処理を施し、ベンゼン環を有するか否かといった構造を特徴づける多数の記述子に変換してから、それらと化学的活性の相関解析を行う。また計算機科学分野でも、同様に記述子を分子の持つ属性や命題とみなし、C4.5のような分類決定木やM5などの回帰分類木、帰納論理プログラミングを適用する方法が試みられている。これらは化学分野の専門家が着目する部分構造に絞り込んだ解析を可能にする半面、人為的に取り上げた部分構造の組み合わせしか抽出できない。これに対し、分子構造上の各パスに沿う原子質量数スペクトルを元にグラフとして分子構造を特徴づけ、検索や類似構造判定を行う手法も提案されている。これは検索や類似性判定の完全性を保証しないものの、全体的なグラフの特徴を表すスペクトルによって高速処理できる実用的方法である.

より直接的に分子構造を扱う手法として、原子鎖に沿う特徴によって目的カテゴリに分類する相関規則を導出するカスケードモデル手法がある[4]. この手法は化学的活性レベルなどの分類を特徴づける精度の高い相関ルールを導出可能である。更に特徴的原子鎖をほぼ完全探索する手法としては、化学分野でCASE及びMultiCASEが提案されている。これは化学分子構造の解析に特化された手法を採用しており、効率的に部分構造を発見できる反面、変則的な構造については若干見落としの可能性がある。これに対して計算機科学分野では、指定した頻度閾値区間に含まれる頻度でグラフデータに表れるパスを完全探索する手法も提案されている。これは一般グラフにおいて完全探索できることが保証されている手法であり、化学分子構造についても原子鎖を完全探索可能であるが、その分膨大な計算を必要とする。

以上は、何らかの近似やある種類の部分構造に限定してグラフ構造を扱う手法であるが、近年、一般グラフの多頻度パターン及び相関規則の完全探索を高速に行うグラフマイニング手法として、グラフの隣接行列表現と相関規則を組み合わせたApriori-based Graph Mining(AGM)手法が開発されている[5]。これは指定された頻度閾値を超える多頻度部分グラフを完全探索する。この探索では多くのグラフ同型問題を解かねばならず、計算量の面で困難を生じる。そこで図２に示すようにグラフを隣接行列として表し、より部分のパターンはそれを含むパターンより出現頻度が必ず高

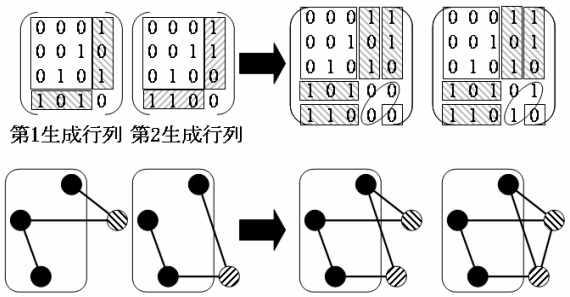

図２　Apriori-based Graph Mining(AGM)の原理

いか少なくとも同じであるという頻度の単調性を利用して問題を解く。グラフ上の第iノードを行列のi行及びi列に対応させ、サイズ（ノード数）kのグラフをk×kとして表す。行列の{i,j}要素で第iノードと第jノード間のリンクの有無及び種類を表す。第１生成行列で表されるグラフと第２生成行列で表されるグラフの両方が多頻度であり、かつ最下及びの最右の行と列のみが異なる行列、すなわち１つのノードのみがトポロジカルな位置が異なる２つグラフをマージし、１サイズ大きな多頻度グラフの候補を生成する。多頻度グラフには，頻度の単調性から第１及び第２生成行列で表されるような１ノードのみ異なる部分多頻度グラフは必ず存在する。したがって、逆に第１及び第２生成行列で表される多頻度グラフの合成から生成される１サイズ大きなグラフは、多頻度である可能性が高い。そこでこのような合成を、上記条件を満たすすべての第１及び第２生成行列の合成によって生成すれば、取りこぼしなく１サイズ大きな多頻度グラフ候補を得ることができる。図２に示すように合成後に元々共通ではなかったノード間にリンクを張るか張らないかによって、またどの種類のリンクを張るかによって、複数の候補が生成される。候補生成後は、それらとグラフデータを隣接行列の性質を使って効率的に照合し、多頻度か否かを確認する。このようにして小さい多頻度部分グラフからより大きい多頻度部分グラフを、それ以上新たな多頻度部分グラフが見つからなくなるまで逐次探索していく。

構造活性相関解析への適用

図３に示すような多数の化合物分子構造と変異原性活性の有無に関する構造活性相関解析に、上述のAGM手法とカスケードモデル手法を適用した。これもカスケードモデル手法を開発しかつ化学分野の専門知識が豊富な研究者との共同研究であり、対象データ分野に関する知識が非常に重要な役割を果たした例である[6]。このデータは変異原性活性が無い、低い、中程度、高いという４レベルに区分けされている。AGM 手法によって、データ中から多頻度の部分分子構造をマイニ

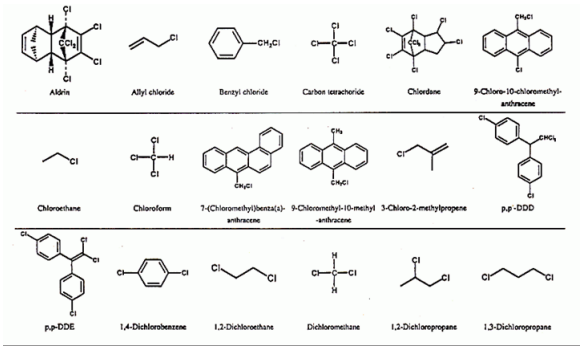

図３　変異原性活性に関する化合物分子構造

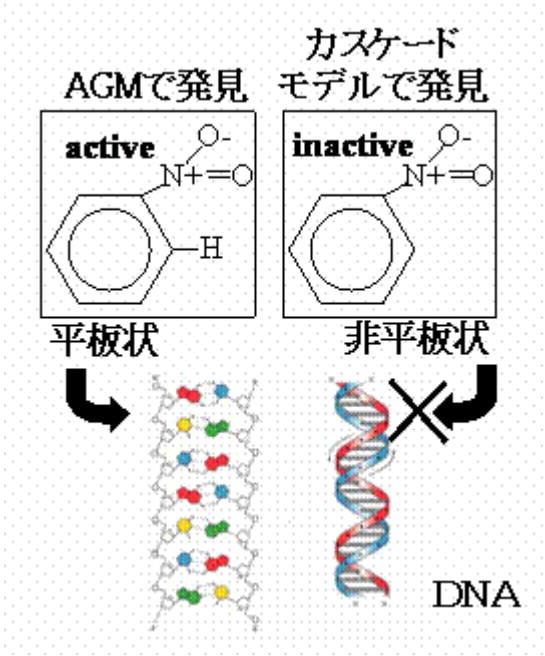

図４　グラフマイニングによる結果

ングし、それらの部分構造が何れの活性をデータ中何%にもたらすかを調べた。一方、カスケードモデル手法により、原子鎖に沿う特徴によって変異原性活性の４レベルに関する分類規則を得た。

図４に結果の１例を示す[6]。これは多数のマイニング結果パターンから、専門家の分析によって注目すべき組合せを選び、更に考察を加えたものである。このような専門知識の駆使無しに、マイニング結果からだけでは価値の高い知識を導くことは困難である。また、このように２種以上の性格の異なる手法を適用し、それらの結果を比較することで注目すべきパターンが得られる場合もある。左右２つのパターンの顕著な違いは、負電荷を帯びた NO_2 基の側に正電荷を帯びた水素基があるか否かである。水素基がある場合、正電荷に NO_2 基は引かれ、その面はベンゼン環平面に近い配位を取る。従って、この部分の構造が平面状になり遺伝子鎖の二重螺旋内にはまり込みやすくなり、高い変異原性を示すと考えられる。これに対し後者は、NO_2 基面が必ずしもベンゼン環平面に一致せず、立体障害によって二重螺旋内にはまり込み難いため、低い変異原性を示すと考えられる。

4.　今後の展望

従来の統計解析では、人間が立てた仮説をモデルを通じて検定する。これに対して、　データマイニングはデータから仮説候補を導出する。専門家はそこから仮説を選択し、何らかの方法で検証を行う。この従来にない機能により、データマイニング技術は必ずしも膨大なデータへの適用に限らず、既に種々の産業界で使われている。

本報で紹介したように、既存のデータマイニング手法でも、今後多くの成果が期待される。一方、構造データのマイニングのように、より複雑なデータに関するデータマイニング技術が急速に発展しつつある[7]。これをベースに各分野に適した手法やデータ処理のノウハウを確立するための努力を、研究レベル、現場レベルで積み上げることにより、更に多くの成果が期待される。化学、製薬、医療分野へのデータマイニングの適用も進みつつあるが、これらの分野には既に種々のデータ分析技術・人材の蓄積があり、今後、データマイニング手法の高度化に伴って、一層、実用化が進むと期待される。

参考文献

[1] J.R. Quinlan: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers (1993).

[2] 鷲尾　隆, 金　智隆, 北風政史: データマイニングとその Biomedical Engineering への適用, 日本エム・イー学会誌, 生体工学, Vol.41, Supp1.1, S4-1, pp.32 (2003).

[3] 金　智隆, 磯村　正, 鷲尾　隆, 北風政史: 医療情報に対する新しいデーや解析手法の適用, 日本エム・イー学会誌, 生体工学, Vol.41, Supp1.1, S4-4, pp.35 (2003).

[4] Takashi Okada: Datascape Survey using the Cascade Model, Discovery Science 2002, pp.233-246, (2002).

[5] Akihiro Inokuchi, Takashi Wasiho and Hiroshi Motoda: Complete Mining of Freqent Patterns from Graphs: Mining Graph Data, Machine Learning, Vol.50, pp.321-354 (2003).

[6] Akihiro Inokuchi, Takashi Washio, Takashi Okada and Hiroshi Motoda: Applying the Apriori-based Graph Mining Method to Mutagenesis Data Analysis, Journal of Computer Aided Chemistry, Vol.2, pp.87-92 (2001).

[7] Takashi Washio and Hiroshi Motoda: State of the Art of Graph-based Data Mining, ACM, SIGKDD Explorations, Vol.5, Issue 1, pp.59-68 (2003).