データマイニング手法の検討を行うための支援業務

報告書

平成17年3月

株式会社三菱総合研究所

目次

図　目次

表　目次

1. 調査の概要

1.1 調査目的

医薬品医療機器総合機構の中期計画（平成 16 年度～平成 20 年度）では、収集した副作用情報等を用いて副作用を早期に発見し、その未然防止策を講ずるためデータマイニング手法を研究し平成 18 年度までに手法を確立し、中期目標期間終了時までに安全対策業務に導入すると定められている。また、当該計画のうち平成 16 年度の計画では、新規手法の導入に向けた検討として、データマイニング手法についての検討を開始し、検討状況について適宜公表するとされている。

本調査は中期計画期間内に導入するデータマイニング手法を明確にし、さらに中期計画期間内の実施スケジュールを策定するためのものである。

1.2 調査内容

調査内容の概要は以下のとおりである。

1. 医薬品副作用情報へのデータマイニング適用検討のための基礎調査

医薬品副作用情報へのデータマイニング適用について検討するための基本情報を得るため、以下の調査分析を行なう。

（1）業務分析

データマイニング適用の対象となる安全分析業務の担当者に対してヒアリングを実施し、業務分析を行なう。これにより、データマイニングによる支援が期待される業務と支援内容を明らかにする。

（2）データマイニング基礎事項整理

データマイニングの基礎事項として、データマイニングの概念、一般的な定義、代表的な手法について調査し、整理する。

（3）異業種成功事例の調査・分析

医薬品副作用情報へのデータマイニング適用検討の参考情報とするため、異業種における類似業務へのデータマイニング導入成功事例について調査する。

（4）諸外国規制当局等の当該業務の現状調査

WHO、米国 FDA、英国 MHRA など諸外国規制当局における医薬品副作用情報の分析業務の現状について調査を行う。シグナル検出に関する検討状況についても調査する。

2.　医薬品副作用情報分析におけるデータマイニングの概念検討

医薬品副作用情報の分析業務に適用するデータマイニングの概念について、中長期的な視点を交えながら検討を行う。

（1）医薬品副作用情報におけるデータマイニングの概念検討

医薬品医療機器総合機構が考える医薬品副作用情報分析におけるデータマイニングについて、調査結果に基づき検討し、概念を明らかにする。

（2）データマイニング適用範囲の検討

検討したデータマイニングの概念の具体的な内容として、医薬品副作用情報の分析業務プロセスのうち、データマイニングの適用による支援が期待される範囲について検討を行う。

（3）シグナル検出からデータマイニングへのアプローチの妥当性検討

現在、医薬品副作用情報分析におけるデータマイニングとして検討が進められているシグナル検出と、2（1）で検討したデータマイニング概念との関係を明らかにすることにより、医薬品医療機器総合機構が考えるデータマイニングの概念の妥当性を示す。

3.　中期計画期間内目標の設定と具体策検討

2 で検討を行った医薬品副作用分析業務へのデータマイニング適用のうち、中期計画期間内に実施する目標を設定し、目標達成のための具体的な方法を検討する。

（1）中期計画期間内目標設定

医薬品副作用分析業務へのデータマイニング適用について、中期計画期間内で実施する目標を設定する。

（2）中期計画目標達成のためのスケジュール策定

中期計画期間内の目標を達成するための、各年度の具体的なスケジュールを策定する。

（3）医薬品副作用情報のあり方に関する検討

データマイニングによる分析を有効なものとするために期待される医薬品副作用情報の情報項目、および入力方法等のあり方について検討を行う。

（4）システム検討

データマイニングを実施するのに必要となる処理能力を充たすためのシステム構成、導入計画について、今後導入が予想される技術内容を視野に入れながら検討する。

1.3 検討委員会

4 章でまとめる中期計画期間内目標の設定と具体策の検討にあたっては、以下の委員からなる「データマイニングに関する検討委員会」を設置し、データマイニングの内容、中期計画期間内の目標、および実施スケジュールの妥当性について検討を頂いた。

【データマイニングに関する検討委員会】

座　長：	藤田　利治	国立保健医療科学院　疫学部　疫学情報室　室長
委　員：	岩崎　学	成蹊大学　工学部　経営・情報工学科　教授
	岡田　孝	関西学院大学　理工学部　教授
	岡田　美保子	川崎医療福祉大学　医療情報学科　教授
	酒井　弘憲	日本製薬工業協会 医薬品評価委員会 統計･DM 部会 副部会長
	櫻井　靖郎	財団法人　日本公定書協会　JMO 事業部　事業部長
	野口　茂	日本製薬団体連合会　安全性委員会　委員
	平山　佳伸	厚生労働省　医薬食品局安全対策課　課長
	望月　眞弓	北里大学 薬学部 臨床薬学研究センター 医薬品情報部門 教授
	鷲尾　隆	大阪大学　産業科学研究所　助教授
	渡邊　伸一	厚生労働省　医薬食品局安全対策課　課長補佐

2. 医薬品副作用情報へのデータマイニング適用検討のための基礎調査

2.1 業務分析

医薬品副作用情報へのデータマイニング適用について検討するにあたり、データマイニングを適用する対象となる安全対策業務について、その業務フロー、体制等について分析を行なった。結果を以下にまとめる。

2.1.1 業務フロー

医薬品副作用情報分析に係る業務の流れのうち、特に医薬品医療機器総合機構の安全対策業務に関わる部分の概要を図 2-1 に示す[1]。

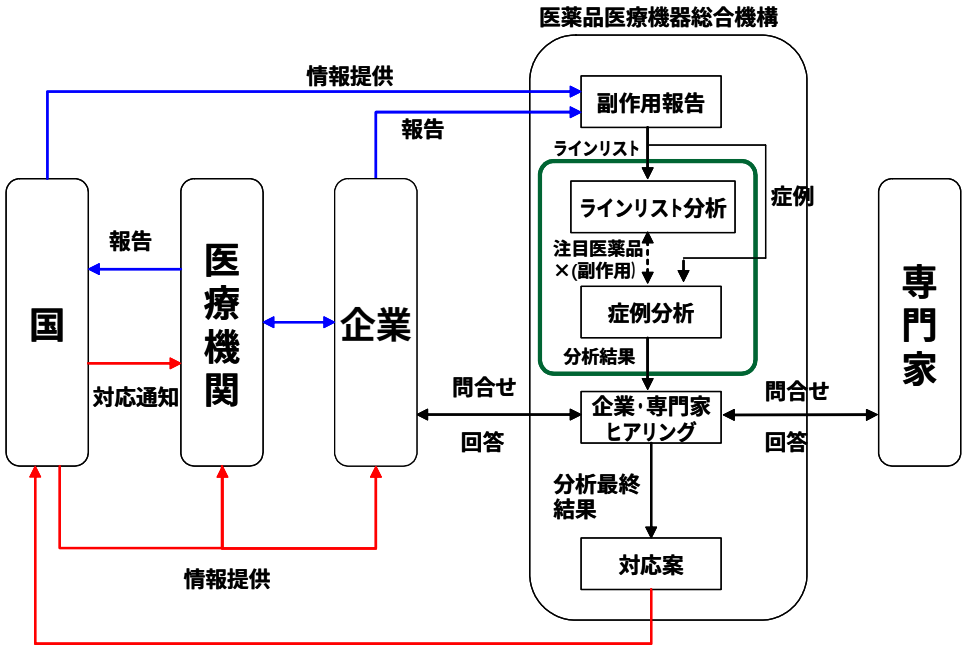

図 2-1 医薬品副作用情報分析に係る安全対策業務の流れ

安全対策業務は以下のプロセスから構成される。

1. 副作用報告収集：企業、医療機関から報告された副作用情報をデータベースに登録する。登録された情報の中から目的に応じて項目を抽出し、日刊、週刊の「ラインリスト」として分析担当者に提示する。FAX 報告は、分析担当者が随時症例票を確認する。
2. ラインリスト分析：ラインリストをもとに、安全性の観点から対応が必要となる

[1]安全対策業務の枠組み全体についての詳細は、医薬品医療機器総合機構ホームページ（http://www.pmda.go.jp/sinsaanzen/anzen_1.html）に公開されている。

可能性がある医薬品および副作用をチェックする。

3. 症例分析：ラインリスト分析の結果、さらに詳細な分析が必要であると判断された医薬品と副作用の組については報告の詳細（症例）を参照しながら、分析を行なう。
4. 企業・専門家ヒアリング：分析の過程で必要に応じて企業に対してヒアリングを行う。企業から、対応したい旨の連絡が持ち込まれることもある。また、専門家に対して詳細な分析を依頼する。
5. 対応決定：分析の結果、対応が必要であると判断されるものについて、添付文書の改訂案の作成等必要な対応を行なう。

このうち、主にデータマイニングの適用対象となる安全分析業務は、ラインリスト分析と症例分析である。安全分析業務を行うにあたっては、最初は全ての医薬品・副作用が検討の対象となるが、このうち詳細調査が必要な医薬品・副作用をラインリスト分析により抽出し、さらに個別症例に基づき分析を行なう症例分析により、対応が必要な医薬品・副作用を抽出する（図 2-2）。すなわち、ラインリスト分析は主に詳細な医薬品・副作用を絞り込むための分析、症例分析はラインリスト分析により絞りこまれた医薬品・副作用に関する詳細な情報をもとに対応の必要性を判断するための分析と見ることができる。

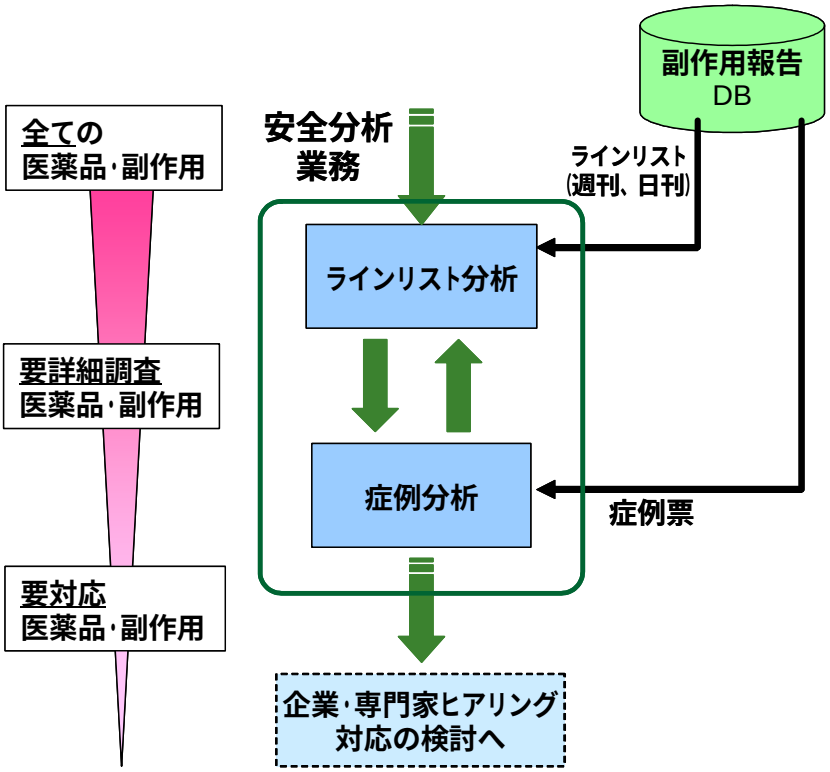

図 2-2　安全分析業務における分析の流れ

ただし、全ての安全分析業務がここで示したラインリスト分析から症例分析へという流れで行なわれているわけではない。症例分析を行なったあと、再びラインリスト分析を行なう場合もあるし、ラインリスト以外から得られた情報をきっかけとして、ラインリスト分析を行なわず直接症例分析を行なう場合もある。

2.1.2 業務へのデータマイニング適用

データマイニングを適用する対象となる業務はラインリスト分析に相当する部分であり、全ての医薬品・副作用の組合せの中から詳細な分析を要する医薬品・副作用を抽出する部分であると考えられる。

ラインリスト分析では、上で述べたとおりラインリスト上の情報を中心に、当該医薬品に見られる同種と推定される副作用の抽出を行なっている。これは、近年、安全分析業務におけるデータマイニングの１つとして注目を集めている対応が必要な医薬品と副作用の組を抽出するシグナル検出手法と同じ処理を行なっているプロセスである。また、現在分析担当者が注目している内容も累積の報告数が多くなった場合、あるいは報告が急増した場合など、比較的データマイニングのアルゴリズムに実装しやすいものである。

一方、業務分析を行なった結果、データマイニングを適用していく上での課題があることが分かった。例えば、分析担当者がラインリストを分析する際には、当該医薬品・副作用に関する過去の対応等の経緯、あるいは症状のコーディングや報告の件数等に含まれるバイアスに関する知識など、ラインリスト以外にも経験等から得た知識を導入しながら実施している。このような知識をデータマイニングのプロセスにいかに取り込んでいくかという点がデータマイニング導入を成功させるための鍵となる。

データマイニングによる分析は、現在の人手による安全分析業務の補助的なツールであり、市販直後調査中品目など特に注目すべき項目については、従来どおりの分析をあわせて行うことにより、さらに質の高い安全分析業務が可能となる。特に、データマイニングによる分析を行なうことにより、現在よりもさらに早期にシグナルを検出する可能性が高まると期待される。また、複数の医薬品を同時に服用している場合の相互作用による副作用発生等の抽出ができると安全分析業務の支援になる。

2.2 データマイニング基礎事項整理

医薬品副作用情報分析業務に適用するデータマイニングについて検討するための基礎資料として、一般的なデータマイニングの定義、データマイニング手法等について調査を行った。以下に調査結果をまとめる。

2.2.1 データマイニングの定義

データマイニングは、計算機の処理能力が向上し、ネットワーク環境が整備され、さらには大容量の記憶装置が安価になったことから、大量のデータを蓄積することが比較的容易になったことを背景に、1990 年代の半ば以後、大量のデータを分析するための技術として注目を集めている技術である。

データマイニングに関連する概念の定義は様々であるが、一般的なものとして以下のようなものがある[11]。

表 2-1　データマイニングに関連する定義

■ データベースからの知識発見(Knowledge Discovery in Databases: KDD) 「データに含まれる、正確、斬新で、潜在的に有用で、かつ究極的に理解しやすいパターンを見出す過程」 ■ データマイニング（data mining） 「データベースからの知識発見」を構成するプロセスのうち、計算効率（時間）として許容できる範囲で動作するデータ分析・知識発見アルゴリズムを用いて、データに含まれているパターンの一覧を抽出するプロセスを指す。

一般にデータマイニングという用語は、表 2-1 の「データベースからの知識発見」に定義される内容を指す場合と「データマイニング」を指す場合がある。「データベースからの知識発見」を広義のデータマイニング、「データマイニング」を狭義のデータマイニングと呼ぶことにする。広義のデータマイニングと狭義のデータマイニングの違いは、前者がデータから有用なパターン（知識）を抽出するプロセス全体をデータマイニングと呼んでいるのに対して、後者はどちらかというとデータに対して手法を適用し、結果を得る部分のみを指している点にある。

「マイニングする」といったように使われている場合には、「マイニング」は広義のデータマイニングを意味している。また、「データマイニングツール」といったように、ツール、あるいはソフトウェアを指す場合には、どちらかというと狭義のデータマイニングを意味している。

このように、データマイニングという用語は、広義、狭義それぞれの意味で用いられて

いる場合があるので、どちらの意味で用いられているのかを正しく理解し、また明確に用語を使い分けていく必要がある。

2.2.2 データマイニングの判別基準

上であげたデータマイニングの定義の他にも、当該内容がデータマイニングであるかどうかを判別するための基準として以下のようなものが存在する。

- データ量：分析の対象とするデータの量が多いこと。特に人手では処理できないようなデータを分析する場合には、データマイニングと判別することがある。データマイニングの多くが大量のデータを扱うことに基づいている判別基準である。
- 手法：後述するデータマイニング手法の代表的なものである決定木、相関ルール分析などを分析に利用している場合、データマイニングと判別される。これらデータマイニング手法の多くは、大量のデータの分析に適したアルゴリズムを採用している。
- 自動的：半自動的な分析により知識（ルール）が抽出されていること。一部では、データマイニングの特徴として統計分析のように専門的な知識を必要としない点が強調される場合がある。

前節であげた定義も含めた一般的なデータマイニングの判別基準を表 2-2 に整理する。データマイニングと呼ばれるプロセスは、多くの場合これら何れかの基準に当てはまる。

表 2-2 当該内容がデータマイニングであるかを判別するための基準

判別基準	内容
【定義】 未知かつ有用な知識を発見できている	結果として業務効率、成果が向上している
【データ量】 大量のデータを分析している	人手では見切れないようなデータを分析している
【手法】 データマイニング手法（ツール）を使っている	決定木、ニューラルネット、相関分析などを使用している、あるいは市販ソフトを利用している
【自動的】 半自動的な分析により知識（ルール）が抽出されている	データをツール（ソフト）に入れると知識が抽出される

2.2.3　データマイニングプロセス

広義のデータマイニング（データベースからの知識発見）のプロセスを図 2-3 に示す。広義のデータマイニングは次のようなプロセスからなる。

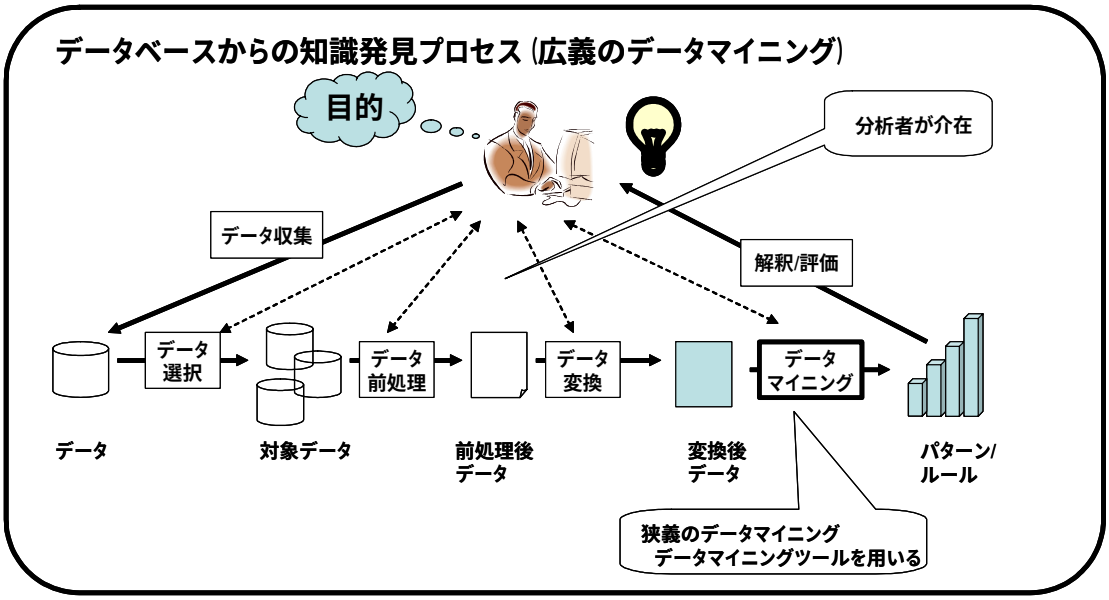

図 2-3　データベースからの知識発見（広義のデータマイニング）のプロセス

1. データ収集：データ分析に利用可能なデータを収集する。
2. データ選択：データ分析（データマイニング）に利用可能なデータから、データマイニングの目的に照らして、必要なデータを選択する。選択されたデータを対象データと呼ぶ。
3. データ前処理：対象データに欠損値、誤り値が含まれる場合には、データの特性に関する知識が得られており欠損値、誤り値を回避することができるならば、データ前処理として欠損値、誤り値を修正し、分析のためにより質の良いデータとする。データ前処理が完了したデータを前処理後データと呼ぶ。
4. データ変換：データの形式、およびデータの値そのものを、データマイニングアルゴリズムに適したように変換する。変換されたデータを変換後データと呼ぶ。
5. データマイニング：目的に適したパターンを抽出することができるデータマイニングアルゴリズムを準備し、これに準備したデータを入力する。データマイニングアルゴリズム（ツール）は、ルール、パターンを抽出する。
6. データマイニングツールにより出力されたルール、パターンを分析者が解釈、評価し、有用な知識を得る。

ここでデータマイニングツールが抽出したルールやパターンが最終的なデータマイニングの結果ではなく、分析者がそれらを解釈、評価した結果として知識を得るまでがデータマイニングであるということに注目することが必要である。データマイニングツールが出力するルールやパターンは、それだけで示唆するものが多いが、さらに分析者が経験から得ている知識を持ってこれらを解釈することにより、さらに深い知識を得ることができる。すなわち、広義のデータマイニングとは、自動的なプロセスというよりは、分析者を支援するマニュアル的なプロセスである。

また、図 2-3 ではデータマイニングの各プロセスを一方向なプロセスとして表現しているが、実際は各プロセスで試行錯誤を行い、以前のプロセスに戻ることも少なくない。データマイニングを通じてより良い質の知識を得るためには、データをただデータマイニングツールにかけるのではなく、データに含まれるバイアスなどの特性を知り、データ選択からデータ変換までを含むデータの前処理に関するプロセスを十分に行うことが重要である。

2.3 データマイニング手法

2.3.1 データマイニングのタスクと手法

データマイニングでは大量のデータを扱うことが多いために、大量のデータから効率良くパターンやルールを抽出することができるアルゴリズムが開発､利用されている｡また、データマイニングでは、最終的に抽出されたパターンやルールを分析者が解釈できることが重要であるため､抽出されたそれらが分析者にとって理解できるものである必要がある。

データマイニングにより抽出したい知識のタイプをタスクと呼ぶ。データマイニングにおける代表的なタスクと適用される手法の例を表 2-3 にまとめる。

表 2-3 データマイニングにおけるタスクと手法例

タスク	手法
分類（クラス判別）	■ 決定木分析 ■ ニューラルネットワーク
予測	■ ニューラルネットワーク
相関性抽出	■ 相関ルール分析
パターン認識	■ ニューラルネットワーク ■ サポートベクターマシン(SVM)
セグメンテーション （クラスタリング）	■ k-Nearest Neighbor (k-NN)
文章データ処理 （テキストマイニング）	■ 形態素解析

2.3.2 機械学習

（1）機械学習アルゴリズム

計算機がそのアルゴリズムに基づいて、入力に対する正しい出力を導出するためのルール（知識、関数）を、事前に与えられた入力、出力の事例データから獲得することを機械学習(Machine Learning)という（図 2-4)。また、そのアルゴリズムを機械学習アルゴリズムという。機械学習アルゴリズムのことを、データマイニング手法（アルゴリズム）とよぶことがある。

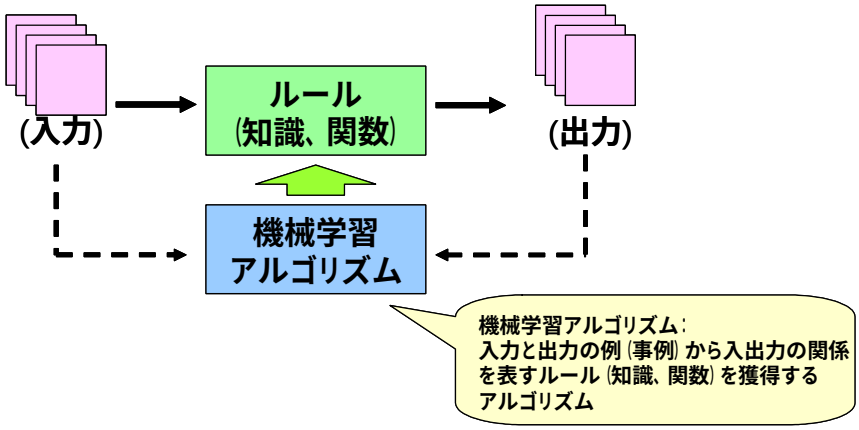

図 2-4　機械学習アルゴリズム

（２）教師あり学習

入力に対する正しい出力（教師信号）が与えられ、正しい出力を出力するようにルール（知識、関数）を調整する学習手法（アルゴリズム）を教師あり学習(Supervised Learning)という。事前に与える入力と正しい出力の集合を学習（訓練）データ（training data)とよぶ。

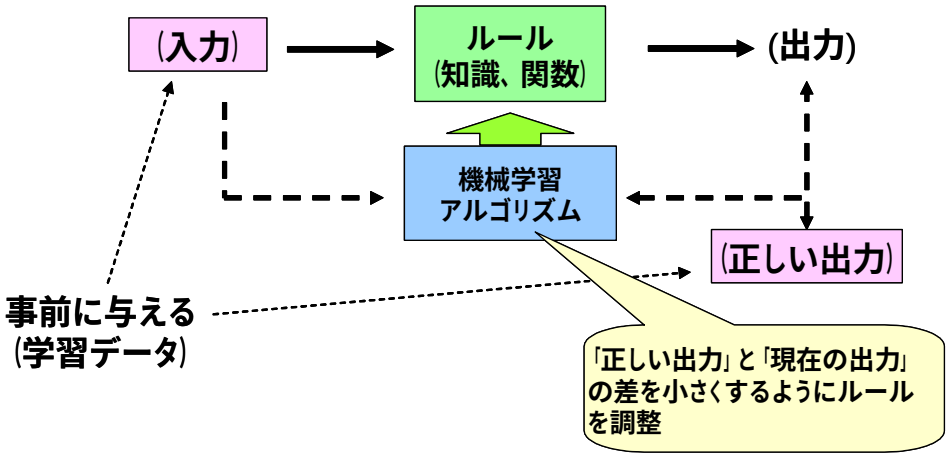

図 2-5　　教師あり学習

（３）教師なし学習

教師なし学習(Unsupervised Learning)とは、正しい出力は与えられず、類似した入力に対して類似した出力をするようにルール（知識、関数）を調整する機械学習アルゴリズムのことである（図 2-6)。類似した事例の集合（クラスタ）を抽出するクラスタリング等は、教師無し学習の代表的な例である。

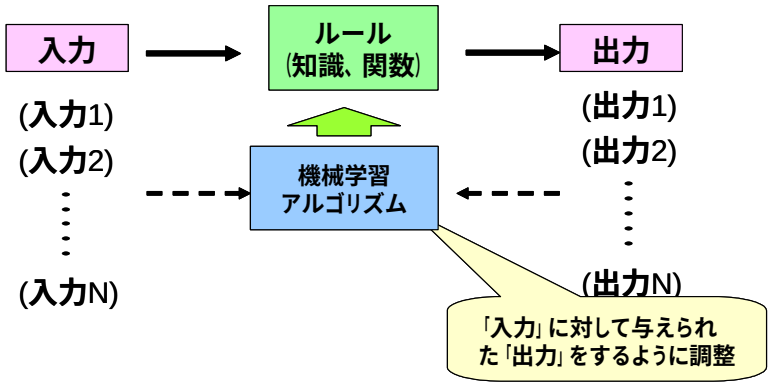

図 2-6　教師なし学習

2.3.3　分類・クラス判別と予測

（1）分類・クラス判別タスク

事例がクラス(class)を判別するための複数の説明属性(attribute)からなっており、与えられた事例について説明属性の値からクラスを判別するためのルールを学習するタスクを、分類・クラス判別という。

分類・クラス判別タスクでは、学習データとして説明属性と正しい出力（正解）に相当するクラスが付与された複数の事例が準備され、この学習データを正しく分類できるようなクラス判定ルールを導出すべく機械学習が行われる（図 2-7）。

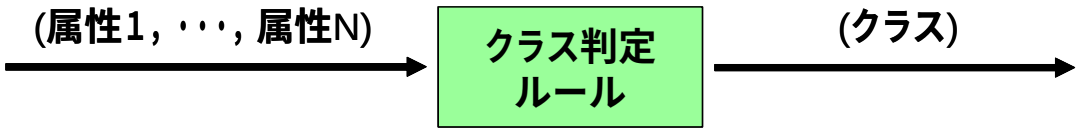

図 2-7　分類・クラス判別タスク

機械学習が行われクラスを判定するためのルールが準備できれば、その後はこのルールに基づき、事例の説明属性さえ与えられれば、クラスを推定することができる。

（2）予測タスク

予測タスクは、分類・クラス判別タスクが事例の説明属性から有限個のクラスを判別するためのルールを学習するのに対して、定量的な出力値（従属変数）を予測するためのルールを学習する。学習は、事前に学習（訓練）データを与えておき、新しい事例の変数に基づき、出力値を予測するための関数を導出する（図 2-8）。例えば、回帰式を求めること

は、この予測タスクを行っていることになる。

図 2-8　予測タスク

（3）手法例：決定木

分類・クラス判別タスクを行うためのデータマイニング手法の代表的なものに決定木アルゴリズムがある。ここでは、カリフォルニア大学アーバイン校の Machine Learning Repository[3]のデータの１つである Hepatitis データに対して決定木アルゴリズムを適用した場合について説明する。

Hepatitis データの概要を表 2-4 にあげる。Hepatitis データは、患者の性別や年齢などの特徴と検査結果が属性として与えられており、その内容から予後の死亡(DIE)と回復(LIVE)を予測するためのルールを抽出するタスクである。

学習データは、表 2-5 のように与えられる。各行が事例、具体的には各患者についてのデータとクラスである予後を表している。決定木アルゴリズムには、このようなデータが与えられ、これら学習データに含まれる事例のクラスが属性値から予測できるようなルールを導出する。

決定木アルゴリズムにより、導出された決定木の例を図 2-9 に示す。決定木はルールが木状に表現されたものであり、最上部を根（ルート）、最下端を葉という。各分岐ノードは、クラスを判別するための属性を表しており、分岐ノードから出ている分岐枝は分岐ノードに採用された属性により判別する際の判別基準値を表している。ルートから葉に至るまでの各分岐ノードと分岐枝により表される判別条件が１つのクラスを判別するルールを表している。

例えば、図 2-9 では”ASCITES（腹水）”がルートノードに表されており、これが”yes（有り）”か”no（無し）”であるかにより、クラスの判別に影響を与えることを表している。”ASCITES”が”no”である分岐枝の下には、分岐ノードに”ALBUMIN”が表れ、この値が”2.8 以下”であるか、”2.8 より大きい”かによりクラスの判別が異なることを表している。”2.8 以下”の場合には、次のノードが葉ノードになり、クラスが”DIE”となっている。このルートノードから分岐ノードを経て葉ノードに至る部分は、「”ASCITES（腹水）”が”no”であり、かつ”ALBUMIN が 2.8 以下”の場合にはクラスが”DIE”である」というルールを表している。

決定木アルゴリズムは、このようなクラス判定のためのルールを、学習データから自動的に学習することができる。

表 2-4　Hepatitis データの属性・属性値とクラス

【クラス】

Class: DIE, LIVE

【属性と属性値】

1. AGE: 10, 20, 30, 40, 50, 60, 70, 80
2. SEX: male, female
3. STEROID: no, yes
4. ANTIVIRALS: no, yes
5. FATIGUE: no, yes
6. MALAISE: no, yes
7. ANOREXIA: no, yes
8. LIVER BIG: no, yes
9. LIVER FIRM: no, yes
10. SPLEEN PALPABLE: no, yes
11. SPIDERS: no, yes
12. ASCITES: no, yes
13. VARICES: no, yes
14. BILIRUBIN: 0.39, 0.80, 1.20, 2.00, 3.00, 4.00
15. ALK PHOSPHATE: 33, 80, 120, 160, 200, 250
16. SGOT: 13, 100, 200, 300, 400, 500,
17. ALBUMIN: 2.1, 3.0, 3.8, 4.5, 5.0, 6.0
18. PROTIME: 10, 20, 30, 40, 50, 60, 70, 80, 90
19. HISTOLOGY: no, yes

表 2-5　Hepatitis データの具体例（抜粋）

CLASS	属性																		
	AGE	SEX	STEROID	ANTIVIRALS	FATIGUE	MALAISE	ANOREXIA	LIVER BIG	LIVER FIRM	SPLEEN PALPABLE	SPIDERS	ASCITES	VARICES	BILIRUBIN	ALKPHOSPHATE	SGOT	ALBUMIN	PROTIME	HISTOLOGY
LIVE	30	F	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	1	85	18	4	?	No
LIVE	50	M	No	Yes	No	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	0.9	135	42	3.5	?	No
LIVE	78	M	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0.7	96	32	4	?	No
LIVE	31	M	?	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0.7	46	52	4	80	No
LIVE	34	M	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	1	?	200	4	?	No
LIVE	34	M	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0.9	95	28	4	75	No
DIE	51	M	No	Yes	No	Yes	No	Yes	Yes	No	No	Yes	Yes	?	?	?	?	?	No
LIVE	23	M	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	1	?	?	?	?	No
LIVE	39	M	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	0.7	?	48	4.4	?	No
LIVE	30	M	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	1	?	120	3.9	?	No
LIVE	39	M	No	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	1.3	78	30	4.4	85	No
LIVE	32	M	Yes	No	No	Yes	Yes	Yes	No	Yes	No	Yes	Yes	1	59	249	3.7	54	No
LIVE	41	M	Yes	No	No	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	0.9	81	60	3.9	52	No
LIVE	30	M	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	2.2	57	144	4.9	78	No
LIVE	47	M	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	?	?	60	?	?	No
LIVE	38	M	No	Yes	No	No	No	Yes	Yes	Yes	Yes	No	Yes	2	72	89	2.9	46	No
LIVE	66	M	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	1.2	102	53	4.3	?	No
LIVE	40	M	No	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	0.6	62	166	4	63	No
LIVE	38	M	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0.7	53	42	4.1	85	Yes
LIVE	38	M	No	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	0.7	70	28	4.2	62	No
LIVE	22	F	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0.9	48	20	4.2	64	No
LIVE	27	M	Yes	Yes	No	No	No	No	No	No	No	Yes	Yes	1.2	133	98	4.1	39	No

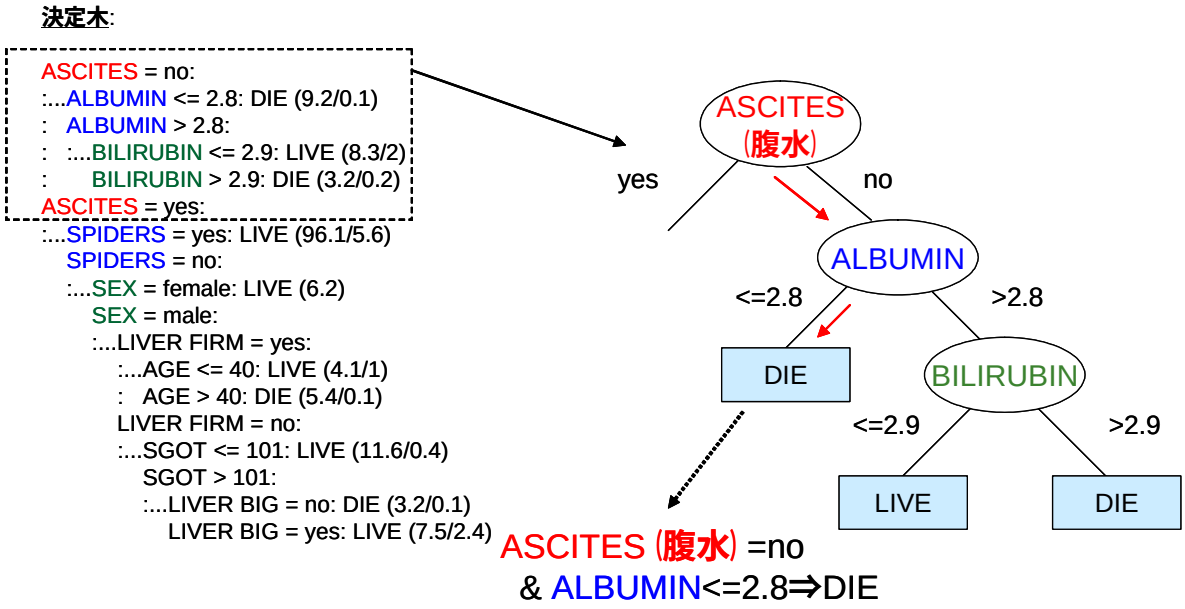

図 2-9　Hepatitis データから導出された決定木

2.3.4 相関ルール抽出

事例の中で頻繁に共起しており、相関があると考えられる項目を相関ルールとして抽出するタスクを相関ルール抽出タスクという。

相関ルール抽出タスクの代表的なものとして、バスケット分析と呼ばれるものがある。バスケット分析とは、顧客が同時に購入する、すなわち同じバスケット（買い物かご）に入れた商品を分析するところから、そのように呼ばれている。バスケット分析では、事例がある客の買い物カゴに相当し、事例の内容はその顧客が同時に購入した商品（牛乳、食パン、ジャムなど）である。最近では、多くの小売店に POS システムが導入されているためこのようなデータが容易に収集できる。

相関は、相関ルールA⇒Bという形で抽出される。ここで、A、Bは商品群であるアイテムの集合である。この相関ルールは、「顧客が商品群Aを購入したときには、他にも商品群Bも購入する（場合が多い）」ということを表している。例えば、相関ルール「ビール⇒紙おむつ」という相関ルールが見つかった場合には、「ビールを購入する顧客は、同時に紙おむつも購入する」ということを表している。このようなルールが見つかれば、商品の配置などに活用することができる。

一般に POS データのような大量のデータから、このような相関ルールを効率よく発見することは困難であったが、Apriori アルゴリズム[9]が開発され、大量のデータから短い処理時間で相関ルールを抽出することが可能になった。Apriori アルゴリズムでは、相関ルールの性能を表す指標として次のようなものが用いられる。

支持度：$\frac{n(A \cup B)}{N}$、

確信度 ：$\frac{n(A \cup B)}{n(A)}$、

ここで N ：事例数、$n(A \cup B)$：A,B ともに含む事例数、$n(A)$：A を含む事例数

支持度は、当該相関ルールが当てはまる事例の割合を、確信度は当該相関ルールの条件部があてはまった場合に、結論部が成立する割合を表している。

相関ルール抽出は、医薬品副作用情報の分析でも、例えば併用薬シグナルの検出等で利用が可能であると考えられる[15]。

2.3.5 パターン認識

（1）パターン認識タスク

パターン認識とは、与えられた入力値（パターン）について有限個のパターンの中から最も近いパターンを出力するタスクである。予め定められたパターンの何れに近いかを判

定するタスクはパターン認識と呼ばれ、パターンそのものを導出するタスクはクラスタリング、またはパターン検出と呼ばれる。例えば郵便番号の自動認識などに用いられている文字認識などは、パターン認識の代表的な例である。パターン認識では、あらかじめ与えられた入力に対して出力すべきパターンが与えられており、これに基づいて学習を行う。すなわち、出力すべきパターン（正解）が与えられる教師有り学習である。一方、パターン検出は、複数ｎ入力のうち類似したものを１つの出力パターンとして同定する教師なし学習である。

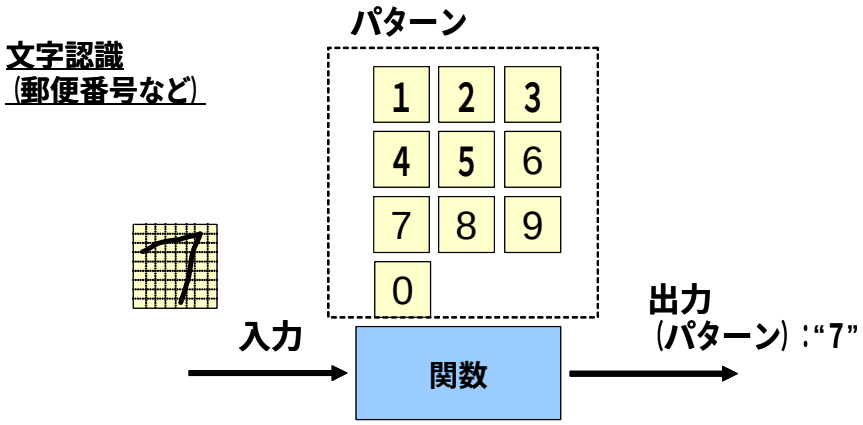

図 2-10　パターン認識タスク

（２）手法例：ニューラルネットワーク

パターン認識や分類、予測等の幅広いタスクに用いられるデータマイニング手法にニューラルネットワーク[12]がある（図 2-11）。ニューラルネットワークは、神経回路網とも呼ばれる。ニューラルネットワークは、ニューロンとよばれる神経細胞をモデル化した一種の関数を組み合わせることにより、複雑な入出力関係をモデル化することができる。ニューロンは、相互に結合されており、入力が一定の閾値を超えると発火する（出力値を出す）。結合は重みとして表現されている。ニューラルネットワークでは、この重みを変化させることにより、モデル化を行う。

ニューラルネットワークは、手法について知識がなくとも扱えること、複雑な入出力関係をモデル化できることからよく利用される。

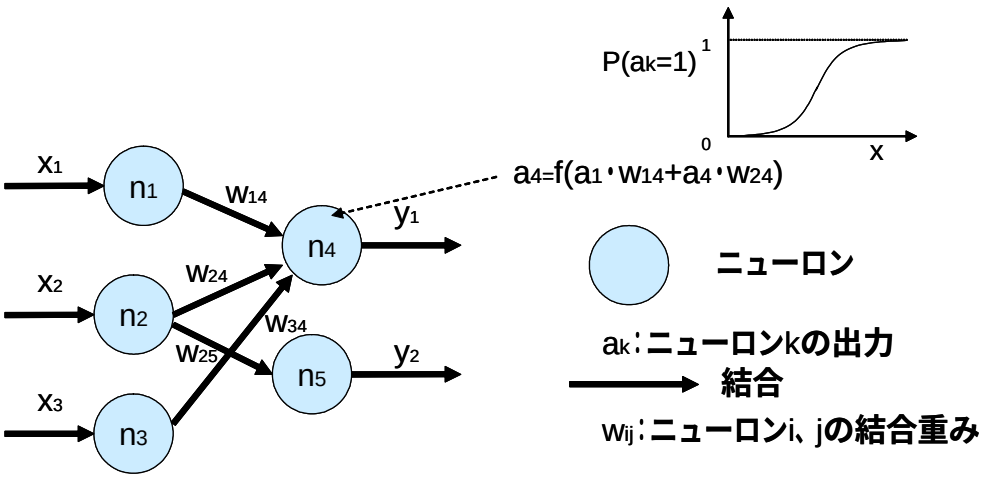

図 2-11　ニューラルネットワーク

ニューラルネットワークには、入力から出力までニューロンが階層をなしている階層型とニューロンが相互に結合している相互結合型がある（図 2-12）。階層型のニューラルネットワークは、主に分類や予測のタスクに用いられるものであり、教師あり学習を行う。一方、相互結合型のニューラルネットワークは、主にパターン認識のタスクに用いられ、教師なし学習を行う。

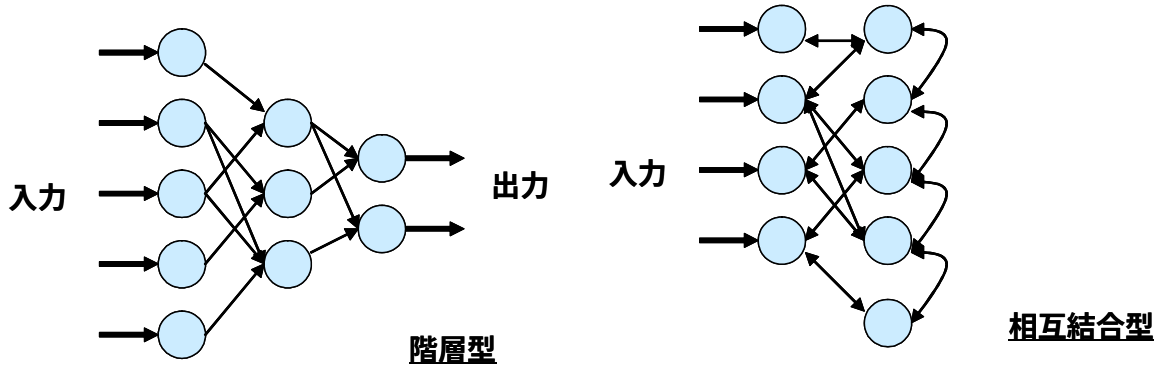

図 2-12　ニューラルネットワークのタイプ

2.3.6　テキストマイニング

一般的なデータマイニングでは、データが数値であったり記号であったり、既に一般的なデータマイニング手法に適用可能な形であるか、あるいは適用可能に近い形である場合が多い。しかし、一般的に分析者が扱うデータはこのような形式でなく、文章などテキスト状態である場合が多い。データが文章そのままでは各種の統計的手法／データマイニング手法に適用しにくい。そこでデータマイニング手法に適用するためには、文章データを数値や記号などによるデータに変換する必要がある。文章データを対象としてデータマイニングを行うことをテキストマイニング[17][13]という。

このような文章データを変換する方法として、形態素解析により、文章を単語に分解することが行われる。例えば、奈良先端大の松本裕治教授が開発した茶筌[2]というソフトを用いると、図 2-13 のように文章を単語に分解することができる。このように単語に分解できれば、文章データを文章中に表れている単語の種類、あるいは頻度等をデータとして、一般的なデータマイニングツールに適用可能なデータ形式とすることができる。

地中にメタンガスが蓄積し、これに引火したのが火災の発生原因であると推定されている。

表層語	基本形	品詞	活用形	
地中	地中	名詞-一般		
に	に	助詞-格助詞-一般		
メタンガス	メタンガス	名詞-一般		
が	が	助詞-格助詞-一般		
蓄積	蓄積	名詞-サ変接続		
し	する	動詞-自立	サ変・スル	連用形
、	、	記号-読点		
これ	これ	名詞-代名詞-一般		
に	に	助詞-格助詞-一般		
引火	引火	名詞-サ変接続		
し	する	動詞-自立	サ変・スル	連用形
た	た	助動詞	特殊・タ	基本形
の	の	名詞-非自立-一般		
が	が	助詞-格助詞-一般		
火災	火災	名詞-一般		
の	の	助詞-連体化		
発生	発生	名詞-サ変接続		
原因	原因	名詞-一般		
で	だ	助動詞	特殊・ダ	連用形
ある	ある	助動詞	五段・ラ行アル	基本形
と	と	助詞-格助詞-引用		
推定	推定	名詞-サ変接続		
さ	する	動詞-自立	サ変・スル	未然レル接続
れ	れる	動詞-接尾	一段	連用形
て	て	助詞-接続助詞		
いる	いる	動詞-非自立	一段	基本形
。	。	記号-句点		

図 2-13　形態素解析の入力と出力例

形態素解析の利用例として、アンケートの自由記述欄の分析や類似文章の検索などがあ

[2] 茶筌は、http://chasen.naist.jp/hiki/ChaSen/で配布されている(平成 17 年 3 月現在)。

る。アンケートの自由記述欄の分析では、例えば商品に関するアンケートであれば、「安い」など商品イメージに関わる形容詞等を抽出し、アンケート中の他の回答項目といっしょに分析を行うというようなことが行われる。また、文書の類似検索では各文書の内容を文書に含まれる各単語の出現頻度を表すベクトルで表現し、文書間の類似度を、各々の出現頻度ベクトルの内積により評価している。

2.3.7 アンサンブル学習

従来の一般的な手法よりも精度のよいルールを学習できる手法として、アンサンブル学習[14]が近年注目を集めており、利用が進められている。

従来の一般的な機械学習では学習データを用いて、データマイニング手法（機械学習アルゴリズム）により、1 つのルール（クラス判定ルール、関数）を学習し、このルールに基づいて分類や予測を行う（図 2-14）。

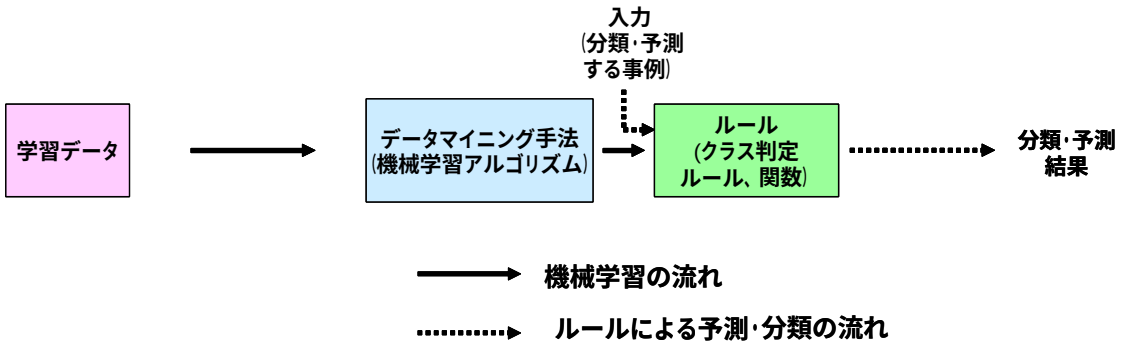

図 2-14 従来の一般的な機械学習手法

これに対して、アンサンブル学習データは以下のような方法で複数（N 個）のルールを学習する。そして、分類・予測を行なう場合には、各ルールの分類・予測結果を多数決や平均により統合し、最終的な分類・予測結果を得る（図 2-15）。手法の詳細は次のようになる。まず元の学習データからサンプリングを行い、複数の学習データ（サンプリング学習データ）を準備する。このサンプリング学習データは、元の学習データから個々に独立にサンプリングされたものであるので、元の学習データも含めて互いに含まれる事例が多少異なる。次に各サンプリング学習データを用いて、データマイニング手法によりルールを学習する。ルールはサンプリング学習データの数だけ生成される。ここで、データマイニング手法が同じであってもサンプリング学習データに含まれる内容が異なるため、生成されるルールも多少異なったものとなる。実際に分類や予測を行う場合には、まず、これら各ルールによる分類、予測を行ない、それらの結果の多数決や平均をとることにより決定する。

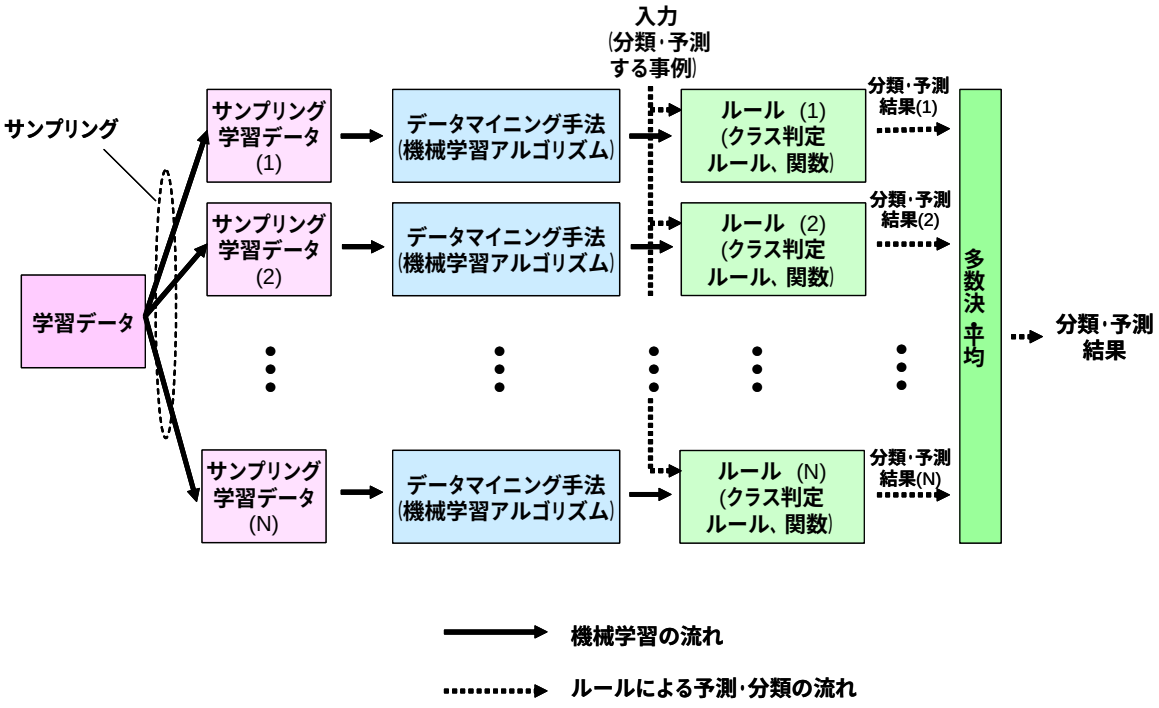

図 2-15 アンサンブル学習

同じ学習データを用いて、同じデータマイニング手法で学習した場合、同じルールが生成される。アンサンブル学習では、元の学習データから独立してサンプリングを行った複数の含まれる事例がわずかに互いに異なるサンプリング学習データを生成し、それぞれを用いて学習を行うことにより、複数の異なるルールを得る。

アンサンブル学習の代表的なものに、Boosting[10]と Bagging[6]がある。Boosting では, 学習データの事例ごとに付与されたサンプリング確率が付与されており、初期は同じ確率となっている。各サイクルでは、サンプリング確率に基づいてサンプリングしたサンプリング学習データを用いて学習を行うことによりルールを導出する。次に、学習に用いたサンプリング学習データを用いて、導出したルールで分類・予測した結果をもとにサンプリング確率を更新する。このとき、誤って分類・予測された事例については、サンプリング確率が他事例に対して相対的に高く更新されるので、次サイクル以後のサンプリングでは、その事例がサンプリングされる確率が高くなる。このサンプリングから、サンプリング確率の更新までのサイクルを、収束条件が充たされるまで繰り返すことにより、複数のルールが生成される（図 2-16)。実際に分類・予測を行う場合には、これら複数のルールの予測・分類結果に対して、ルール生成時の導出ルールの予測精度を反映した重み付け平均をとることによって、全体の予測・分類結果とする。Boosting を適用することにより，全学

習データを用いて１つのルールを生成する従来の一般的な手法に比べて分類精度・予測が向上することが報告されている。ただし、複数のルールを順々に導出することが必要であるために全ルールを導出するまでに、従来の一般的な手法に比べて多くの処理時間を要するという問題点がある.

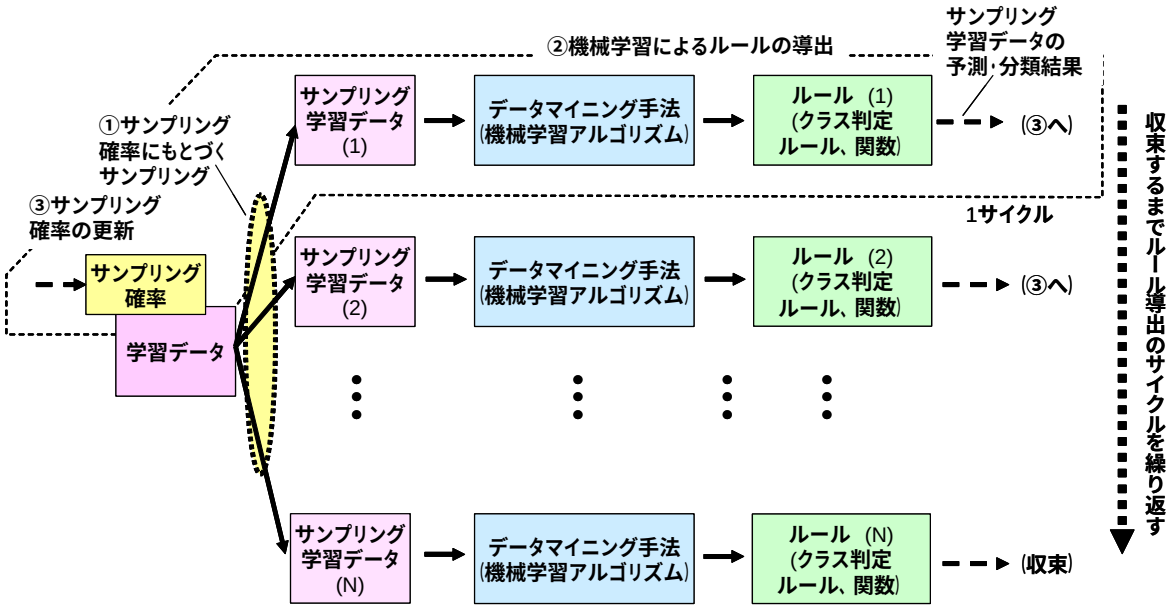

図 2-16　Boosting によるルールの導出

一方、Bagging とは元学習データに対してブートストラップサンプリングと呼ばれる手法でサンプリングを行うことにより複数のサンプリング学習データを準備する。各サンプリング学習データを生成するためのサンプリングは互いに独立して行なうことができる。次に各サンプリング学習データにより学習を行い、複数のルールを導出する（図 2-17)。そして、これら複数のルールの各々が分類・予測した結果の多数決や平均をとることにより、最終的な分類・予測結果とするものである。

Bagging は Boosting に比べると導出される分類･予測の精度の面で劣ることが多いが、バイアスを含むデータでも比較的安定して１つのルールを生成、分類・予測に用いる従来の一般的な手法よりも精度を改善することができる。さらには Boosting と異なり、各ルールの導出プロセスが互いに独立したものであるため、並列的に導出することも可能である。このため、並列処理を行えば１つのルールを導出するのと処理時間が変わらないという特長をもつ。

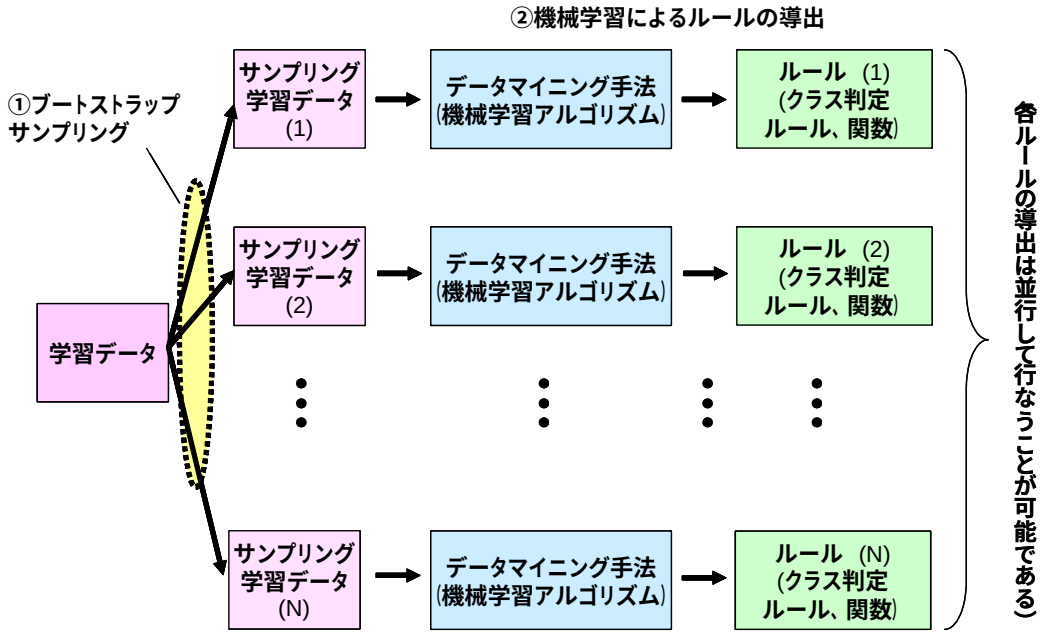

図 2-17　Bagging によるルールの導出

アンサンブル学習は、既に手法が確立されており[5]、何れのデータマイニング手法にも適用可能で、かつ比較的実装が容易な手法であることから、安全分析業務に導入するデータマイニングにおいても、実装を検討すべき手法である。

2.3.8　まとめ

以上のようにデータマイニング手法では、データマイニングにより抽出したい知識（タスク）ごとに手法が開発されている。

医薬品副作用情報の分析においても、例えば以下のような手法が適用可能であると考えられる。

① 相関ルール抽出：併用薬をはじめ副作用と因果関係があると考えられる複数の要因を同時に扱う場合に、相関ルール抽出手法が利用できると考えられる。相関ルール抽出により、因果関係に関するラフな分析を行うことにより仮説を立てた後、統計的手法を用いた詳細な分析を行なうのが有効である。

② ニューラルネットワーク：医薬品と副作用の因果関係の強さの指標値の予測や、医薬品群、副作用群など類似した群（クラスタ）の抽出などにニューラルネットワー

クの適用が有効であると考えられる。例えば、WHO では、ニューラルネットワークを用いたシグナル検出や ADR 群の抽出に関する手法の開発を行っている。

③ アンサンブル学習：アンサンブル学習は、多くのデータマイニング手法のみでなく、既存のシグナル検出手法で用いられているような統計的手法に適用可能な手法であり、かつ精度を高めることができる手法である。シグナル検出手法に適用することにより、シグナルの精度を高めることができると考えられる。

2.4 異業種成功事例の調査・分析

医薬品副作用情報の分析へのデータマイニングについて検討するための基礎資料の１つとして、異業種においてデータマイニングを導入し成功している事例を調査した結果をまとめる。ここでは特に、市販された製品の不具合に関するデータに対してデータマイニングを適用することにより、原因の発見や製品の改善に活用された２事例をとりあげた。

2.4.1 自動車製造業の事例

米国運輸省高速道路安全局（The National Highway Traffic Safety Administration: NHTSA)が、ファイアーストーン社のタイヤのリコール問題を契機として、自動車およびタイヤやチャイルドシートなど自動車関連部品について、不具合に関して収集された多くの項目を定期的に報告する仕組みを法制化した。これが Transportation Recall Enhancement, Accountability, and Documentation (TREAD) Act、いわゆる TREAD Act である。

TREAD Act 自体は、自動車および自動車関連部品（チャイルドシート、タイヤなど）に関する、事故情報、クレーム情報、関連製品の米国外での事故情報について、NHTSA に定期的に報告するように義務化しただけのものであり、それら情報の分析は義務づけられていない。一方、これら収集した情報を自社の製品品質の改良に生かす取り組みが自動車業界自ら行われ始めている。

例えば、GM やフォードにおいて不具合データの分析に取組み始めていることが報告されている(日経産業新聞 2004 年 11 月 18 日 1 面より)。米国においても、全国のディーラーから、不具合に関する報告が電子メールなど文章で大量に届いていたが、文章であるために分析に必要なデータの抽出が困難であるという問題があった。例えば「雨になるとエンジンがかからない」という報告があがった場合、分析を行うためには「条件＝雨」、「部品名＝エンジン」、「不具合＝かからない」と抽出する必要がある。これに対して、文章データにテキストマイニングを適用することにより、文章から分析に必要な情報を抽出することに成功している。さらに、データマイニング手法を用いて、特定の「条件」－「部品」で発生しやすい「不具合」を相関ルールとして抽出することにも成功した。

本事例は、文章データに対するテキストマイニングの適用事例の１つであり、文章による記述が多い症例票に基づいて分析を行う症例分析の部分などへのデータマイニングの適用について検討する際に参考になると考えられる。

2.4.2 家電製造業の事例

市販後に不具合が発生したテレビなどの電気製品について早急に問題を解決するためには、市場で発生した不具合の状況を修理サービスの結果から把握し、製造、設計部門にフィードバックすることが重要である。

データマイニングが導入される以前から、市場で発生した不具合の内容が記された修理

伝票は工場に回されていたし、重大なものについては修理を行ったサービスマンから工場に直接報告されていた。さらに、ペーパーレス化が進み修理伝票はデータベースに格納され、共有することができるようになっていた。このように情報の共有が進められたにもかかわらず、品質に関わる問題が発見できなかったり、発見が遅れたりという問題点が指摘されていた。堀は、その理由について次のように分析している[19]。

① サービスマンは製品の動作原理を定性的に理解しており、その理解に基づき不具合発生要因の仮説を立てている。これにより、部品故障が設計不良などの重要な原因により生じたかどうかを推定できる。一方、人間が仮説を与えてデータベースを検索するときには、該当する修理件数を数え上げるだけである。
② サービスマンは、設計変更・一般故障率など多くの背景知識を持っている。したがって、彼らは故障の真の原因（部品不良か設計不良かなど）を推察し、その重大さを的確に判断できる。計算機は背景知識をもっていないので、発生頻度でしか重要性の判断ができない。

堀らは、このような既存のデータベースによる品質に関わる問題の発見の課題を解決するために、市場品質監視システムを開発、導入した。市場品質監視システムは、データマイニングの一手法であるバスケット分析（相関ルール分析）を採用しており、製品の不具合に関わる、設計段階、製造段階、検査結果など全ての要因と不具合の相関関係（因果関係）を分析し、相関ルールとして抽出するものである。バスケット分析を適用することの利点として、定量的基準に基づき市場の品質動向をもれなく監視できること、不具合に関わる要因群の抽出が可能であることが指摘されている。

2.4.3 まとめ

異業種におけるデータマイニングの導入事例のうち代表的なものとして、医薬品副作用情報の分析と市場に製品が出たあとの不具合情報の分析という点で共通点がある２つの事例を抽出し、調査を行った。

自動車製造業の事例からは、データマイニングを有効なものとするためにテキスト情報を扱うことが重要であることが確認された。多くの場合、重要な情報はコーディングされたデータではなく、テキストで記述された情報に含まれる。医薬品副作用情報についても、詳細な情報が含まれる症例報告票は、特に重要な部分はテキストで記述されている場合が多い。将来的に導入するデータマイニング手法については、テキストを扱えることが重要な要件になると考えられる。

家電製造業の事例からは、分析の担当者は分析時に背景知識を利用しており、これが計算機よりも高度な分析を可能にしている点であるということが指摘されている。医薬品副作用情報の分析においても、ラインリストに含まれる情報以外にも、これまでの経験や得

られている情報に基づく知識や薬理作用などに関する知識を用いて、担当者は分析を行なっていることが業務分析から明らかになっている。データマイニングに、このような分析担当者が持つ背景知識をいかに利用していくかがデータマイニング導入を成功させるための鍵となる。

2.5 諸外国規制当局等の当該業務の現状調査

2.5.1 主な諸外国規制当局における適用手法

欧米の規制当局では、すでに副作用報告の症例の情報を有効に活用するためのシグナル検出手法が開発され、実際に使用されている[22]。

1998 年に Bate らによって WHO の Uppsala モニタリングセンター（UMC）で用いられている BCPNN が発表された。1999 年には FDA の DuMouchel が GPS Program を、2001 年には英国 MCA（現在の MHRA）の Evans らが PRR の方法を発表した。さらに DuMouchel らは 2001 年に GPS Program を拡張した MGPS を発表し、現在では FDA では MGPS を使用している。また、オランダの Lareb では最もシンプルな方法である ROR を適用しており[4]、オーストラリアの TGA では PROFILE と呼ばれるシグナル検出手法の適用を検討している[8]。これらの諸外国の規制当局で適用または検討されている手法をまとめたものを表 2-6 に示す。

表 2-6 主な諸外国規制当局において適用または検討されている手法

規制機関	国	発表年	シグナル検出手法
WHO	―	1998	BCPNN：Bayesian Confidence Propagation Neural Network
FDA	米国	1999	GPS：Gamma-Poisson Shrinker Program
		2001	MGPS：Multi-Item Gamma-Poisson Shrinker Program
MHRA	英国	2001	PRR：Proportional Reporting Ratios
Lareb	オランダ	―	ROR：Reporting Odds Ratio
TGA	オーストラリア	―	PROFILE：Probability Filtering Method

2.5.2 シグナル検出手法

シグナル検出の元となるデータは、行に医薬品を列に副作用を取り、その報告件数を度数とした表 2-7 に示す度数表である[18]。

また、表 2-7 において特定の医薬品と副作用に注目すると、注目する医薬品とその他の医薬品、注目する副作用とその他の副作用からなる表 2-8 に示す 2×2 分割表ができる。さらに、このそれぞれのセルを確率で表した場合には表 2-9 に示す表が作成できる。

表 2-7　シグナル検出の元となるデータ

	副作用 1	副作用 2	…	副作用 p	合計
医薬品 1	n_{11}	n_{12}	…	n_{1p}	n_{1+}
医薬品 2	n_{21}	n_{22}	…	n_{2p}	n_{2+}
⋮	⋮	⋮	…	⋮	⋮
医薬品 m	n_{m1}	n_{m2}	…	n_{mp}	n_{m+}
合計	n_{+1}	n_{+2}	…	n_{+p}	n_{++}

表 2-8　2×2 分割表のセル度数

	注目する副作用	その他の副作用	合計
注目する医薬品	n_{11}	n_{12}	n_{1+}
その他の医薬品	n_{21}	n_{22}	n_{2+}
合計	n_{+1}	n_{+2}	n_{++}

表 2-9　2×2 分割表の確率

	注目する副作用	その他の副作用	合計
注目する医薬品	P_{11}	P_{12}	P_{1+}
その他の医薬品	P_{21}	P_{22}	P_{2+}
合計	P_{+1}	P_{+2}	P_{++}（=1）

以下では、表 2-8 および表 2-9 の表中に示した記号を用いて、表 2-6 に示した手法を中心にシグナル検出の手法について述べる。

（1）ROR[4][22]

この手法は、通常のオッズ比を用いたものであり、最もシンプルなシグナル検出手法である。オランダの Lareb では ROR を副作用自発報告システム（SRS）に適用している。

表 2-8 に基づいて期待値は次のように計算される。

$$ROR = \frac{(n_{11}/n_{21})}{(n_{12}/n_{22})} = \frac{n_{11}n_{22}}{n_{12}n_{21}} \qquad (2\text{-}1)$$

また、95%信頼区間は次式で与えられる。

$$95\%CI = e^{\ln(ROR)\pm 1.96\sqrt{\left(\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}\right)}} \qquad (2\text{-}2)$$

シグナル判定基準は、95%信頼区間の下限が 1 より大きい場合に、「シグナルあり」と判断している。

Lareb ではこの ROR を次の 3 つの目的のために使用している。

1. 従来のケースごとにチェックする方法を補完する。
2. 副作用報告データベース内で定期的に発生する医薬品と副作用の組の関係の不均衡を発見する。
3. 併用薬などの、より複雑な関係を含めた検討を行う。

（2）PRR[22]

この手法は、現在 MHRA で用いられている手法である。PRR は疫学における特定死因死亡率（Proportional Mortality Ratio）に類似している。PRR では、表 2-8 に基づいて、評価がなされ、医薬品ごとの報告割合の比を次に示す PRR 期待値として与えるというものである。

$$PRR = \frac{n_{11}/n_{1+}}{n_{21}/n_{2+}} = \frac{P_{11}/P_{1+}}{P_{21}/P_{2+}} \qquad (2\text{-}3)$$

シグナル検出基準として、様々な値を用いた研究が報告されているが、MHRA では次の 3 つの条件を満たす場合を基準としている。

$$\begin{aligned} &(1)\ PRR \geq 2 \\ &(2)\ \chi^2 = \frac{n_{++}\left(\left|n_{11}n_{22} - n_{12}n_{21}\right| - n_{++}/2\right)^2}{n_{1+}n_{2+}n_{+1}n_{+2}} \geq 4 \\ &(3)\ n_{11} \geq 3 \end{aligned} \qquad (2\text{-}4)$$

(1)では、注目した医薬品と副作用の組の報告割合が、その医薬品と他の副作用との報告割合の 2 倍以上であることを示している。また、(2)における χ^2 とは、期待度数と観測度数のズレを表す数値であり[20]、この数値が 4 以上の場合は偶然起きたとは考えにくいと判断し、シグナルと判断している。また、(3)は注目した医薬品と副作用の組の報告数が 3 件以上の場合である。他の基準としては、$PRR > 3, \chi^2 > 5$ や $PRR > 2, \chi^2 > 4$ (すべての副作用報告数>2 の場合)というものがある。3 つ目に示した基準値を英国の ADROIT データベ

ースに対して適用した場合、約 60%は検出済みの副作用、約 15%は間違ったシグナル、残りの約 25 %は詳細な分析が必要と判断される新たなシグナルであったという報告もなされている[7]。

（3）BCPNN[22]

この手法は WHO で採用されている。BCPNN では注目する医薬品に関する報告がなされる確率 P_{i+}、および注目する副作用に関する報告がなされる確率 P_{+j} から、注目する医薬品と副作用の組が報告される確率を予想し、この計算された確率と実際の報告数から得られる確率 P_{ij} を比較することで、シグナルを検出する。

医薬品と副作用の組に対して、評価のために IC_{ij} を次のように定義する。

$$IC_{ij} = \log_2\left(P_{ij} / P_{i+}P_{+j}\right) \qquad (2\text{-}5)$$

分母は P_{i+} と P_{+j} の積であることから、注目する医薬品と副作用の組が報告される確率の予想値とみることができる。分子は実際の報告による確率であり、観測値と推定値の比を取っていることがわかる。

IC_{ij} を計算するためには、P_{i+}, P_{+j} の確率、および P_{ij} の確率の分布を仮定する必要があるが、この手法では標本サイズの少ないセルでの 1 例の影響を小さくするためにベイズ流のアプローチを用いている。具体的には、$P_{i+}\left(i=1,2\right)$ と $P_{+j}\left(j=1,2\right)$ に関する分布にはベータ分布を用い、$P_{ij}\left((i,j)=(1,1),(1,2),(2,1),(2,2)\right)$ に関する同時分布としては、ディリクレ分布（ベータ分布を多変量に拡張したものとみなすことができる）を用いている。それぞれの分布を決定するためには、以下のパラメータを決定する必要がある。

- P_{i+}に関するベータ分布のパラメータ：$\alpha_1, \alpha_2\left(\alpha_1+\alpha_2=\alpha\right)$
- P_{+j} に関するベータ分布のパラメータ：$\beta_1, \beta_2\left(\beta_1+\beta_2=\beta\right)$
- P_{ij} に関するディリクレ分布のハイパラメータ：$\gamma_{11,}\gamma_{12,}\gamma_{21,}\gamma_{22}$

注目する医薬品と注目する副作用の組の評価を行うためには IC_{11} を求めればよい。この計算のために必要となる P_{11} の事後分布の期待値と分散に関しては事前分布に用いた 4 つのパラメータではなく、2 つのパラメータ（γ_{11} と γ）で表すことができる。これらの事前分布としては無情報事前分布を用いており、$\alpha_1=\beta_1=1, \alpha=\beta=2,\ \gamma_{11}=\gamma_{12}=\gamma_{21}=\gamma_{22}=1$ としている。

通常使われるオッズ比を用いた場合には、オッズの推定において 4 つのセルのうち 1 つでも 0 のセルがある場合、オッズは 0 または∞になり、周辺和に 0 がある場合はオッズ比が定義できない。また、標本サイズが少ない、あるセルにおいてオッズ比を推定する場合、標本の少ないセルでの 1 例の影響が非常に大きくなるため、推定が不安定になるなどの問

題があるが、前述したようなベイズ流のアプローチを行うことで、安定したパラメータの推定を可能としている。

各分布のパラメータを決定することで、事後分布における IC_{11} の期待値とその分散は次のように計算される。

$$E(IC_{11}) = \log_2 \frac{(n_{11}+\gamma_{11})(n_{++}+\alpha)(n_{++}+\beta)}{(n_{++}+\gamma)(n_{1+}+\alpha_1)(n_{+1}+\beta_1)} \tag{2-6}$$

$$V(IC_{11}) = \left(\frac{1}{\log_2}\right)^2 \left[\frac{n_{++}-n_{11}+\gamma-\gamma_{11}}{(n_{11}+\gamma_{11})(1+n_{++}+\gamma)} + \frac{n_{++}-n_{+1}+\alpha-\alpha_1}{(n_{1+}+\alpha_1)(1+n_{++}+\alpha)} + \frac{n_{++}-n_{+1}+\beta-\beta_1}{(n_{+1}+\beta_1)(1+n_{++}+\beta)}\right] \tag{2-7}$$

ここで、

$$\gamma = \gamma_{11}\frac{(n_{++}+\alpha)(n_{++}+\beta)}{(n_{1+}+\alpha_1)(n_{+1}+\beta_1)},\ \gamma_{11}=1, \alpha_1=\beta_1=1,\ \alpha=\beta=2 \tag{2-8}$$

である。

以上から、データ取得後の事後分布を用いて確率区間（信頼区間に相当）を求め、そこから外れたものをシグナルとして検出する。具体的なシグナル検出基準は（おおよその）95%信頼区間の下限が 0 より大きい場合としており、次の場合にシグナルとして検出している。

$$E\left(IC_{11} - 2\sqrt{V(IC_{11})}\right) > 0 \tag{2-9}$$

（4）GPS[1][22]

この手法は、FDA で以前用いられていた方法である。GPS では背景の情報を加えた層別による評価を行っている。背景情報としては性（男、女、不明）、年齢、報告された日（5 年毎、あるいは 1 年毎）を用いており、これを s 層とした場合、期待値 E_{ij} を次のように与えている。

$$E_{ij} = \sum_s E_{ijs} = \sum_s \frac{n_{i+s}n_{+js}}{n_{++s}} \tag{2-10}$$

ただし、E_{ijs} は s 層における期待値、n_{++s} は s 層の報告件数を表す。

シグナルを検出する際の評価指標としては相対リスク $RR_{ij}=n_{ij}/E_{ij}$（実際の報告数と期待値の比）がある。この相対リスクに対して $\lambda_{ij}=\mu_{ij}/E_{ij}$ を定義し、実際のシグナル検出の評価にはこの λ_{ij} を用いている。ただし、μ_{ij} は観測値が n_{ij} となるポアソン分布の平均である。この λ_{ij} の事前分布を 2 つのガンマ分布の混合分布からの観測値であると仮定し、5 つのハイパラメータ $\theta=(\alpha_1,\beta_1,\alpha_2,\beta_2,P)$ を推定する。α_1,β_1 および α_2,β_2 は 2 つのガンマ分布におけるパラメータ、P はその混合率である。パラメータを決定するために、尤度 $L(\theta)$ を用いる。尤度 $L(\theta)$ は母数空間での様々な θ の値に対するもっともらしさを表す関数とみなすことができる[20]ため、この値を最大とする θ を探し、これを推定値とすればよい。尤度は次のように与えられる。

$$L(\theta)=\Pi_{ij}\left\{Pf(n_{ij};\alpha_1,\beta_1,E_{ij})+(1-P)f(n_{ij};\alpha_2,\beta_2,E_{ij})\right\} \qquad (2\text{-}11)$$

$$f(n;\alpha,\beta,E)=(1+\beta/E)^{-n}(1+E/\beta)^{-\alpha}\times\Gamma(\alpha+n)/\Gamma(\alpha)n!$$

（$f(\ )$ は負の二項分布を表す）

反復法によって、この尤度 $L(\theta)$ を最大にするような 5 つのパラメータ $\theta=(\alpha_1,\beta_1,\alpha_2,\beta_2,P)$ を求める。次に推定された θ の値を用いて、λ の事後分布 Q_n を求める。

$$Q_n=Pf(n;\alpha_1,\beta_1,E)/\left[Pf(n;\alpha_1,\beta_1,E)+(1-P)f(\lambda;\alpha_2,\beta_2,E)\right] \qquad (2\text{-}12)$$

以上により求められる λ の事後分布の累積分布関数を用いて 5%点（EB05）を算出し、EB05 をシグナル判定値とする。

なお、FDA では $EB05\geq 2$ をシグナル判定基準としている。

また、λ の事後分布から以下に定義される λ の幾何平均（EBGM）が求められる。

$$EBGM_{ij}=\exp\left\{E\left[\log(\lambda_{ij})\,\middle|\,n_{ij},\theta\right]\right\} \qquad (2\text{-}13)$$

$$E\left[\lambda\,|\,N=n,\theta\right]=Q_n(\alpha_1+n)/(\beta_1+E)+(1-Q_n)(\alpha_2+n)/(\beta_2+E)$$

$$E\left[\log(\lambda)\,|\,N=n,\theta\right]=Q_n\left[\psi(\alpha_1+n)-\log(\beta_1+E)\right]+(1-Q_n)\left[\psi(\alpha_2+n)-(\beta_2+E)\right]$$

ここで、$\psi(x)=d(\log\Gamma(x))/dx$ である。

EBGM の値は $E_{ij}\to\infty$ の場合には、相対リスク $RR_{ij}=n_{ij}/E_{ij}$ と一致する。EBGM は次の（５）で示す MGPS において用いられる値である。

（５）MGPS[1][22]

この手法は、現在 FDA で用いられている方法であり、GPS を 2 つ以上の医薬品相互作用等についても検出できるように拡張した方法である。なお、GPS と MGPS を総称して EBS(Empirical Bayes Screening)と呼ぶこともある。この手法では、医薬品相互作用を考慮した場合の期待値を得るためのモデルを対数線形モデルと仮定し、このモデルのパラメータを決定し、評価を行う。

s 層におけるアイテム（医薬品もしくは副作用）i の報告割合を P_i^s、s 層における報告数を n_s とする。2 因子の場合（医薬品 1 種類と副作用の組の場合）、期待値 E0 は、

$$E0_{ij} = \sum_s n_s P_i^s P_j^s \qquad (2\text{-}14)$$

独立と仮定した三因子の場合（医薬品 2 種類と副作用の組の場合）は

$$E0_{ijk} = \sum_s n_s P_i^s P_j^s P_k^s \qquad (2\text{-}15)$$

と計算される。

因子が 3 以上の場合にはすべての 2 つの因子の組について対数線形モデルを定義し、このモデルから与えられる期待値 E2 を求める。例えば、3 因子の場合には $E2_{ijk}$ は次の 3 つのペアから推定される。

$$\lambda_{ij} E0_{ij},\ \lambda_{ik} E0_{ik},\ \lambda_{jk} E0_{jk} \qquad (2\text{-}16)$$

同様に、4 因子では 6 つのペアから推定されることになる。なお、パラメータλは（４）GPS で示した EBGM（λの幾何平均）として与えることができる。

よって、3 因子以上の場合の期待値と 2 因子から推定される報告度数は単純な引き算により次のように比較することができる。

$$EXCESS2_{ijk} = \lambda_{ijk} \times E0_{ijk} - E2_{ijk} \qquad (2\text{-}17)$$

（６）PROFILE[8]

PROFILE 法は TGA が導入を検討している手法である。副作用報告には副作用の原因となる被疑薬が複数存在しており、どれが実際に副作用を引き起こしているのかを判断しなくてはならない。PROFILE 法は、このような場合に、ある副作用の症状と対となる医薬品を特定する手法である。

注目する副作用に対して表 2-8 を作成し、各医薬品との関連性を求める。関連性を求める方法としては、次の３つの方法を提案している。

・ 報告に記載されている第１被疑薬をそのまま対とする
・ フィッシャーの正確確率検定値（p 値）を計算し、この値が最も小さい医薬品を対とする
・ Peto 法を用いたオッズ比を計算し、この値が最も小さい医薬品を対とする

これら 3 つの方法のうち、3 つ目の Peto 法を用いたオッズ比を用いた場合は有用な結果が得られず、PROFILE には適さないことが確認されている。

フィッシャーの正確確率検定値（p 値）は、次式により求まり、$n_{11}>2$の場合に計算することとしている。

$$p=\frac{n_{1+}!n_{2+}!n_{+1}!n_{+2}!}{n_{11}!n_{12}!n_{21}!n_{22}!} \tag{2-18}$$

p 値は 2×2 表における因子同士が独立であるかを判断する際に用いる値であり、この値が小さいほど、独立ではないと判断できる。したがって、p 値が小さい医薬品ほど注目した副作用を引き起こしている可能性が高い、つまりシグナルと判断できる。医薬品と副作用の組が決定したものについては、同様の副作用における他の被疑薬は関連がないと判断できる（つまりn_{11}が減り、n_{21}が増える）。この操作を繰り返すことで、すべての報告について原因となる医薬品を特定することができる。

（７）その他[22]

（ア）Yule's Q

これは、連関を表す指標であり、期待値とその標準偏差および 95%信頼区間は以下の通りになる。

$$Q=\frac{n_{11}n_{22}-n_{12}n_{21}}{n_{11}n_{22}+n_{12}n_{21}}=\frac{ROR-1}{ROR+1} \tag{2-19}$$

$$SE_Q=\frac{1}{2}\left(1-Q^2\right)\sqrt{\left(\frac{1}{n_{11}}+\frac{1}{n_{21}}+\frac{1}{n_{12}}+\frac{1}{n_{22}}\right)} \tag{2-20}$$

$$95\%CI=Q\pm 1.96SE_Q \tag{2-21}$$

95%信頼区間の下限が 0 より大きい場合に、「シグナルあり」と判断される。

（イ）Poisson[7]

この手法は、統計に基づく容易な手法である。

注目する医薬品と副作用の組について報告件数 k とその確率 p を下記のようなポアソン分布で仮定する。

$$p = 1 - \sum_{k=0}^{a-1} \frac{e^{-\mu} \times \mu^{k}}{k!} \quad （a \text{ は観測度数、 } \mu \text{ は期待度数}）$$ （2-22）

シグナル検出の閾値となる確率 p_{th} をあらかじめ決めると、報告確率が p_{th} となる報告件数 k_{th} が決定する。この報告件数 k_{th} を超える報告がなされた場合に、この医薬品と副作用の組をシグナルとして検出する。具体的な p_{th} の値としては、p_{th}=0.005 とした研究が報告されている。

（ウ）Chi square（Yates の補正）

この手法は 2×2 分割表の検定統計量を用いる方法であり、表 2-8 に対して χ^2 値（観測度数と期待度数のズレを表す指標[20]）

$$\chi^2 = \frac{n_{++}\left(\left|n_{11}n_{22} - n_{12}n_{21}\right| - n_{++}/2\right)^2}{n_{1+}n_{2+}n_{+1}n_{+2}}$$ （2-23）

を計算し、この値が 3.84 より大きい場合に、「シグナルあり」と判断する。

（エ）ロジスティック回帰分析

これまでの手法とは異なり、この手法は一般的なロジスティック回帰分析を用いて医薬品と副作用の関係のモデルを構築し、これを元にシグナルを検出する。

対数オッズを用いたモデルとして、ロジスティック回帰式を用いて β_1 を推定する。注目する医薬品に関する変数 x と注目する副作用に関する変数 y を用いた場合、この関係を次の式で表すこととする。

$$\log it(y) = \beta_0 + \beta_1 x_1$$ （2-24）

ここで、x_1 は注目する医薬品の時に 1、その他の医薬品の時に 0 をとる変数、y は注目する副作用の時に 1、その他副作用の時に 0 を取る変数、β_0，β_1 はモデルのパラ

メータである。このとき、β_1は注目する医薬品と注目する副作用の関係を表す指標となり、このβ_1を指数変換することによりオッズ比$\exp(\beta_1)$の推定が可能となる。

シグナル検出は、β_1の95%信頼区間の下限が0よりも大きい場合に、「シグナルあり」と判断する。

また、この手法では共変量の影響を調整することもできる。例えば共変量として性、年齢、報告された日を仮定する。この場合は、ロジスティック回帰式は次のようにおくことができる。

$$\log it(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \qquad (2\text{-}25)$$

ここで、x_1は注目する医薬品の時に1、その他の医薬品の時に0をとる変数、x_2,x_3,x_4は性、年齢、報告された日の共変量、yは注目する副作用の時に1、その他副作用の時に0を取る変数、β_0,β_1,β_2,β_3,β_4はモデルのパラメータである。この関係式は共変量の影響も考慮した場合の注目する医薬品と注目する副作用の関係を表すため、先程と同様に、β_1を指数変換することによりオッズ比$\exp(\beta_1)$の推定が可能となる。

（オ）医薬品相互作用を検討したロジスティック回帰分析

（エ）と同様、注目する医薬品と注目する副作用の関係式を表すモデルを構築するが、この手法では医薬品相互作用の影響を検討するために「注目する医薬品の組合せと注目する副作用との組」に注目する。2種類の医薬品（A,B）の医薬品相互作用を想定した場合、モデルは以下のロジスティック回帰式を用いて表すことができる。

$$\log it(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 \qquad (2\text{-}26)$$

ここで、x_1は注目する医薬品Aのときに1、注目する医薬品A以外の医薬品のときに0をとる変数、x_2は注目する医薬品Bのときに1、注目する医薬品B以外の医薬品のときに0をとる変数、yは注目する副作用のときに1、その他の副作用の時に0をとる変数、β_0,β_1,β_2,β_{12}はモデルのパラメータである。

上式におけるβ_{12}は注目する副作用に医薬品AとBのどちらも影響を及ぼしていた場合に、その医薬品の組と副作用の関係を表す指標である。したがって、このβ_{12}を推定し、指数変換することにより医薬品相互作用のオッズ比$\exp(\beta_{12})$の推定が可能になる。

β_{12}の95%信頼区間の下限が0よりも大きい場合、「シグナルあり」と判断できる。

3剤以上の医薬品相互作用に関しても上記の式を拡張することで検討が可能とな

る。

（カ）逐次確率比検定(Sequential Probability Ratio Test, SPRT) [21]

SPRT とは、仮説 H_1 と対立仮説 H_2 の尤度比により検定を行う手法である。今、仮説 H_1 では報告件数 n_{ij} が期待値 E_{ij} のポアソン分布に従うものとし、一方、対立仮説 H_2 では報告件数 n_{ij} が E_{ij} の定数倍の期待値 $R \cdot E_{ij}$ に従うものとする。ここで、尤度比を λ とすると、

$$\log \lambda = n_{ij} \log(R) - E_{ij}(R-1) \qquad (2\text{-}27)$$

となる。SPRT では、検定の基準として仮説 H_1 が真であるのに H_2 を採択する第１種の過誤の確率 α 、仮説 H_2 が真であるのに H_1 を採択する第２種の過誤の確率 β を指定し、これに基づき以下のように仮説を採択する。

$\lambda \leq \dfrac{\beta}{1-\alpha}$ のとき、仮説 H_1 を採択する

$\lambda \geq \dfrac{1-\beta}{\alpha}$ のとき、対立仮説 H_2 を採択する

ある医薬品・副作用の報告件数の尤度比が $\lambda \geq \dfrac{1-\beta}{\alpha}$ のとき、期待値に対して R 倍多いとしてシグナルとして検出する。

2.5.3 シグナル検出手法以外の事例

医薬品副作用情報の分析へのデータマイニングの適用事例でシグナル検出手法以外のものは非常に少ない。WHO の Uppsala モニタリングセンター（UMC）では相互結合型のニューラルネットワークを用いた ADR パターンの検出手法の開発を行っている[2]。シグナル検出を行う際に医薬品と副作用の因果関係を適切に抽出するためには、同じ症例報告に見られる複数の ADR のパターン（共起パターン、群）をデータから抽出することが重要である。

WHO UMC では、データマイニング手法として、Recurrent Bayesian Neural Network という相互結合型のニューラルネットワークを用いた検討を行い、従来医師ら専門家が臨床的な視点に基づきデータベースを検索しながら抽出していた ADR の共起パターンを、ニューラルネットワークにより抽出できることを確認している。

2.5.4 **まとめ**

諸外国の規制当局では、シグナル検出手法の開発が行われ、安全分析業務への導入が進められている。シグナル検出手法については、いくつかの手法が提案され、性能の比較・評価が進められている。

しかしながら､シグナル検出手法の性能は分析するデータの特性に大きく依存するため、医薬品医療機器総合機構への導入を検討する場合には、機構が保有する実データの特性の分析と、実データを用いたシグナル検出手法の性能の検証が必須である。また、現在までに提案されているシグナル検出手法の多くは、主に単一の医薬品と副作用に関するシグナルを検出するものである。安全分析業務においては、併用薬や年齢などの層別のシグナルを検出する高度なシグナル検出手法の導入が要請されており、これを実現するために新たな技術的検討が必要である。

さらに、安全分析業務にシグナル検出手法を導入するためには、技術的な面以外に安全分析業務への適用の観点からの検討が必要である。シグナル検出手法はシグナルという示唆を分析担当者に与えるのみであり、安全分析業務において、検出されたシグナルに基づき、どのように判断していくのかという業務フローと判断基準については別途検討する必要がある。逆に、安全分析業務の業務フローの観点から、シグナル検出手法が検出するシグナルの量、および質にも要請が発生すると考えられ、これに関わるシグナル検出手法の判定基準の設定変更についても検討を要する。

これら安全分析業務への適用の観点からの検討を行うためには、実際の業務への試適用を通じた評価が必要となる。

3. 医薬品副作用情報分析におけるデータマイニングの概念検討

3.1 医薬品副作用情報におけるデータマイニングの概念検討

2.2 でまとめたように、近年データマイニングに注目が集まっており、他産業分野でのデータマイニングの概念は固まりつつある。一方、医薬品副作用情報の分析におけるデータマイニングについては、シグナル検出とよばれる手法の開発が進められており、当該分野におけるデータマイニング手法の中心として捉えられている。また、2.5 でまとめたように諸外国の規制当局では、シグナル検出手法の導入、および導入の検討が進められている。

シグナル検出とシグナルの定義に関する代表的なものを表 3-1 にあげる。

表 3-1　シグナルとシグナル検出の定義

■ シグナル 「それまで知られなかったか、不完全にしか証拠付けられていなかった有害事象と薬との因果関係の可能性に関する情報」(WHO[16])
■ シグナル検出 「詳細な調査を必要とする自発報告の発見およびその優先順位付けを行うこと」（藤田他[18]） 「詳細な調査が必要な自発報告の発見と調査の必要な優先順位付け」（久保田[16]）

シグナルとは、医薬品と副作用について因果関係がある可能性のあるものに関する情報であり、安全分析業務でいうところの症例分析を行うべき医薬品・副作用が抽出されているとみることができる。シグナル検出とは、まさにこのシグナルを抽出するための手法、アルゴリズムのことを指す。

3.2 データマイニング適用範囲の検討

導入するデータマイニング手法により支援を行う業務の範囲について検討を行った。2.1 で行った安全分析業務の分析結果によると、安全分析業務におけるラインリスト分析は、人手によってこのシグナル検出を行っていると見ることができる。よって、現在、諸外国の規制当局等で導入されているシグナル検出手法を導入した場合には、安全分析業務におけるラインリスト分析の部分を支援することになると考えるのが妥当である。

ラインリスト分析とシグナル検出手法を比較したものを図 3-1 に示す。ラインリスト分析では、全ての分析プロセスを分析担当者が実施する。分析においては、ラインリストの情報の他、分析担当者が持つ知識や情報を動員して、症例分析など詳細な分析が必要となる医薬品と副作用の組を抽出している。一方、シグナル検出手法では、全ての分析プロセスをシグナル検出手法が自動的に行う。シグナルを抽出する部分のアルゴリズムについて

は、前節でみたようにいくつかの手法が準備されているが、基本的な枠組みとしては２×２分割表を準備して、当該医薬品・副作用の報告件数が他の医薬品・副作用に比べて有意に多くないかどうかを判定し、多い場合にはシグナルを出すというものである。

一般的には、ラインリスト分析により抽出される詳細調査が必要な医薬品・副作用の組よりも、シグナル検出で抽出されるそれの方が、量が多くなる。これは、１つにはシグナル検出手法が基本的には報告数という限られた情報のもとで分析が行われているのに対して、ラインリスト分析では、分析担当者の知識など、より豊富な情報のもとで分析が行われるため、より精緻な分析が行えることによるものである。

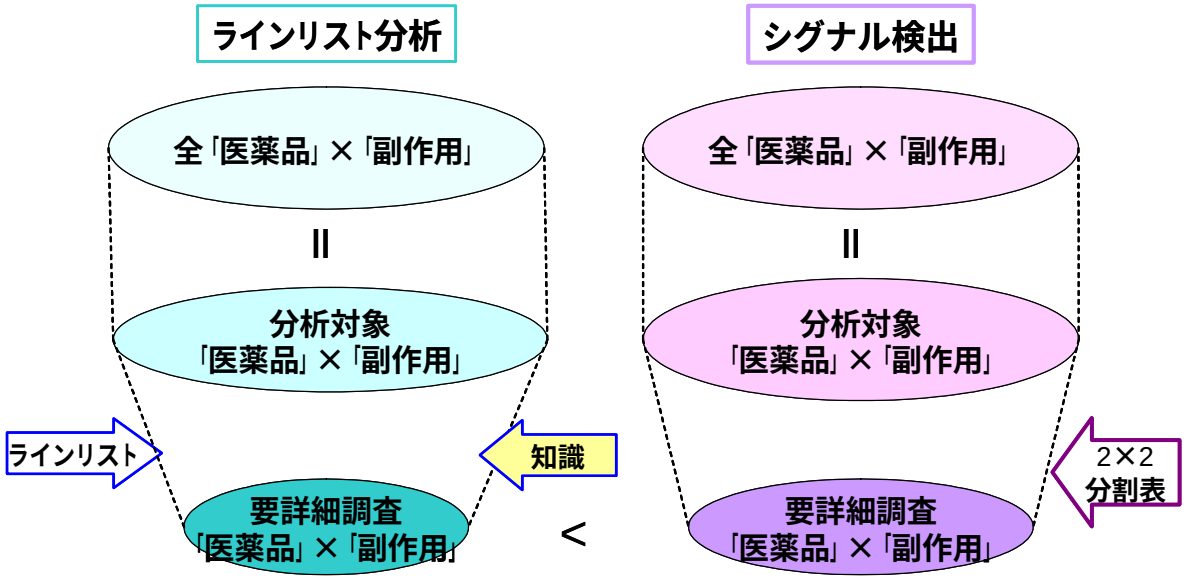

図 3-1　ラインリスト分析とシグナル検出手法の比較（１）

シグナル検出手法では、誤って詳細調査が必要でない医薬品と副作用の組についてシグナルを出してしまうことがある。これは、一般的に「False Positive (FP)」と呼ばれるものであり、余分な誤ったシグナルを出すことにより、分析担当者の負担を大きくすることが問題となっている。

ラインリスト分析とシグナル検出手法について、使用する情報、検討対象とする医薬品・副作用、検出基準、分析結果の観点から整理したものを表 3-2 に示す。シグナル検出手法のラインリスト分析に比べた場合の特長は、シグナルの検出が客観的な統計的指標に基づきなされること、シグナルの検出が自動で行われることにある。一方、シグナル検出の短所は、シグナルの検出のために使用している情報が基本的には報告件数に限られるため、過剰にシグナルを検出する False Positive の問題が発生することである。

表 3-2　ラインリスト分析とシグナル検出手法の比較（2）

	ラインリスト分析	シグナル検出手法
使用する情報	• 「ラインリスト」	• 「当該医薬品／その他医薬品」×「副作用／その他副作用」の 2×2 分割表
検討対象とする「医薬品 × 副作用」	• 全て（ただし、対象と検討優先順位は担当者の知識に基づく）	• 全て
検出の基準	• 指標:「累積報告数」、「最近の報告数」等 • 判断基準：例えば「報告数が多い」、「最近報告数が急増」など	• 指標：「累積報告数」 • 判断基準：統計的指標・閾値に基づく
分析結果	• 要詳細調査「医薬品」×「副作用」（詳細調査の優先順位）	• 要詳細調査「医薬品」×「副作用」 • 詳細調査の優先順位

安全分析業務に導入するデータマイニングの１つとしてシグナル検出手法を考えると、現状ではラインリスト分析を自動化できるレベルのものではなく、ラインリスト分析のための情報の１つとして捉えるのが妥当であると考えられる。シグナル検出手法を安全分析業務に導入した場合のプロセス図 3-2 に示す。

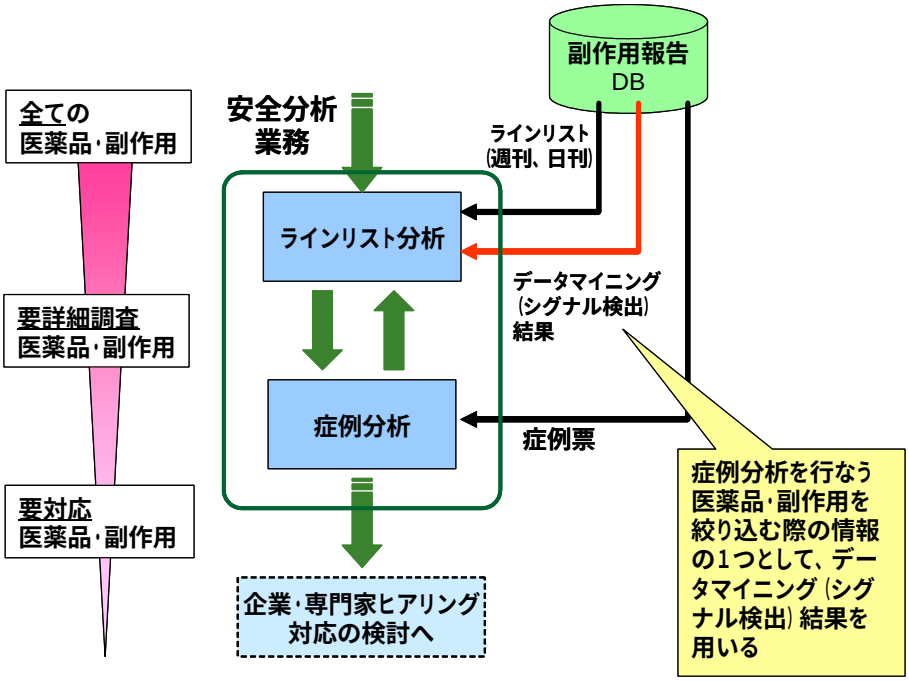

図 3-2　シグナル検出手法を導入した場合の業務プロセス

このように、シグナル検出手法をデータマイニング手法として導入した場合には、現状の安全分析業務のうち主にラインリスト分析を支援することになる。

一方、症例分析の部分では、症例票に含まれる文章を含む情報に基づいて分析が行われている。よって、データマイニング手法による支援を検討する場合には、データマイニング手法が文章を含む情報を扱えることが可能になる。すなわち、2.3.6 で説明したテキストマイニングのような技術を実装したデータマイニング手法が必要となる。近年、テキストマイニングに関する技術が進められており適用事例も増加しているが、症例分析では症例票の行間に含まれる情報、すなわち分析担当者が持つ知識に基づいて分析が行われている部分もあるので、これをデータマイニング手法として実装することは現状では困難であり、更なる技術開発が必要となる。

以上から、中期計画期間中に導入するデータマイニング手法により支援する業務は、ラインリスト分析に相当する部分を主とするべきであると考えられる。

3.3　シグナル検出手法からデータマイニングへのアプローチの妥当性検討

3.3.1　シグナル検出手法のデータマイニングとしての妥当性

医薬品副作用情報の分析においては、シグナル検出手法がデータマイニングであると認識されているが、表 2-2 に示した一般的なデータマイニングであるかを判別する基準に照らし、シグナル検出をデータマイニングとすることの妥当性について検証を行った。

まず、判別基準の妥当性について検討した。判別基準として最も一般的かつ重視される

べきは定義である。また、定義の次に一般的である判別基準は、データ量である。それぞれの判別基準の妥当性を最も妥当なものとして◎､次に妥当なものとして○とした｡また、その他の手法、自動的であることという判別基準については、先の２つの判別基準に比較すると、一般的ではないことから妥当性のランクをさらに１つ下の△とした。

次に、シグナル検出手法の各判別基準への適合性を検討した。まず、定義については、安全分析業務に有用な、それまでに分析担当者が気づいていなかったような因果関係が疑われる医薬品と副作用の組をシグナルに関する情報（知識）として抽出することを目的としており適合すると考えられるが、シグナルはあくまでも詳細な検討をするためのきっかけの情報であり深い知識ではないことから、適合性を最も高いＡから次に高いＢに該当するとした。次に、データ量については、元になる副作用情報データベース中のデータ量は大きいものの、シグナル検出を行う際には、報告数というデータに落としこんでいるために手法として扱うデータ量は一般のデータマイニングに比べると多いとは言えない。そこで、データ量に関する判別基準への適合度をＢ～Ｃとした。また、シグナル検出手法に適用されている手法の多くはいわゆるデータマイニング手法ではないこと、また、医薬品と副作用の因果関係について最終判定に近いところまで自動的に判断を行うものではないことから、それぞれ適合度をＣとした。

以上のように、シグナル検出手法は、特にデータマイニングの一般的な定義に比較的よく適合しており、データマイニングであると判断することができると考えられる。

表 3-3　データマイニングの判別基準とシグナル検出の適用

データマイニングの判別基準		判別基準の妥当性*	シグナル検出の判別基準への適合性**
【定義】未知かつ有用な知識を発見できている	結果として業務効率、成果が向上している	◎	Ａ～Ｂ
【データ量】大量のデータを分析している	人手では見切れないような量のデータを分析している	○	Ｂ～Ｃ
【手法】データマイニング手法（ツール）を使っている	決定木、ニューラルネット、相関分析などを使用している、あるいは市販ソフトを利用している	△	Ｃ
【自動的】半自動的な分析により知識（ルール）が抽出されている	データをツール（ソフト）に入れると知識が抽出される	△	Ｃ

*最も妥当であるものから、◎、○、△、×とした。

**最も判別基準に適合するものからＡ、Ｂ、Ｃとした。

3.3.2 既存のシグナル検出手法以外で導入を検討すべきデータマイニング

既存の医薬品と副作用の因果関係に関するシグナルを抽出するシグナル検出手法（基本的シグナル検出手法とよぶ）以外で導入を検討すべきデータマイニング手法について検討する。

（1）抽出知識の高度化

シグナル検出手法により抽出されるシグナルの内容を高度化した手法の導入が考えられる。例えば、次のような高度なシグナルの抽出が考えられる。

- 層別シグナル：年齢別や性別などの層別の医薬品と副作用の因果関係に関するシグナルの抽出が考えられる。
- 併用薬シグナル：現在は単独の医薬品と副作用の因果関係に関するシグナルを抽出しているが、併用薬も含めた医薬品と副作用の因果関係のシグナルとしての抽出も期待されるシグナル内容高度化の１つである。

これら層別のシグナルや併用薬のシグナルの検出は、分析担当者が分析業務で実際に抽出している内容でもあり、シグナル検出手法により抽出することが可能になれば、基本的なシグナル検出手法に比べて支援する業務の範囲を拡大することができる。また、抽出される情報も基本的シグナル検出手法に比べて高度なものとなることから、よりデータマイニング的な手法になると見ることもできる。

（2）分析データの高度化

基本的シグナル検出手法では、報告数の情報（データ）のみに基づきシグナルの抽出を行っている。3.1 で整理したように、分析担当者は報告数だけでなく、当該医薬品や副作用に関する過去の対応などの知識を用いている。例えば、シグナル検出手法において、次のような情報（データ）を利用することが考えられる。

- 既知・未知データ：分析担当者は念頭に置いている既存文書に反映済み（既知）であるか、未知であるかなどの情報をシグナル検出手法でも利用する。
- 医薬品の流通量に関する情報：報告数は医薬品の流通量の影響を大きくうけるはずである。医薬品の流通量に関する情報を用いることにより、報告数の多さをより正しく評価することが可能になる。
- 重み付け情報：症例票に含まれる情報の量に基づいて付与される重み付けを報告数の集計に反映することにより、各報告の内容の重要度を反映したシグナル検出が可能となる。

その他、現状の技術レベルでは導入が困難であるが、継続的に検討すべき内容として以下のようなものがある。

- 化学構造・薬理作用・医薬品動態データベースの利用：分析担当者が利用している化学構造、薬理作用、薬物動態に関する知識をデータベース（知識ベース）化し、これをシグナル検出に用いる。
- 症例報告票記述（テキスト）データの利用：症例分析の業務について支援を行うためには、症例報告票に含まれるテキストデータを分析できるデータマイニング手法の開発が必要である。テキストマイニング手法を含むデータマイニング手法の研究開発が必要となる。

（3）手法の高度化

基本的シグナル検出手法では、ある１つの手法により抽出された統計的指標に基づき、予め定められた閾値で評価している。しかし、医薬品群ごとに流通量が異なる等の理由から、閾値を変更した方がよい場合がある。基本的シグナル検出手法からの手法の高度化として、以下のような手法について検討を行い、導入することが考えられる。

- 複数手法・基準の利用：医薬品群等ごとに異なるシグナル検出基準、手法を利用することについて検討する。
- 複数手法の併用：2.3.7 で述べたアンサンブル学習の枠組みを導入し、複数のシグナル検出手法によりシグナル検出を行い、その結果に基づいて最終的なシグナル検出を行う手法について検討する。

上にあげた基本的シグナル検出手法に加えて、シグナル検出手法を高度化すべく導入の検討が期待されるデータマイニング手法について図 3-3 にまとめる。

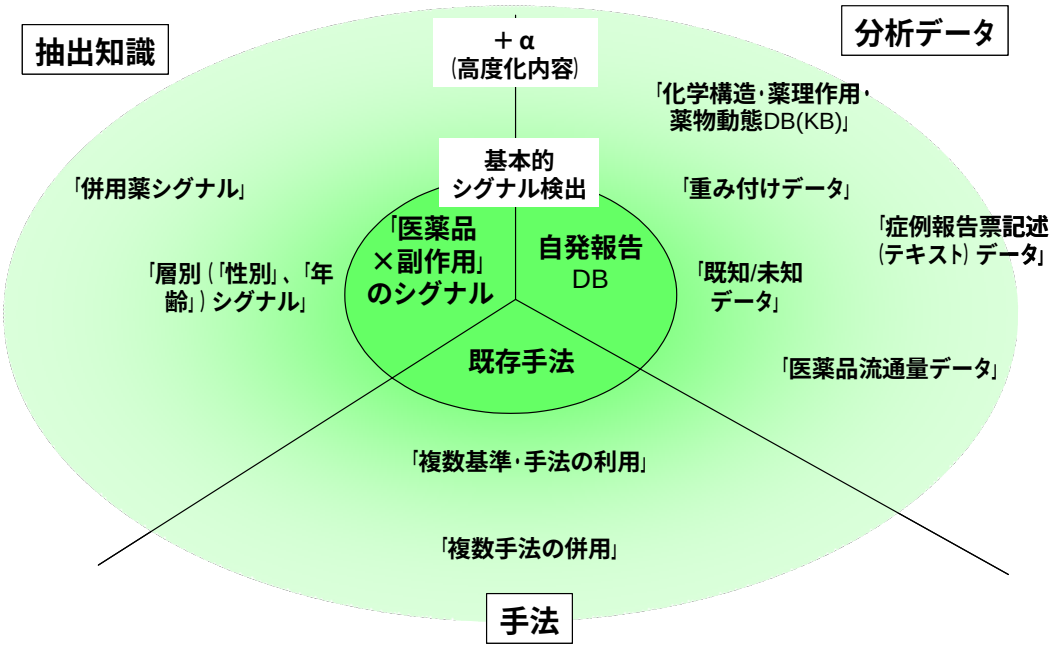

図 3-3　基本的シグナル検出手法と導入が期待されるデータマイニング

4. 中期計画期間内目標の設定と具体策検討

4.1 検討体制

中期計画期間内目標の設定と具体策検討にあたっては、下記の委員からなる「データマイニングに関する検討委員会」を設置し、中期計画期間内の目標、実施スケジュールの妥当性について審議頂いた。

（1）委員会体制

座　長：	藤田　利治	国立保健医療科学院　疫学部　疫学情報室　室長
委　員：	岩崎　学	成蹊大学　工学部　経営・情報工学科　教授
	岡田　孝	関西学院大学　理工学部　教授
	岡田　美保子	川崎医療福祉大学　医療情報学科　教授
	酒井　弘憲	日本製薬工業協会 医薬品評価委員会 統計･DM 部会 副部会長
	櫻井　靖郎	財団法人　日本公定書協会　JMO 事業部　事業部長
	野口　茂	日本製薬団体連合会　安全性委員会　委員
	平山　佳伸	厚生労働省　医薬食品局安全対策課　課長
	望月　眞弓	北里大学 薬学部 臨床薬学研究センター 医薬品情報部門 教授
	鷲尾　隆	大阪大学　産業科学研究所　助教授
	渡邊　伸一	厚生労働省　医薬食品局安全対策課　課長補佐

（2）委員会開催日程

第1回：平成１６年１２月１４日
第2回：平成１７年　1月２４日
第3回：平成１７年　3月　7日

（3）検討結果の公表

委員会により検討された内容については、本報告書でまとめる内容以外にも医薬品医療機器総合機構より、公表資料として公開される。

4.2 中期計画期間内目標設定

（1）中期計画期間中に導入するデータマイニング手法

前節までに検討を行った内容に基づき､中期計画期間内に導入するデータマイニングに関する目標を設定する。

中期計画期間内に導入するデータマイニングについては､諸外国規制当局が基本的シグナル検出手法の導入と活用を進めている現状も鑑みて、

「基本的シグナル検出手法を中心として安全対策業務の向上に資するように高度化したもの」

を対象とする。

この目標を達成するために、以下の（ア）から（ウ）の検討を行う。

（ア）データの持つ特性把握のための検討

2.2.3 でまとめたとおり、シグナルをはじめとする有益な情報を得るためには、データマイニング手法を適用する前にデータの持つ特性を十分に分析､理解することが必要となる。これに関わる内容として以下の検討を実施する。

- データの特性に関する分析：報告者毎に報告の記述やコーディングに異なる傾向が見られる、あるいは市販後の時期に関係して報告数が異なる等の理由により、自発報告データには一定の傾向、特性が含まれると考えられる。データマイニングを適用し、結果を理解する上では、この特性を十分に理解することが前提となる。データの持つ特性とこれが生じる背景を理解するために、データの分析を行なう。
- 背景知識の利用検討：2.1 の業務分析で明らかになったように、専門家や分析担当者は、上記のようなデータの持つ特性をはじめとしてデータを正しく解釈するのに必要な知識（背景知識）を用いている。データマイニングのプロセスに分析担当者が持ち、分析に用いている背景知識をデータの前処理、およびシグナル検出手法を中心としたデータマイニング手法に導入、利用する方法について検討を行う。
- データの前処理に関する検討:データマイニングの結果の質を向上させるために、データの誤り、欠損値、およびデータの持つ特性に対処するための前処理方法について検討を行う。
- データベース構成に関する検討：シグナル検出を中心としたデータマイニングに必要なデータのデータベースでの管理方法について検討を行う。

（イ）基本的シグナル検出手法に関する詳細検討

基本的シグナル検出手法とは、自発報告データに基づき、安全対策業務の観点から医薬品・副作用の組を抽出する手法である。基本的シグナル検出手法については、2.5 で調査結果をまとめたように既に研究開発された既存手法があるため、これらを中心に業務への導入に必要な詳細検討を行う。

詳細検討で実施する内容としては、以下のようなものがあげられる。

(a) 手法に関する検討

- 適用手法（アルゴリズム）の選択：海外の規制当局で利用されている PRR、BCPNN、EBM など既存手法については、諸外国の規制当局が保有するデータ、および手法を評価するための人工データによる評価が行われているが、医薬品医療機器総合機構が保有する自発報告データを用いた評価は行われていない。そこで、まず医薬品医療機器総合機構の保有する自発報告データに対して既存手法を適用し、シグナル検出の試行、および評価を行う。
- 検出基準の設定：シグナル検出を行う際の検出基準（閾値）の設定を行う。

(b) データに関する検討

- 医薬品、副作用に関する集計単位：医薬品、および副作用についてシグナル検出のために報告数を集計する際の集計単位を検討する。特に副作用の集計単位については、MedDRA（医薬規制用語集）における用語階層レベルや SMQ（Standardized MedDRA Queries、MedDRA 標準検索式）の利用等について実験による評価を通じて、最も適したものを見いだす。
- データの利用期間：シグナル検出に利用する報告データの期間を設定する。
- 既知・未知データの利用：添付文書のうち重大な副作用に関する項目への反映済（既知）、未反映（未知）のデータを利用するシグナル検出手法の検討を行う。

（ウ）シグナル検出手法に関する高度化検討

基本的シグナル検出手法を安全対策業務の向上に資するものとして高度化する方向としては、以下の２つに分類される。

① 抽出する知識・情報（シグナル）の内容を高度化する
② 抽出する知識・情報（シグナル）の精度を高度化する

①の内容の高度化については、安全対策業務に必要な情報という観点からは、基本的シグナル検出手法により抽出される医薬品と副作用のシグナル以外にも、安全対策業務に資するものとして以下の情報の抽出が期待される。

1. 層別シグナル

「性別」や高齢者、小児といった「年齢」などの層別のシグナル情報。安全対策業務では、これら層別に注目した分析を行っている。

2. 併用薬シグナル

併用薬の使用状況を考慮したシグナル情報。安全対策業務では、実際に併用薬を念頭

に置いた分析を行っている。

基本的シグナル検出手法からの高度化の内容については、以下で挙げる高度化の候補内容について技術的検討を通じて、中期計画期間中の導入の可否の判断を行う。

②の抽出する知識・情報の精度向上については、分析に利用する手法（アルゴリズム）、分析対象データについて以下のような検討を行う。

(a) 手法（アルゴリズム）

- 複数手法・基準の利用：医薬品群等ごとに異なるシグナル検出基準、手法を利用することについて検討する。
- 複数手法の併用：複数手法によりシグナル検出を行い、さらに各手法によるシグナル検出結果に基づいて、最終的なシグナル検出を行う手法を検討する。機械学習手法の分野で注目されている Boosting 手法、Bagging 手法などアンサンブル学習法の適用の検討を含む。

(b) 分析対象データ

- 重み付けデータの利用：症例報告の情報量に基づいて付与された重み付けデータを利用するシグナル検出手法について検討を行う。
- 医薬品の流通量に関するデータの利用：処方情報など各医薬品の流通量に関するデータの利用についても実現可能性の検討を行う。

（2）今後の導入を目指して検討を行うデータマイニング手法

中期計画期間中の安全対策業務への導入は困難だと判断されるものの、今後の導入を目指して検討を行うべきであるデータマイニング手法として以下のようなものが挙げられる。

- 症例報告票記述データの分析：症例報告票に記述されている情報のうち、コード化されていない記述（テキスト）データをシグナル検出手法に利用するための方法について検討を行う。
- 化学構造・薬理作用・薬物動態データ（知識）ベースの適用：分析担当者が分析時に用いている当該知識について知識ベース化し、これを分析に利用するシグナル検出手法について検討を行う。

4.3　中期計画目標達成のためのスケジュール策定

（１）実施項目

平成１８年度までの手法確立、平成２０年度までの業務への導入に向けて、検討会を設け専門家の意見を聞きつつ、以下の項目について実施する。（図 4-1）。

【平成17年度～平成18年度】

平成１８年度末までの手法の確立を目的として以下の検討を行う。

(ア) データマイニングを用いた安全業務プロセスの検討

今回導入するデータマイニング手法（シグナル検出手法）が提供する情報に基づき安全業務プロセスの信頼性向上を実現するための具体的な方法について検討を行う。検討する内容には以下の点を含むものとする。

- 検出されたシグナルをはじめとする情報に基づく安全対策業務上の判断基準、業務フローの検討
- 次項（イ）で検討されたデータマイニング手法の業務への試適用と評価
- 検出されたシグナル、およびシグナル検出のために算出された指標などの情報の分析担当者への提示インタフェース、提供リストに含むべき情報項目の検討
- データマイニング手法の運用を含む人員体制

(イ) 業務に導入するデータマイニング手法の確定

業務へ導入する「基本的シグナル検出手法を中心として安全対策業務の向上に資するように高度化した」データマイニング手法を確定するために、以下の検討を行う。

- データの持つ特徴を把握するための検討
- 基本的シグナル検出手法に関する詳細検討
- シグナル検出手法に関する高度化検討

これらの検討を行うことにより、平成１８年度末には、手法の確立を完了し、業務への仮導入が可能な状況とする。

【平成19年度～平成20年度】

(ウ) 業務システムの開発

平成１８年度末までに確立されたデータマイニング手法を実際に業務に導入するための業務システムの開発と業務の試運用を平成２０年度末までに実施し、業務への導入を完了する。

（２）実施状況の公表

実施状況については適宜公表する。

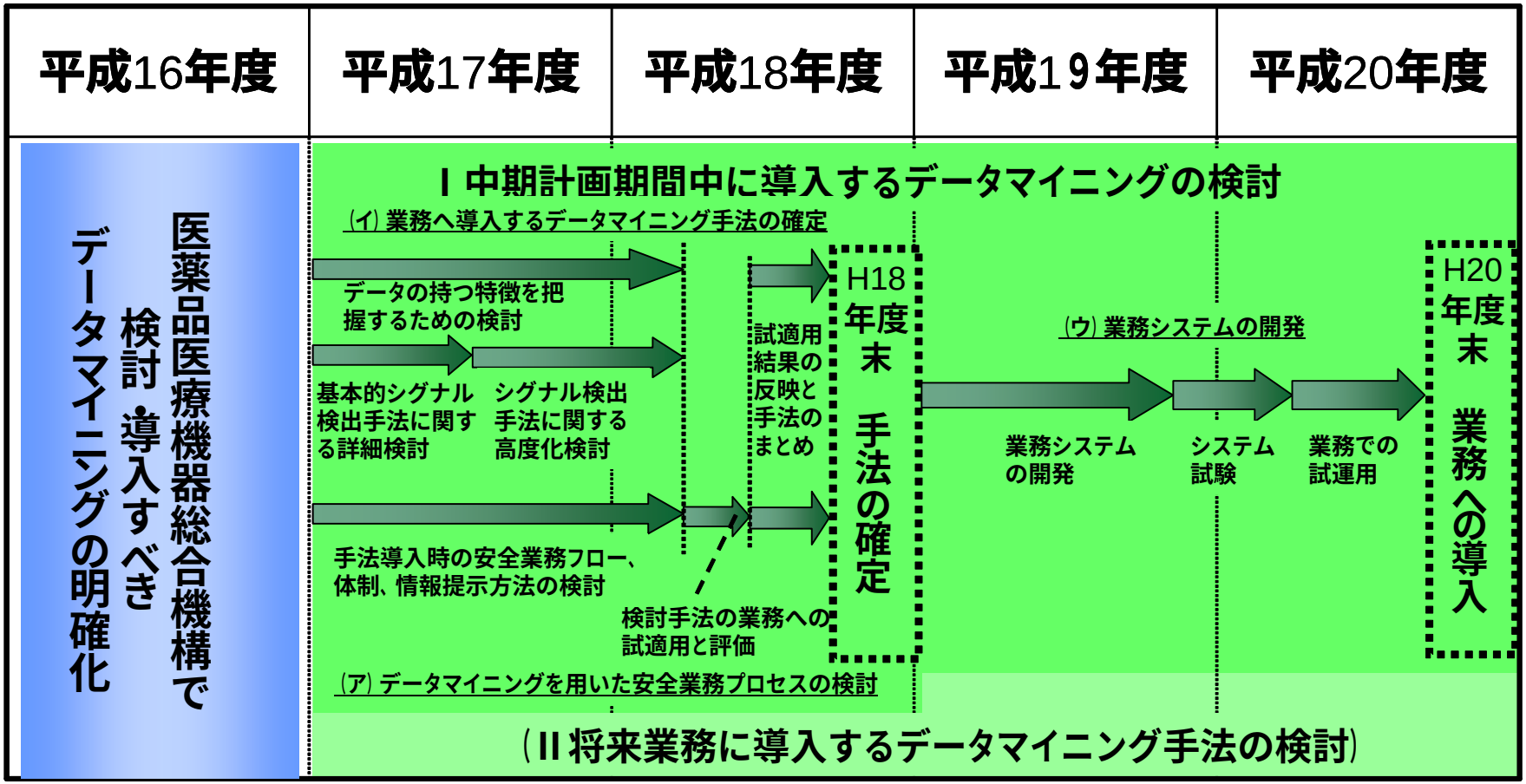

図 4-1　中期計画期間中のスケジュール

4.4　医薬品副作用情報のあり方に関する検討

シグナル検出をはじめとするデータマイニング手法を適用することに限らず、安全対策業務の高度化のためには、自発報告データの質および量の向上が期待される。

自発報告データの質の問題については、例えば以下のような問題がある。

- 当該医薬品の上市初期に報告が集中する傾向がある
- 添付文書の改訂など対応が行われたことにより注目された事象に報告が偏在する傾向がある
- 報告者である製薬企業間、および報告者個人ごとにコーディングが異なる傾向が見られる
- 報告に記述されている情報が分析に必要な量に比べて不足している

また、自発報告データの量の問題については、報告を行う基準が報告者ごとにまちまちであり、少なく報告される傾向があるという点があげられる。

これら自発報告データの質、量の問題を解決するために医薬品医療機器総合機構として、製薬企業・医療機関に対して必要な働きかけを行っていくことが必要である。

4.5　システム検討

（1）システムの概略

基本的シグナル検出を行うシステムの概要を図 4-2 に示す。基本的シグナル検出は、上流側のシステムである副作用報告受付システムから作られる、副作用報告データベースのデータが利用される。また、基本的シグナル検出の結果は、安全分析業務全般の情報提供を実施する、情報提供システムに引き渡される。図中の点全部分が基本的シグナル検出を行う部分で今回のシステム検討の対象である。以下、この部分をシステムと呼ぶ。

（2）データに関する検討

システムで利用されるデータの種類と内容及び検討事項を表 4-1 に示す。

表 4-1　データの概要

データの種類	内容	検討事項
副作用報告 DB	現在の安全分析業務で利用されている副作用報告のデータベース	業務への導入にあたって、副作用報告受付システムとの連携を検討する必要がある。
シグナル検出用 DB	シグナル検出を行うために、副作用報告データベースから必要なデータのみを抽出したデータベース	データマイニング手法の検討が既存の安全分析業務に影響を及ぼさないようにするために、副作用報告データベースとは独立させる方式が考えられる。なお、平成 17 年度以降の検討によっては、必ずしもデータベース構成とする必要は無く、フラットファイルの抽出処理機能を定型化する方式も考えられる。
フラットファイル	・ 医薬品コード ・ 副作用コード ・ 報告件数 ・ その他（性別，年齢，併用薬等） で構成されるデータ	
２×２分割表	シグナル検出の入力データ	
副作用報告 DB 以外のデータ	副作用報告 DB に含まれないデータ	シグナル検出手法に関する高度化検討により新たに必要とされるデータであり、医薬品医療機器総合機構内部に新たにデータベースとして整備する方式や、インターネット等を利用して、外部のデータを利用する形式が考えられる。

上記の検討事項は、平成１７、１８年度に実施される「データマイニングを用いた安全業務プロセスの検討」を通じて実施されるものとする。

（３）処理に関する検討

システムの処理と概要及び検討事項について表 4-2 にまとめる。

表 4-2　システムの処理概要

処理	概要	検討事項
データ抽出	副作用報告データベースからシグナル検出用のデータベースを作成する。	必ずしもデータベースを作成する必要は無い。直接フラットファイルを作成する処理方法も考えられる。
フラットファイル作成	シグナル検出用データベースからフラットファイルを作成する。	副作用報告データベース以外のデータを利用する場合には分散データベースへの処理やインターネットを通じたデータのやり取りの実施等を行う可能性がある。
分割表作成	フラットファイルから２×２分割表を作成する。	
シグナル検出	基本的シグナル検出を行う。	
結果表示	シグナル検出の結果を分かりやすく表示する。	表示方法に工夫が必要になる。

上記の検討事項は、平成１７、１８年度に実施される「データマイニングを用いた安全業務プロセスの検討」を通じて実施されるものとする。

（４）安全分析業務全体のシステムとの連携

図 4-2 に示すとおり、基本的シグナル検出を行うシステムは、上流側である副作用報告受付システム及び下流側である情報提供システムとの連携が円滑であることがきわめて重要となる。通常の安全分析業務に加えて、基本的シグナル検出の結果をどのように安全分析業務に反映するのかを具体的に検討したうえで、それぞれのシステムを連携させる必要がある。最適な連携方法については、平成１９、２０年度に実施予定の「業務システムの開発」において詳細検討及び開発が実施される必要がある。

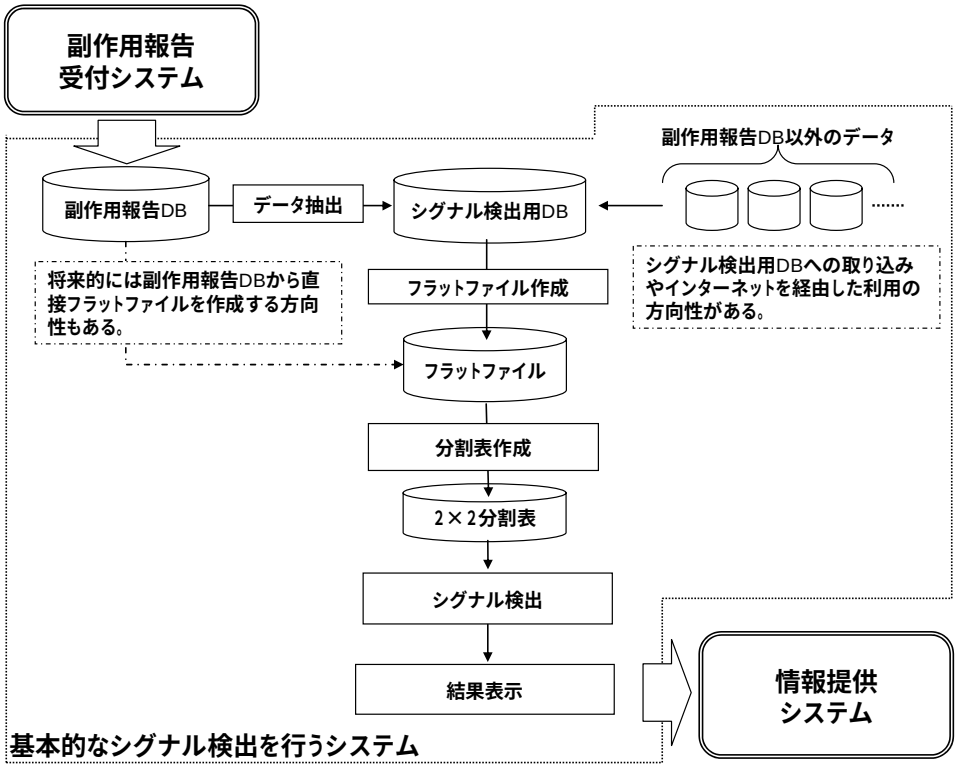

図 4-2　システムのデータ構成と処理フローの概略

（5）大量データ処理と関連する IT 技術

副作用報告データベース以外のデータを利用する場合には、シグナル検出用データベースへのデータの追加やインターネットを利用したデータの利用が考えられる。利用するデータが物理的に複数個所に分散していてネットワークまたはインターネットを利用する場合には、グリッドやクラスタを利用する可能性がある。

（ア）グリッド

グリッドの定義としては、グリッド協議会によるものがある[3]。その内容は、以下のとおりである。

「グリッドは広域ネットワーク上の計算、データ、実験装置、センサー、人間などの資源を仮想化・統合し、必要に応じて仮想計算機（Virtual Computer）や仮想組織（Virtual Organization）を動的に形成するためのインフラです。」

[3] グリッド協議会ホームページ http://www.jpgrid.org/about/definition.html

現在、産業技術総合研究所のグリッド研究センター[4]等で最先端の研究が実施されている。グリッド研究センターの研究テーマの１つに、データベースグリッドの研究がある。これは、グリッド環境での大規模なデータ応用を支援する、分散データベース統合や情報サービスへの応用を目指した研究開発である。現時点ではグリッドは研究開発のフェーズであるが、この分野の進展は早い。平成 21 年の新たな業務システムの本格導入の際には、この技術が実用化され､広く利用されている可能性もある｡今後のシステム検討においては、このような技術の利用可能性も念頭において進める必要がある。特に、医薬品医療機器総合機構の外部にあるデータベースを利用する場合など、グリッドが有効利用できる可能性がある。

（イ）クラスタ

ネットワーク接続された複数のコンピュータのことであり、分散処理用のソフトウェアを用いることによって、一つの計算機資源として利用可能な並列もしくは分散処理システムを意味する。多数のコンピュータを協調させる技術という点ではグリッドと類似しており、グリッドで用いられる分散処理技術がクラスタを構成するときに用いられる例もあることから、グリッドという言葉がクラスタと同じ意味で使用されることも多くなってきている。グリッドはグローバルな協調分散を志向したネットワーク・ソフトウェア技術が主体であり、クラスタはネットワーク接続された並列・分散処理が可能な複数のコンピュータを意味すると考えてよい。医薬品医療機器総合機構内部にあるデータベースの高速分散処理などには、ハードウェアとしてクラスタの利用が有力である。並列処理用のソフトウェアとしては、MPI という通信ライブラリの使用実績が豊富であるが、グリッドの分散処理技術を利用することも可能である。クラスタは既に豊富な実績があるが、実際にシステムに取りいれる際には、上流や下流のシステムとの連携なども含めて十分な検討が必要である。

今後、さらに上記以外の高速な分散処理技術の研究開発も進むと考えられるが、システムの本格的な導入に当たっては、将来に渡る保守や拡張を考慮し、導入準備段階で最先端の技術動向を把握した上で主流となる技術の検討を行うことが重要である。

[4] 独立行政法人産業技術総合研究所グリッド研究センターホームページ、センター概要パンフレット http://www.gtrc.aist.go.jp/jp/intro/041019gridJ.pdf

5. 調査のまとめ

データマイニング手法の検討を行うための支援業務として、医薬品医療機器総合機構で中期計画期間内に導入するデータマイニング手法を明確にし、さらに中期計画期間内の実施スケジュールを策定することを目的として調査を行った。

諸外国の規制当局で導入されているデータマイニングに関する調査から、諸外国では医薬品と副作用の因果関係が疑われるものをシグナルとして抽出するシグナル検出手法（基本的シグナル検出手法)がデータマイニング手法として導入されていることが確認された。まず、この諸外国の規制当局の動向にキャッチアップするため、医薬品医療機器総合機構では未導入であるシグナル検出手法を導入することを中期計画期間中に導入するデータマイニング手法の中心部分として設定するものとした。導入するシグナル検出手法については、対象としているデータが今回分析の対象とするデータではないが基本的シグナル検出手法については諸外国の規制当局のデータなどを用いた研究開発が既に進められていること、一方、層別シグナルや併用薬シグナルの抽出、およびシグナル検出の高精度化など基本的シグナル検出手法をさらに高度化した手法の導入が期待されていることを鑑みて、中期計画期間中に導入するデータマイニング手法を「基本的シグナル検出手法を中心として安全対策業務の向上に資するように高度化したもの」と設定した。

データマイニングに関する基本事項、異業種におけるデータマイニング導入の成功事例に関して調査した結果、データマイニングを行なうにあたり、データの特徴を十分に把握すること、分析担当者が持つ背景知識をデータマイニングに導入することが導入成功の鍵となることが明らかとなり、これらの検討の実施には２ヵ年度を要するため、平成１７年度から２ヵ年の手法の開発では、手法の高度化のための各種の検討を行うとともに、データの特徴を把握するための検討と背景知識をデータマイニングに導入するための検討を行うものとした。

さらに、データマイニングは基本的に自動的に知識を抽出するプロセスではなく、分析者を支援し、最終的に分析者が知識を発見するプロセスであることから、業務への親和性が重要であることを再確認した。この業務への親和性を高めるための検討を、平成１７年度からの検討において、データマイニングを用いた安全業務プロセスの検討として実施することとした。

平成１８年度までに開発されたデータマイニング手法を業務に導入するため、平成１９年度から平成２０年度までの２ヵ年にわたり業務システムの開発を行うものとした。業務システムの開発は、既存のシステムとの整合性をとることが必要であるため、これに関わるシステムの開発、改修を含むものである。また、平成２０年度の中間期以後は、開発した業務システムを用いたデータマイニング手法を用いた安全対策業務についてテストを行う、平成２０年度末の時点では、データマイニング手法の業務への導入が完了するものとした。

参考文献

[1] A. Szarfman et. al., “Use of Screening Algorithms and Computer Systems to Efficiently Signal Higher-Than Expected Combinations of Drugs and Events in the US FDA's Spontaneous Reports Database”, Drug Safety , Vol.25, No.6, 381-392, (2002)

[2] Bate et. al., ”Pattern detection for celecoxib and refecoxib in the who database” (http://www.who-umc.org/index2.html).

[3] C.L. Blake, C.J. Merz, ”UCI Repository of machine learning databases” [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science., (1998).

[4] E. P. van Puijenbroek et. al., “Application of Quantitative Signal Detection in the Dutch Spontaneous Reporting System for Adverse Drug Reactions”, *Drug Safety* , Vol.26, No.5, 293-301, (2003)

[5] H. Kargupta et. al., Data Mining: Next Generation Challenges and Future Directions, The AAAI Press, (2004).

[6] L. Breiman, “Bagging Predictors”, *Machine Learning*, Vol. 24, No. 2, 123-140, (1996).

[7] M. Hauben et. al.,”Quantitative Methods in Pharmacovigilance Focus on Signal Detection ”, *Drug Safety* , Vol.26, No.3, 159-186, (2003)

[8] P. Purcell, S. Barty, “Statistical Techniques for Signal Generation The Australian Experience”, *Drug Safety* , Vol.25, No.6, 415-421, (2002)

[9] R. Agrawal, R. Srikant, “Fast Algorithms for Mining Association Rules”, *Proc. of the 20th Int'l Conference on Very Large Databases*, 487-499, (1994).

[10] R.E. Schapire, Y. Freund., P. Bartlett, and W.S. Lee, Boosting the margin, “A new explanation for the effectiveness of voting methods”, *Proc. of 14th International Conference on Machine Learning*, 322-330, (1997)

[11] U. M. Fayyad et. al., “From Data Mining to Knowledge Discovery: An Overview”, Advances in Knowledge Discovery and Data Mining, 1-34, The AAAI Press, (1996).

[12] 阿部重夫，ニューラルネットとファジィシステム，近代科学社，(1995).

[13] 石井哲，テキストマイニング活用法，リックテレコム，(2002).

[14] 上田修功，アンサンブル学習，計測と制御，Vol.41，No.3，248，(2002).

[15] 岡田孝他，アクティブマイニングによる化学物質群からのリスク分子発見，人工知能 学会誌，VOl.20，No.2，211-218，(2005).

[16] 久保田潔，自発報告からのシグナル検出，－英国 MCA，米国 FDA，WHO の新しい方法－，薬剤疫学，Vol.6，No.2，101-108，(2001).
[17] 林俊克，Excel で学ぶテキストマイニング，オーム社，(2002).
[18] 藤田利治他，医薬品の副作用自発報告によるシグナル検出の実用化に向けての検討，厚生労働科学研究補助金（医薬品等医療技術リスク評価研究事業）分担研究報告書，(2004).
[19] 堀聡，修理伝票のデータマイニング－市場品質管理の効率化－，計測自動制御学会誌，Vol.41，No.5，342-344，(2002)
[20] 松原望他，　統計学入門，東京大学出版，(2000).
[21] 松原望，意思決定の基礎，朝倉書店，(2001)
[22] 渡邉裕之他，重要な安全性情報を早期に検出する仕組み－シグナル検出の最近の手法について－，計量生物学，Vol.25，No.1，37-60，　(2004).