汎用解析ツールを用いたデータマイニングの基礎考察

A Basic Consideration on Data Mining using All-Purpose Analyzing Software

須藤 敦史*
Atsushi SUTOH

＊ 工学博士（株）地崎工業 土木技術部 主席研究員(〒105-8488 東京都港区西新橋2-23-1)

In recent years, data accumulation has been remarkably increased, and it is pointed out that it is necessary to restore data in order, and to discover rules hidden behind data. The notion of data mining is first clarified and is interpreted as a non-structural inverse problem. In this study consists of the following two topics, one is a basic consideration on a data mining which is the power of current data-processing functions. Then, it is obtained that interesting knowledge rules from database of seismic damages of water supply system. And the other, we introduce data mining with a view to discuss applications of an all-purpose analyzing software. Finally, it is found that the usefulness of these data mining procedures using an all-purpose analyzing software, for finding out rules of seismic damages of water supply system.

Key Words: *data mining, water supply system, seismic damage, all-purpose software*

1．はじめに

近年,コンピュータ性能の向上やネットワーク技術の発達により,データウエアハウスの活用要求が増加してきている.しかし一般的にデータベースは定性・定量的データが混在し,かつデータ間の相互関係が複雑であるため,確率・統計,機械学習など多様な枠組みで試みられている [1]が現状では効果的な利用がなされていないのが現状である.

そこでデータ間の相関関係や新たな知識・ルールなどの発見・発掘を目的としたデータベースからの知識発見（KDD：Knowledge Discovery in Databases）あるいはデータマイニング（DM：Data Mining）[2,3]が注目されている.

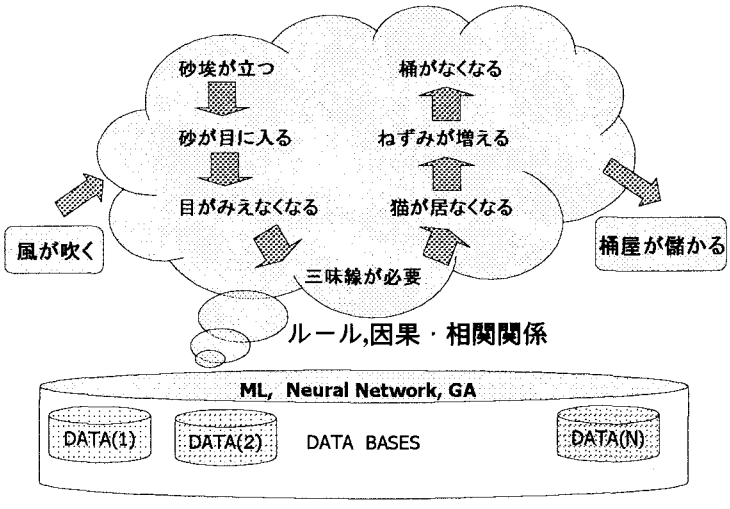

図-1 データマイニングの概念図

そもそもDMは図-1に示すようにデータベース中の隠れているデータ間の相関関係やルールを客観的に発見するための現実問題として新たなプロセッシング技術あるいは考え方であり,特にマーケティングの分野などで多くの応用され [4,5],著者らも様々な事象・現象の解析に適用している [6,7,8].

しかし,その反面取り扱うデータ量の増加や用いる解析手法の複雑化・ブラックボックス化に伴い,本来要求されている迅速なデータ間の相関関係や情報,知識やルールの検索が難しくなってきているのも事実であり,そのため簡単に分析を行うための汎用解析ツールを用いたDMやデータ分析の手順・方法および抽出された相関関係やルールに対する客観的な評価方法が求められている.

そこで本研究では以下に示す事項をKey Wordsとして,阪神大震災における神戸市・芦屋市・西宮市の上水道配水管の地震被害データをもとに最もポピュラーな汎用的な統計解析ツールであるMicrosoft Excelを用いたデータマイニングにより,上水道配水管の被害状況や影響を与える様々な地盤条件についての基礎解析を行っている.

①簡便な（ユーザーフレンドリーな）DM
②データ解析の手順（ドリルダウンの方法）と結果の評価

2．データマイニング

DM が現れた背景には以下の社会的・技術的要因が影響していると考えられる．

(a)データウエアハウスの発展・発達

コンピュータの進歩で膨大でかつ多様なデータの蓄積が行われ,これらの有効活用が社会的・技術的に求められている．

加えて,マーケティングの分野などでは社会構造の変化により消費者の嗜好の移り変わりが早くなっているため,蓄積されたデータを迅速に解析しないとデータベースの有効活用が難しくなりつつある．

(b)解析理論・技術の統合

従来のデータ解析理論もしくは技術を統合・システム化した新しい解析手法が求められている．

しかし,理論の複雑化・ブラックボックス化などにより,ユーザーは新しい解析手法に対して若干の戸惑いを示しているのも事実である．

(c)解析技術・手法の汎用化・ソフトウェア化

実際のデータを解析するため汎用化されかつ操作方法が簡単なツールもしくはソフトおよび標準的なデータ解析の手順（ドリルダウンの方法）と簡単な結果の評価方法が求められている．

一般的に DM は「仮説検証型」と「仮説生成型」に分けられ,前者は仮説のデータによる検証すを,後者はデータを単純な表現形式に変換して隠れたルールの発見することを目的としている．しかし,両者とも従来のデータ解析に比較して「状況予測」より「結果解釈」に重点を置いているところが特徴である．

ここで本研究では,汎用的な統計解析ツールを用いた初期解析とデータ解析の手順（ドリルダウンの方法）の検討を目的としているため,「仮説検証型」を使用している．

3．解析ツールとデータマイニングの手順

（1）汎用解析ツール（Microsoft Excel）

最近はパソコンの急激な普及とネットワーク技術の発達に伴って,汎用的な統計ツールもしくはソフトが手軽に使用できるようになってきており,その中で代表的な統計ソフトは STATISTICA,SPSS,SAS,S-PLUS などがあげられる．ここで汎用的かつ簡易なデータマイニングを行うにはパソコンで稼動する統計ソフトをいかに活用するかが重要なことであり,データマイニングを実現するパソコンの汎用的な統計ソフトが有する要件は以下となる．

a)多くの統計手法がサポートされている．

b)多量なデータを高速で処理できる．

c)データ数・変数等の制約が少ない．

d)安価で手軽であり,かつ多数が利用している．

e)必要に応じて統計プログラム（マクロ等）の作成ができる．

f)他の表集計ソフトとの連携がよい．

g)マニュアル・ヘルプ機能等が完備・充実している．

h)グラフ機能が充実している．

以上より,現在の環境では汎用的な統計解析ツールであるため,Microsoft Excel を用いてデータマイニングを行っている．

（2）データマイニングの手順

DM とは,直訳すれば「データの発掘」であり膨大なデータの中に隠れている知識や規則を客観的に発見することであり,その標準的な手順は図-2 に示す 6 段階のプロセスが基本となっている．

(a)データの選択(selection)

事象・現象とその解析目標を設定し,基本データベースから必要なデータを選定してマイニング用のデータを構築する．

(b)データの洗浄(cleaning)

構築された解析用データベースから,ノイズや異常値を除去してクリーンなデータを構築する．ただし,この洗浄は事前にできるものもあれば,データのコード化や知識・規則の発見の段階になってから実行する場合も生じる．

(c)データの補強(enrichment)

有用なデータや新たなデータを追加し,他データベースとの結合を考慮した構造形式にする．

(d)データのコード化(cording)

データをマイニングしやすい形式に変換する．

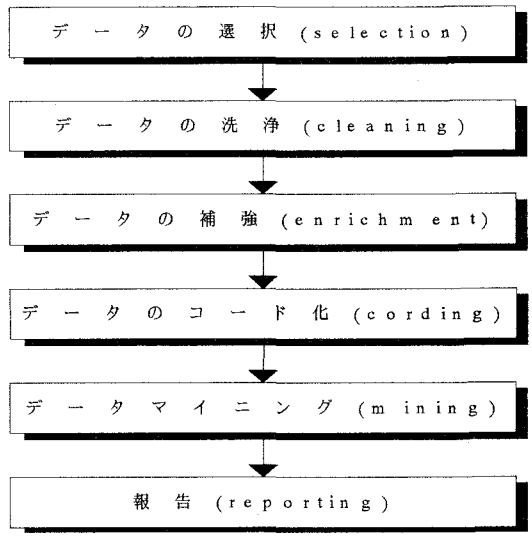

図-2 データマイニングの手順

(e) データマイニング(mining)

前処理が済んだデータベースからの知識・規則（ルール）の発見を行うものであり,最も重要なプロセスである.

4.上水道供給システムの地震被害解析

（1）使用データおよび観測項目

DMに用いたデータベースは,阪神大震災の神戸市・芦屋市・西宮市における上水道配水管の地震被害調査データ[9]を用いており,観測項目は震度,地盤種類・地形,管径・管種と被害の有無などである.

ここで日本水道協会報告書の調査データを用いているため, データの選択(selection)とデータの洗浄(cleaning)の一部は完了しているものと考えられる.

（2）データマイニングにおける目標の設定

ここで,これらの配水管における被害のデータ（項目）に対して「どのように解析を行っていくか？」が問題となる.

そこで,まず以下に示すもっとも単純（常識的）な地震と配水管における被害データの関係が成立するかどうかを検証する.

「震度が大きくなるに伴って被害も増加する」

ここで観測データの各項目ごとに被害件数,被害率（異なる被害の指標）において仮説との相関関係を3ランク（段階）を（Yesは71～100%で仮説成立, Noは0～30%で仮説成立, －は31～70%で仮説成立か比較不可能）を設定した.

また観測項目のランク分けに対しては,データの平均値以上・以下とするブール属性化（0-1）が最もシンプルであり,その後3ランク（たとえばgood（100～70%）, fair（70～30%）, poor（30～0%）など）4ランク（excellent（100～75%）, good（75～50%）, fair（50～25%）, poor（25～0%）など）と細分化して行くのが常套手段である.

ここでもっとも単純（常識的）な解析結果を都市別にまとめたもの表-1,管径別を表-2,管種別を表-3に示す.

5．データマイニングの評価について

まず,表-1に示す都市別被害表からは沖積平野（液状化無し）と良質地盤には仮説が成立し,改変山地,段丘,谷・旧水部,沖積平野（液状化有り）では仮説が成立してない.

つまり, 沖積平野（液状化無し）と良質地盤は震度が大きくなるに伴って被害率・件数ともに増加する(相関関係を有する)ため,被害予想がしやすい地質・地形であるが,それ以外は震度が大きくなるに伴って被害が増加するとは限らないため,別の項目・要因との（詳細検討事項を抽出,さらに都市による相違も!）相関関係が考えられ,また震度による被害予測が困難な地形であることも示唆し

表-1　都市別仮説検証

	被害件数			被害率		
	神戸市	芦屋市	西宮市	神戸市	芦屋市	西宮市
Ⅰ改変山地	No	No	No	No	－	No
Ⅱ段丘	No	No	No	No	No	No
Ⅲ谷・旧水部	No	No	No	No	No	No
Ⅳ沖積平野(液状化無し)	Yes	Yes	No	No	Yes	Yes
Ⅴ沖積平野(液状化有り)	No	－	No	No	No	No
Ⅵ良質地盤	－	－	－	Yes	Yes	No

表-2　管径別仮説検証

	被害件数				被害率			
	75~75	100~150	200~450	500~800	75~75	100~150	200~450	500~800
Ⅰ改変山地	No	No	No	－	－	－	－	－
Ⅱ段丘	No	No	－	－	－	No	No	－
Ⅲ谷・旧水部	No	No	No	Yes	No	No	No	－
Ⅳ沖積平野(液状化無し)	－	－	Yes	No	Yes	Yes	Yes	－
Ⅴ沖積平野(液状化有り)	No	No	No	No	No	No	Yes	－
Ⅵ良質地盤	No	－	－	Yes	－	－	－	－

表-3　管種別仮説検証

	被害件数		被害率	
	DIP-A.K.T	CIP	DIP-A.K.T	CIP
Ⅰ改変山地	No	No	No	－
Ⅱ段丘	No	No	No	No
Ⅲ谷・旧水部	No	No	No	No
Ⅳ沖積平野（液状化無し）	－	－	No	Yes
Ⅴ沖積平野（液状化有り）	No	No	No	No
Ⅵ良質地盤	－	－	－	Yes

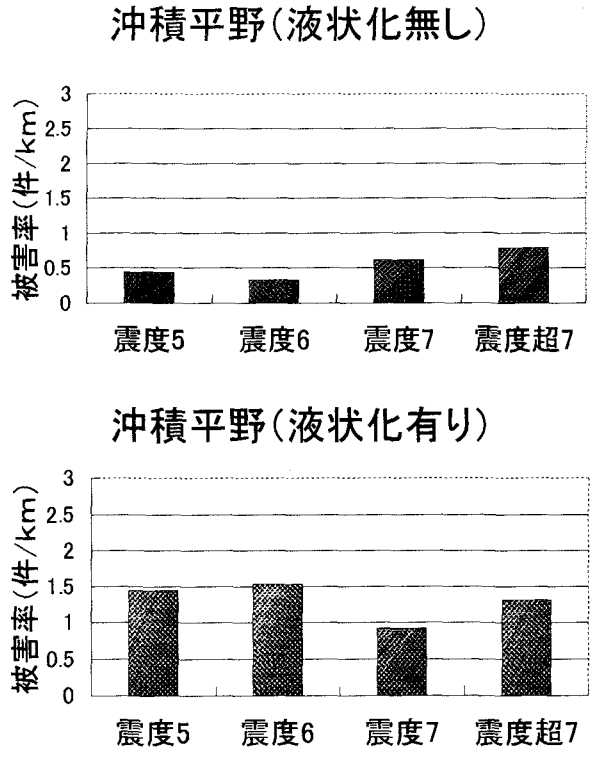

図-3(a)　地盤種類と震度-被害の関係

ている.

加えて,表-2 からは管径が 200～450mmが他の管径よりも被害予測しやすく,さらに表-3 からは管種 CIP が DIP よりも被害の予測しやすい.

次に,さらに DM 結果を詳しく考察するために被害率による図-3(a),(b)を作成した.

ここで図-3 より,改変山地・段丘・谷・旧水部では震度 5 程度から被害が大きいため,地盤状況と震度の関係など詳細な解析が必要となり,同時に震度 5 程度からの対策を考えていくことを示唆している.

また沖積平野では震度 5 程度から液状化の発生が示唆され,加えて全体的に大きな被害が出ているため災害予測も困難な地形であると言える.

6. ま と め

汎用統計ツール（Microsoft Excel）を用いて上水道配水管の被害に影響を与える様々な要因についてデータマイニングを行った結果,仮説「震度が大きくなるに伴って被害も増加する」は管種や管径よりも,地形の種類に大きくかかわってくることが導き出された.よって地形の種類により地震の被害予測・対策を考えていくのが,効率もよく適切ではないかと考えられる.

これにより,データマイニングの長所である専門（地震工学）知識がなくても簡単に解析が行えることが証明されたのではないかと考える.

最後にこの研究は科学研究補助,基盤研究(B)(1),都市ライフラインの地震時性能照査技法の開発およびその応用（代表者:星谷勝）の一部として行ったものである.

参考文献

1) 徳山豪：データマイニングに使われる最適化の数理,応用数理,VOL.6,NO.4,pp303-313,1996.12.
2) Pieter Adrians, Dolf Zantinge 著　山本英子・梅村恭司訳：データマイニング,共立出版,1998
3) 大規模データベースからの知識獲得,人工知能学会誌,Vol.12,No.4,pp496-549,1997.7
4) 中林三平：データマイニング価値ある情報を掘り当てる,NIKKEI COMPUTER,pp142-147, 1996.9.30.
5) 須藤敦史,高須光朗,星谷勝：ニューラルネットワークを用いたデータマイニングによる非構造システムの同定,応用力学論文集,Vol.2,pp.83-90,1999.
6) 須藤敦史,星谷勝：決定木・GA を用いたデータマイニングによる赤潮発生要因の同定,応用力学論文集,Vol.3,pp.99-106,2000.
7) 須藤敦史,佐藤大介,星谷勝：非構造システムの同定・逆解析におけるデータマイニングによる相関ルール抽出について,応用力学論文集,Vol.4,pp.33-40,2001.
8) 須藤敦史,吉尾薫光,三上隆：データマイニングにおけるルール抽出の考察とトンネル覆工の微細ひび割れ発生要因への適用,応用力学論文集,Vol.5,pp.45-52,2002.
9) 地震による水道管路の被害予測,社団法人日本水道協会,1998.

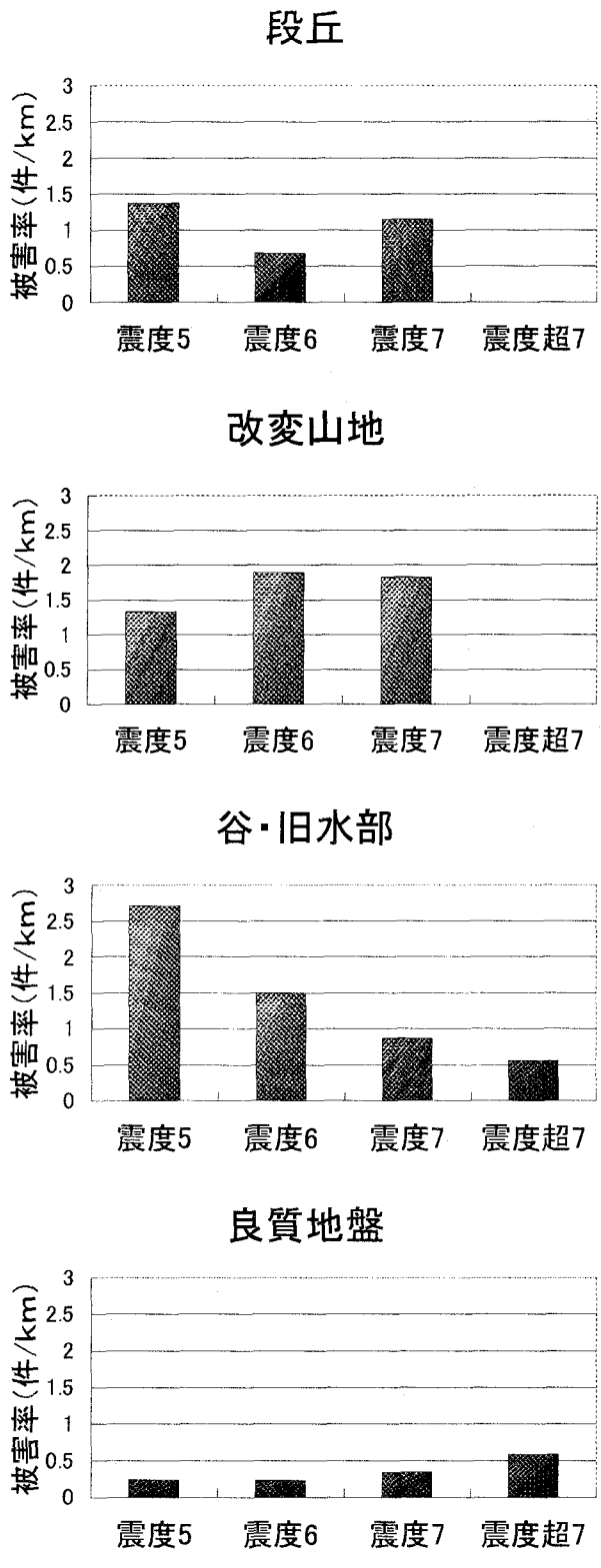

図-3(b)　地盤種類と震度-被害の関係