データマイニングを用いた交通事故分析

A Study on traffic accidents analysis using the data mining methods

鹿野島 秀行*
Hideyuki KANOSHIMA*

1. はじめに

交通事故の発生は人，車両，道路交通環境という３つの要因が複雑に関連して発生していること，また交通事故が非常に稀に発生する事象であり統計的に有効な要因に結びつけることが難しいこと等の理由から，交通事故分析は困難な作業である。しかしそのような困難な中でも交通事故分析はこれまで種種のものが行われており，それぞれ成果を挙げてきている。

ところで従来のマクロ交通事故分析は予め分析者が事故発生要因を想定・仮定し，それをデータを用いて検証してゆく方法が一般的である。この方法は分析効率がよいという長所を持つ一方，分析者の想像できなかった想定・仮定を排除してしまう一面をもっている。

そこでデータから仮説を自動的に抽出する「データマイニング」と呼ばれるツールを利用してマクロ交通事故分析を行ったので，その結果について報告する。

2. データマイニングの概要

（1）データマイニングの定義

データマイニングは近年様々な産業分野，特にマーケティング分野での活用が活発化している。現状では明確な定義・合意はないと見受けられるが，ここでは「大規模なデータの集まりから価値があり，自明でない情報を効果的に発見する手法」と定義する[1]。

（2）特徴

○膨大な量のデータの取り扱いが可能
○データの集合から（半）自動的に発見的なルール（仮説）を導き出せる
　－事故発生要因や事故の背景の組み合わせを客観的に抽出できる
○データ項目が多い場合でも，網羅的な分析が可能
○視覚化による表現（分析結果を視覚化し，理解しやすい表現が可能）

（3）活用方法と交通事故分析への適用

データマイニングを「情報の効果的な発見」と定義すると，交通事故分析への適用例は表１のようにに整理される。

表１　データマイニングの交通事故分析への適用例

場　面	適　用　例
原因と結果の発見	・事故率の高い道路での道路状況のパターンの発見 ・致死率の高い道路での道路状況のパターンの発見 ・事故類型別の道路交通状況の把握
原因・背景の構造発見	・若者が関与している事故の道路状況 ・夜間事故の特性
データの検索	・上記条件の具体的な区間の抽出 ・交通安全対策を実施すべき地点の抽出

3. データマイニングの方法

データマイニングは前処理であるデータ作成，後処理であるレポート出力作業も必要となる。特にデ

キーワード：情報処理，交通安全
*正会員 工修 建設省土木研究所 道路部 交通安全研究室
〒305-0804 茨城県つくば市大字旭１番地
Tel (0298) 64-2211 Fax (0298) 64-0178
E-mail kanosima@pwri.go.jp

ータ作成は分析結果の信頼性を大きく左右する重要な作業で，一連の分析作業の終了後であっても結果如何ではデータ作成に遡って再分析の必要性が生ずる場合もある。

（1）データ作成の流れ

データ作成の流れは，通常以下のステップにより行われる[2]。

① データの選択（data selection)

分析対象となるデータベースより，分析の視点を考慮し必要となるデータの選択を行う。

② データの洗浄（cleaning)

基本的な条件（同じデータが複数存在している等）によりデータをチェックし，誤データを除去する。

③ データの補強（enrichment)

分析の視点より必要と考えられるデータを外部データベース等から追加し，データの拡充を行う。

④ データのコード化（coding)

データのセグメンテーションやコード化を行う。

（2）データマイニングの実行

データマイニングとは単一の技術ではなく，データからデータ以上のものを引き出す技術の集団といってもよく，その機能は相関，クラス分類等様々である[2]。またこれら技術に対応して，アルゴリズムも統計的手法，決定木，NN，GA等が用いられる。

ここでは後に示す適用例に併せて，判定根拠帰納ツール，回帰ツリー帰納ツールの2つについて示す。

① 判定根拠帰納ツール（Evidence Inducer）

複数のデータ項目の相関の高さを導き出す。指標は"Evidence"と呼ばれ，1つの選択された項目の下でのその他の項目の条件付確率から計算される指標であり，具体的には以下の手順で計算される[3]。

・1つの項目 i を選択する。
・項目 i の下での項目 j が生起する条件付確率 Pij を求める。
・下記の式に基づき Evidence を算出し，重要な項目を抽出する。

$$Evidence_{ij} = -\log\left(1 - \frac{P_{ij}}{\sum_i P_{ij}}\right)$$

ここに $Evidence_{ij}$：選択した項目 i かつ j の Evidence

P_{ij}：選択した項目 i かつ j の生起確率

＜Evidence による評価例＞

・事故率の高い地点では，交通量と信号交差点の密度が高い。
・そのうち交通量が×××台／日で最も危険度が高くなる。
・その確率は○○％であり，信頼区間は±○○％である。

② 回帰ツリー帰納ツール（Regression Tree Inducer）による評価

回帰ツリー（Regression Tree）は推定しようとするカテゴリーを最もよく表現する階層構造を発見するものである。その手法は以下の通りである。

・1つの項目を選択する。
・選択した項目を最も分離できる（表現できる）データを見つける。
・連続量であれば，自動的に区分する。
・次の階層も同じことを実施し，それ以上分類できなくなれば止める。

なお連続量を区分する場合の境界値は与えられるデータから計算された値であるから，参考的に用いるべきものであり，普遍的な値ではないことに注意が必要である。

＜Regression tree による評価例＞

・夜間事故は，幅員××m以上で交通量が○○台／日以下で多く発生する。
・若者事故は，夜間での幹線道路に関係が大きい。

（3）レポーティング

分析結果について出力すべき内容の加工を行い，必要に応じて分析結果の表示を行う。

4. データマイニングを用いた交通事故分析

Evidence による評価

ここでは Evidence による評価の一例として事故類型別に交通事故発生と相関の大きい項目を抽出した結果を記す。対象となる道路種別は一般国道，主要地方道，一般都道府県道，政令指定市の一般市道であり，概ね幹線道路として分類できる。なお今回用いたのは警察庁，建設省の交通事故統合データベースである。

単独事故と昼夜の関係を検討する場合を例に説明すると，以下の手順を踏む。

・事故類型毎に，発生頻度に関係する項目を抽出し，条件付き確率を計算する。

例：単独事故 50 件（うち昼間 10 件，夜間 40 件）
　　出合頭 20 件（うち昼間 8 件，夜間 12 件）
の場合，
　単独かつ昼間 p=0.2　単独かつ夜間 p=0.8
　出合頭かつ昼間 p=0.4　出合頭かつ夜間 p=0.6

・事故類型毎に Evidence を計算し，比較評価する。

例：単独事故と夜間の関係
　　Evidence = -log(1-0.8/(0.8+0.6))=0.85
　単独事故と昼間の関係
　　Evidence = -log(1-0.2/(0.2+0.4))=0.41
　同様に出合頭事故と夜間・昼間の Evidence はそれぞれ 0.56，1.10 と計算されるから，単独事故と夜間，出合頭事故と昼間の関連性が高いといえる。

このようにして分析・整理した結果を表 2 に示す。

5. データマイニングを用いた交通事故分析 回帰ツリーによる評価

（1）方法

当事者，事故類型による事故の特性，道路状況を把握するために，回帰ツリーによる評価を行った。分析は事故類型別，センサス区間毎に死傷者数を階層に分け死傷者の多くなる道路状況を算出した。また利用したデータは交通事故統合データベースであるが，分析に気象条件を反映させるために，都道府県毎の降雪日数，最低気温データを付加した。

（2）分析結果

図 1 は追突事故の回帰ツリーを示したものである。追突事故は全センサス区間では平均 21.24 人の死傷者が発生しているが，これを日交通量 20,429 人を境に区分すると，交通量の多いセンサス区間では 91.39 人，少ないセンサス区間では 11.27 人と大きな差が現れる。つまり日交通量の違いは追突事故の発生と相関が高いと言える。日交通量の多いセンサス区間を更に昼夜率 1.51 を境に区分すると，昼夜率の高いセンサス区間では 174.2 人，少ないセンサス区間では 77.47 人と大きな差が現れる。つまり日交通量が多く，昼夜率が高い区間では追突事故による死傷者数が多い区間と言える。ところで日交通量や昼夜率の境界値は選択項目を最もよく分離（表現）できるようにソフトウェアが自動的に計算した結果である。また表 3 はこうして得られた事故類型毎の特徴を整理したものである。

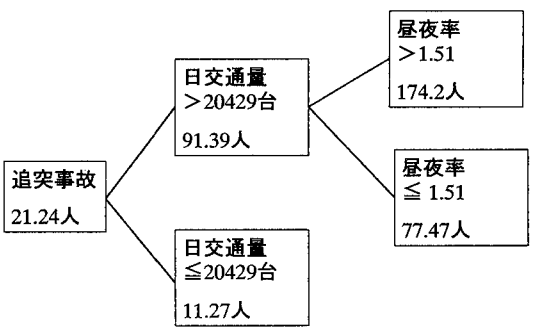

図 1　追突事故の回帰ツリー

6. まとめ

今回はデータマイニングのうち，Evidence による評価，回帰ツリーによる評価を，マクロ交通事故分析に適用した。その結果データの範囲内で網羅的に交通事故の特徴を把握することができた。

ところでこの方法は当然のことながらデータベースに含まれない事故発生要因を探し出すことは不可能である。また結果はあくまで相関関係を表すものであり，因果関係を表すものではないことに注意が必要である。したがってこの結果は万能なものではなく，結果を足がかりにした因果関係の分析が別途必要となる場合があることに留意すべきである。

参考文献

1)Joseph P.Bigus 著，株式会社社会調査研究所，日本 IBM 共訳：「ニューラル・ネットワークによるデータマイニング」，1997

2)Pieter Adriaans, Dolf Zantige 著，山本英子，梅村恭司訳：「データマイニング」，1998

3)日本 SGI：「MineSet2.6 日本語マニュアル」

表2　事故類型と関連性の高い指標（Evidence による評価）

事故類型 / 指標	歩行者通行中	横断歩道横断中	その他横断中(乱横断)	路上作業中	正面衝突	追突	出合頭	左折時	右折時	工作物衝突	路外逸脱	備　考
道路種別	○											○主要地方道，都道府県道との関連高
死傷者数					○	○			○	○	○	○1事故当たりの死傷者数の多さ
季節	冬	秋	春	春	冬	夏	夏	夏	冬	夏	夏・冬	
昼夜	夜			昼		昼	昼	昼		夜	夜	
路面状態	AB	AB		B	B	B					B	A 湿潤　B 積雪凍結
道路線形				A	B	A		A		BC	BCD	A 直線　B 左カーブ C 右カーブ D 直線（下り）
沿道状況	AC	AC		AB	CD	B	AC	A	A	BCD	BCD	A DID B DID 以外の市街地 C 平地　D 山地
信号交差点密度		大		小	小		大	大	大	小	小	
無信号交差点密度	大	大				大						
日交通量	小	中大	中	大	小	大		大	中大	小中	小	
混雑時旅行速度	低	低			高		高	低	低	高	高	
昼夜率	○	○		○	○	○		○				○昼夜率の高い箇所との相関高
大型車混入率				○	○						○	○大型車混入率の高い箇所との相関高
混雑度				低	低	高	高	高	高	低	低	

表3　事故類型毎の特徴（回帰ツリーによる評価）

事故類型	特　徴
歩行者通行中	日交通量 3,317 台超，かつ車道部幅員 7.83m 以下の区間で多く発生
歩行者横断中	歩行者交通量 505 人超，かつ昼夜率 1.5 超の区間で多く発生
その他横断中（乱横断）	日交通量 8,467 台超，かつ昼夜率 1.5 超，かつ歩行者交通量 262 人超の区間で多く発生
正面衝突	日交通量 5,856 台超，かつ年間最低気温－4.5℃以下の区間で多く発生
追突	日交通量 20,429 台超，かつ昼夜率 1.5 超の区間で多く発生
出合頭	日交通量 9,896 台超，かつ昼夜率 1.5 超の区間で多く発生
左折時	昼夜率 1.5 超，かつ二輪車交通量 500 台超の区間で多く発生
右折時	日交通量 20,429 台超，かつ昼夜率 1.5 超の区間で多く発生
工作物衝突	日交通量 5,647 台超，特に 21,920 台超の区間で多く発生
路外逸脱	降雪日数 60 日超，かつ混雑時旅行速度 36km／h 超の区間で多く発生