1E2-1

データマイニングによる障害継続時間予測モデルの検討 – 阪神高速道路における交通データウェハウスの活用 –

Models of Incident Clearance Duration by Practical Data Mining Tools

河野 浩之[*1] Hiroyuki KAWANO　高田 裕之[*2] Hiroyuki TAKADA　石井 康裕[*2] Yasuhiro ISHII　久利 良夫[*3] Yoshio HISARI

[*1]南山大学数理情報学部 Dept. of Info. and Telecom. Eng., Nanzan University

[*2]阪神高速道路株式会社 Hanshin Expressway Company Limited

[*3]阪神高速道路管理技術センター Hanshin Expressway Management Technology Center

The Hanshin Expressway Company Limited has been gathering and providing traffic information since 1970. At present, congestion, incidents and other attributes are integrated in the traffic data warehouse. We analyze 58,738 records stored from 2004 April to next March. By using CfsSubsetEval of mining tool WEKA, we select major attributes, "routes, number of lanes, causes, trouble level, day/night", and remove construction records using EM algorithm. We derive mathematical models of duration of incidents and discuss the applicability of mining algorithms for the traffic data warehouse.

1. はじめに

阪神高速では 2003 年の交通管制システム更新時に，道路交通情報を管理するデータベースシステムの再設計・再構築を行い，車両検知器のデータや情報板表示記録，事故記録等を蓄積するデータベースを一元的に管理するデータウェアハウスを構築している．本稿では，交通データウェアハウス [Al-Deek 02] に一元化され集約されつつあるデータに対するデータマイニング技術 [Shekhar 03] の適用可能性についての検討状況を報告する．2 章では，障害継続時間に対するモデル生成を行うデータマイニングプロセスを述べ，WEKA[Weka 06] を用いた実験結果を 3 章で示し，4 章において検討を行う．5 章「むすび」では，今後の課題についても述べる．

2. 交通データウェアハウスとマイニング技術

阪神高速道路で測定され，データウェアハウスへと蓄積されているデータは多種多様であり，その利用目的も様々である．管制業務に大きな影響を及ぼす障害継続時間の予測モデルを生成するにあたって，抽出データに対するデータクリーニングやデータトランスフォーメーションなどの典型的なデータマイニングプロセスを実行する．

2.1 障害継続時間モデル生成のためのデータ処理

障害継続時間を分析するために，交通データウェアハウスから，次の属性をもつ交通障害日報を抽出する．

対象路線 … 全地区全路線の本線（出入り口を除く）
対象期間 … 2004 年 4 月～2005 年 3 月の 1 年間
属性 … 月, 障害 SQ, 地区, 障害登録番号, 登録日時, 登録時, 解除日, 解除時, 記録日付, 系統, 車線数, 原因, 程度, 場所, 入口出口区分, 復旧見込み, 車種, 車線, 形態, 障害継続時間, 有効データ, 昼夜, 障害継続時間 (分)

連絡先: 〒 489-0863 愛知県瀬戸市せいれい町 南山大学数理情報学部 河野浩之，Tel:0561-89-2000，Fax:0561-89-2082，kawano@it.nanzan-u.ac.jp

データ抽出後,「同一障害番号で複数回の登録がある障害レコードのうち，登録時間間隔が 5 分未満のレコード」の条件によりデータ統合を行った．さらに,「車線数，昼夜区分，障害継続時間」属性をもつ 58,737 レコードを整備した．

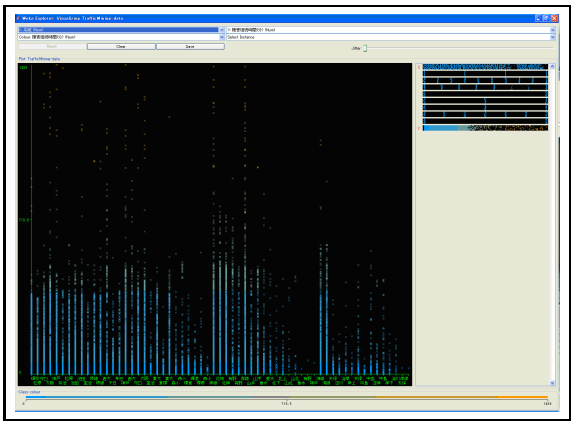

図 1: 各系統の障害継続時間分布（左から順に，環状線，松原線上り，守口線下りなど）

図 1 をはじめとする可視化 (Visualization) を行い，データ分布に偏りのある属性値の有無を確認した．例えば，夜間には「路肩障害，半車障害」に比較して,「1 車障害，2 車障害」が多い．その他，程度属性（1 車障害，路肩障害，半車障害，2 車障害，通行止，3 車障害，特通行止，支障無）に対して，復旧見込み属性（しばらく，まもなく）の「まもなく」の属性値のほぼ全てが,「1 車障害」と記録されている．

WEKA の視覚化機能を援用することで，各属性の属性値が他の属性に及ぼす影響や，2 種の属性間の影響を把握し，初期段階のデータクリーニングを試みた．この際，ランダムサンプリングを行った 3,000 レコードに対して，主要アルゴリズム（分類・数値予測スキーム）を用いることで，ルール (rules)，決定木 (tree)，関数型 (functions)，インスタンスベース (lazy) のモデルを試験的に生成し，次の出力モデル（一部）を得た．

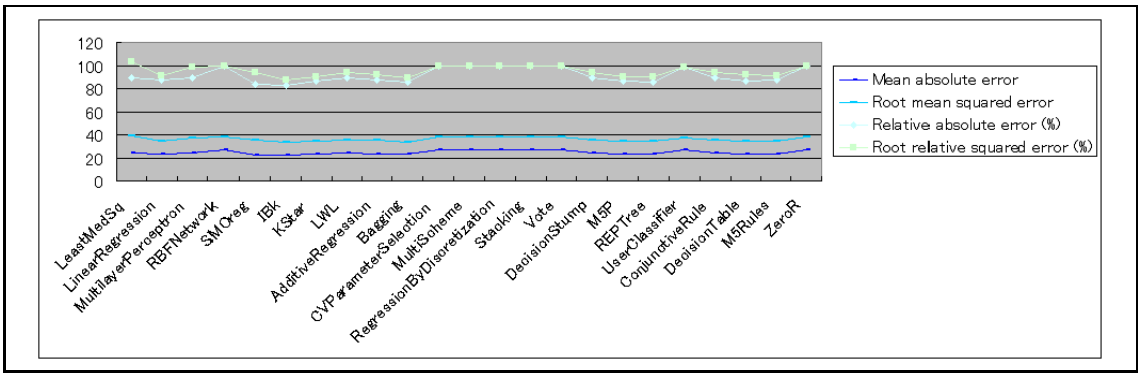

図 2: 「原因 (工事), 昼夜 (夜)」データに対する各アルゴリズムの性能指標

出力モデル:
*Linear Regression Model:
障害継続時間 (分)=
107.6131*原因= “工事, その他”+34.7619

*M5 pruned model rules (using smoothed linear models):
Number of Rules: 1
障害継続時間 (分)=
93.0528*系統= “松原線上り, 池田線下り”+
88.3424*原因= “点検, 工事, その他”+
17.9418 [29/120.204%]

また，抽出データのどの属性を中心に解析するべきかを検討するため,「Select Attributes 機能」を用いて処理した．属性選択手法を複数適用したところ,「原因，昼夜」属性が候補となった．以後,「原因，昼夜」属性の特徴に注意し，モデル生成においても「原因，昼夜」属性値に着目しながらデータ処理を行う．

ところで，今回の調査データに対する障害継続時間モデル生成には，相関ルールに関するアルゴリズムは利用できない．また，数値属性である「障害継続時間」を対象とすることから，文字属性に対するモデル生成手法も適用できない．

さらに，今回の検討を進める上で，その出力モデルの適用領域を考えると，解釈可能なルールや規則を得ることが可能な出力モデルを得ることが望ましい．従って，Quinlan J. R. により提案された M5 を発展させた M5Rules アルゴリズム, M5P アルゴリズム，AIC 基準を用いた Linear Regression Model アルゴリズムなどによるモデル出力に注意を払う．

2.2 属性分割データに基づくスキーム選択

前節のサンプルデータに対する実験結果を補強するため，全データの属性「系統，車線数，原因，程度，復旧見込み，車種，車線，形態，昼夜，障害継続時間 (分)」に対し，CfsSubsetEval を実行した．その結果，障害継続時間 (分) に対する主要要因は,「車線数，原因，復旧見込み，昼夜」となった．

また，業務判断により障害処理中（処理後）に入力される詳細属性「復旧見込み，車種，車線，形態」を除いたモデル生成が必要であると考えられることから，初期入力データとして「系統，車線数，原因，程度，昼夜，障害継続時間 (分)」属性に対して同様の属性選択を実行し,「車線数，原因，昼夜」の属性を得た.「車線数，原因，昼夜」属性は，次の属性値からなる．

車線数 ··· 1,2,3,4
原因 ···· その他, 火災, 緊急工事, 故障, 工事, 事故, 清掃, 点検, 落下物, 脇見
昼夜 ···· 昼, 夜

ここで,「原因」属性の特性を考慮し，円滑な交通のかく乱要因となりやすい「故障（火災・故障・事故・脇見)」，計画的に実施される可能性の高い「工事（工事・緊急工事)」，一時的な要因が多い「その他（その他，清掃，点検作，落下物)」により分類する．そして，データ処理時間を短縮するために,「原因（故障，工事，その他)，昼夜（昼，夜)」に含まれる諸属性値によりデータを 6 分割し,「系統，車線数，原因，程度，昼夜，障害継続時間 (分)」属性をもつデータとして処理した．ただし，故障に分類されるデータには，詳細属性「復旧見込み，車種，車線，形態」が付与されることが多いため,「系統，車線数，原因，程度，復旧見込み，車種，車線，形態，昼夜，障害継続時間 (分)」属性をもつデータに対する処理も行う．

また，前節の議論に基づいた主要アルゴリズムを中心に，各アルゴリズムによる出力モデルの性能を比較する．その際，6 分割データと 2 詳細属性データに対してモデル生成を行い，小さな relative absolute error を与えるアルゴリズムを中心に選択する．図 2 は,「原因（工事)，昼夜（夜)」データに対する各アルゴリズムの性能指標をグラフにしたものである．

8 種のデータに対する実験に基づいて,「SMOreg，IBK，Bagging，DecisionTable」が比較的好ましい性質をもつことが分かった．そこで，業務上利用しやすい性質をもつ「M5P, M5Rules，LinearRegression」を中心に,「SMOreg，M5P, M5Rules，LinearRegression」を用いた実験を進める．

3. 主要アルゴリズムによるモデル生成

前章の一連の実験から，障害継続時間の予測値のエラーは，500 分を超過する障害継続時間データの影響を非常に大きく受けることが分かった．さらに，著しい大きな外れ値であるため，事後的に入力される「詳細属性（復旧見込み，車種，車線，形態)」データによっても，その予測モデルの精度を大きく改善することができない．さらに，M5Rules は 13 ルールを出力し，M5P は 26 クラスターに分割する．これらの結果を踏ま

表 1: 各アルゴリズムによるモデル評価

Algorithms	LinearRegression	M5P (26 分割)	M5Rules（19 ルール）	SMOreg
Mean absolute error	17.394	16.6133	16.594	16.7252
Relative absolute error (%)	87.0141	83.1084	83.0144	83.6683

えて，以下の方針によるデータクリーニングを行った後に，モデル生成を行うことにした．

1. 「工事」属性値の除外
   360 分を超える長時間の障害継続を引き起こす主要な原因は,「工事」である．また，一般的に計画的に実施される「工事」が多く，工事種別や規制方法など，今回利用できなかったデータを加え，障害継続時間を工事時間から予測することが望ましい．
2. 障害継続時間の上下限の設定
   非常に短い障害継続時間をもつデータが多数存在するが，管制員は障害登録時点で数分内の解除を想定している場合が現実には多い．よって，5 分未満の障害継続時間を除外する．また，阪神高速道路におけるトリップ完結時間である 180 分以上も除外する．

これらのクリーニング後のデータに対して実験を行ったところ，Linear Regression よりも，M5P，M5Rules の方が多少好ましいモデルを生成する結果が得られた (表 1)．ただし，相対誤差が 80%を超えていることから，より精度の高い予測モデルを生成する必要がある．なお，各アルゴリズムにより生成されたモデルは非常に複雑であるので，以下は，M5Rules の一部分のみを示す．

障害継続時間 (分)=
-7.7721*系統=“湾線下り, 池田線下り, 神戸線上り, 湾線上り, 神戸線下り, 西大阪線上り, 山手線下り, 池田線上り, 守口線上り, 淀 川左岸線下り, 松原線下り, 環東渡り, 松原線上り, 守口線下り, 堺線下 り, 北神戸線下り, 堺線上り, 蛍池線下り, 環堺渡り, 北下山渡り, 山北上 渡り, 山手線上り, 北神戸線上り, 中島線上り, 垂水線上り, 淀岸下渡り, 淀岸上渡り, 神戸山手線, 北上山渡り, 蛍池線上り, 有野線下り, 有野線, 淀川左岸線上り”+
14.8144*原因=“脇見, 事故, 火災, 点検”+
31.5147 [389/130.421%]

ここで，より精度の高いモデル生成を行うために，データクリーニング後のデータの特徴を再度調べることにし，クラスタリングアルゴリズム（EM）を適用した．その結果，図 3 に示すような「系統，車線数」に強く影響される複数のクラスターが存在する．

・路線系統，特に環状線のクラスター（Cluster 3）は固有の性質をもつ
・車線数が 4 車線も固有の性質をもつ（環状線と同義）
・障害継続時間 100 分以下に，複数クラスターが存在する
・50 分を超えて，1 クラスター（Cluster 4）が存在する

したがって，他の系統と異なる障害継続時間の特性をもつ環状線に対して，別途モデル生成を考慮する必要があると思われる．現時点で，十分な分析を行うことはできていないが，車線数が 4 車線であること，交通量が多く障害の影響が大きいこと，交通流が異なること，もしくは，障害排除に時間を要していること等に起因するものと予想している．

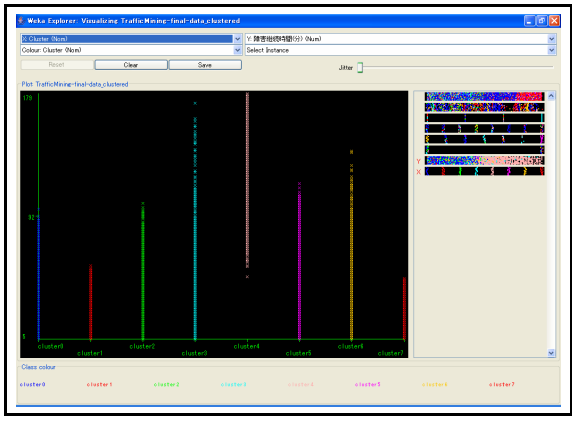

図 3: データクリーニング後のデータに対するクラスタリング結果

4. 生成モデルに関する考察

前章までにおいて，データウェアハウスに蓄積された「障害日報」データに対して，適用対象データとマイニング手法を変化させつつ複数のデータクリーニングなどを行なったデータを利用し，障害継続時間予測モデルの生成を試みた．そして，図 4 に示すような M5Rules などによる路線系統別障害継続時間の予測モデルを得ることができた．

そして，M5P もしくは M5Rules によるルール生成や，データ分類を伴う線形回帰モデルの利用により，他のアルゴリズムに比較して誤差を抑える可能性があることが明らかになった．また，生成モデルの RMS 誤差が 20 分強であることから，30 分単位程度の予測精度の情報提供等では利用できる可能性がある．よって，本稿における実験は，交通管理業務における障害継続時間予測に利用可能な記述モデルを与えるワンステップとなった．

なお，実験結果に基づいて，データと障害継続時間の関係に対して，以下の性質を検証する必要があると考えている．

- 「工事」は，計画的に実施されること，今回対象とした属性以外の情報も十分入手可能であることから，別途詳細な分析を行う必要がある．実際，工事には多くの属性やデータや計画要因，工事運用情報などが付随しており，運用記録情報なども有力な説明要素になる可能性がある．
- 180 分以上の長い継続時間，5 分未満のみ時間継続時間の障害については，交通管制業務や利用目的などの運用面での情報を分析する必要がある．
- 少なくとも環状線は，独立したモデルとして分析する必要があると考えられる．

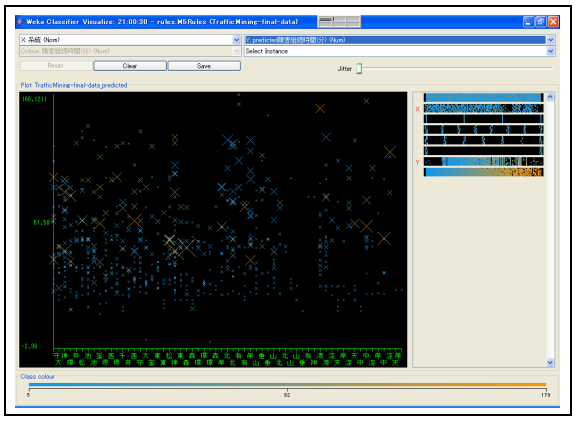

図 4: M5Rules による路線系統別障害継続時間の予測値

- クラスタリングの結果において，100 分程度以下の障害継続時間に対して多くのクラスタが存在すること，クラスタとデータ属性が一部を除いて明確に対応していないことなどから，他にも付加することが可能な気象などの属性情報を含めた分析が必要である．

ところで，障害継続時間に対する要因については，今回のクラスタリングの結果からも分かるように，データとして明示的に記録されていない（あるいは記録不可能な）要因が影響している可能性がある．もっとも，この種の発見をデータマイニングにより行うことは困難と思われる．

なお，交通データに対するデータマイニングは，合衆国の道路交通関連機関でも検討しており，分析テーマと検討結果は，本研究を進める上でも参考にした [Conerly 00, Coufal 03, Shekhar 03, Turochy 02]．特に，先行事例である研究 [Smith 02] において，精度の高い障害継続時間の推定モデルは相当程度に困難であることが報告されている点に注意した．

もっとも，データマイニングにより得られた結果を詳細に分析し，精度を高める以外の方針も重要である．すなわち，利用目的を勘案しながら，生成されたモデルを具体的に運用する方法について検討を進めるべきである．実際の運用方法としては，予測情報を提供する相手，提供のタイミング，提供媒体の考慮の他にも，情報提示の工夫の検討が必要である．特に，今回のように誤差が大きい予測値を提供する場合には，予測値を特定の値でなく分布で与えるなどが有効と思われる．

5. むすび

阪神高速道路では，交通データウェアハウスの有効活用に関する検討を行っており，現状の交通管制データベースシステムを拡張する方向性のひとつに，データマイニング機能の導入の検討がある．本稿では，適用可能性に関する課題整理のために，基本的なデータマイニングプロセスに従って処理を試みた．

その結果，データ準備段階で，交通データウェアハウスから抽出したデータに対して必要な属性を外生で加える必要が生じた．分析効率を向上させるには，より適切なスキーマ設計（スノーフレークスキーマの採用など）に取り組む必要がある．また，今後，障害継続時間と密接に関係する可能性のあるデータをさらに整備すると共に，繰り返し分析を行う必要があることが明らかになった．

なお，今回の管制業務における交通障害記録データから障害継続時間を推定するモデル構築のプロセスにおいて，利用データの特性や分析目的に合わせて，数多くのアルゴリズムから適切な手法を選択する部分は，日常業務に導入する上で大きな障害となる．しかし，データマイニングツール WEKA を用いることによって，比較的短期間のうちに，数多くのデータマイニングアルゴリズムを適用し，適合性に関する見通しを得られるメリットは実務への導入の実施を検討する上で大きい．また，論文中には記載しなかったが，SPSS Clementine を用いた予備実験も同時に行っており，適切なデータマイニングツールの導入が重要である．

今後，交通管制におけるデータマイニング適用について，データ拡充後に同テーマの再検討，もしくは，異なる分析テーマを選定し，この種の経験の蓄積を行いたい．また，渋滞や所要時間などを予測する場合は，現在採用しているアルゴリズムより優れた予測精度があるかどうかの検証を進める必要があると考えている．

参考文献

[Al-Deek 02] Al-Deek, H. and Abd-Elarhman, A., “An Evaluation Plan for the Conceptual Design of the Florida Transportation Data Warehouse, Phase-1-, Final Report,” Center for Advanced Transportaiton Systems Simulation, Department of Civil & Environment Engineering, University of Central Florida, 2002.

[Conerly 00] Conerly, M., Gray, B., Busby, K. and Mansfield, E., “Data Mining and Visualization of Alabama Accident Database,” University Transportation Center for Alabama, 2000.

[Coufal 03] Coufal, D. and Turunen, E., “Short Term Prediction of Highway Travel Time Using Data Mining and Neuro-Fuzzy Methods,” Proc. of the 10th IFSA World Congress, pp.175-186, 2003.

[Shekhar 03] Shekhar, S. and Chawla, S., “Spatial Database: A Tour,” Prentice Hall, 2003.

[Shekhar 03] Shekhar, S., Lu, C. T., Chawla, S. and Zhang, P. “Data Mining and Visualization of Twin-Cities Traffic Data,” Technical Report TR 01-015, Dept. of CSE, Univ. of Minnesota, 2001.

[Smith 02] Smith, K. and Smith, B. L., “Forecasting the Clearance Time of Freeway Accidents,” National ITS Implementation Research Center, 2002.

[Turochy 02] Turochy, R. E. and Smith, B. L., “Alternative Approaches to Condition Monitoring in Freeway Management Systems, Final Report,” Virginia Transportation Research Council, 2002.

[Weka 06] Weka,“ Weka 3: Data Mining Software in Java, ” (http://www.cs.waikato.ac.nz/ml/weka/, 2006 年 4 月 16 日参照)