1H3-1

教師あり機械学習による工事データからの道路開通予測

Predicting Road Openings from Construction Data by Supervised Machine Learning

小林亘[*1] Wataru Kobayashi
柴崎亮介[*2] Ryosuke Shibasaki
関本義秀[*3] Yoshihide Sekimoto

[*1] 東京電機大学 Tokyo Denki University
[*2] 東京大学 Tokyo University
[*3] 東京大学 Tokyo University

Road maps and road navigation data require to be updated as soon as new road is available. Information on road openings of all road administrators should be collected to reflect the changes in the maps, though it is not arranged. This paper supplies experimental results of predicting location and time of road openings from construction data by supervised machine learning. Variables created by methods of construction in a period predicted road openings per location. Some of the True Positive Rate by machine learning showed over 70% even for unknown test data, but the result was unstable compared to prediction by the numbers of construction contracts.

1. はじめに

新たな道路に対応して速やかに地図やカーナビデータの更新がなされるためには，全国 1,700 以上の地方自治体等の道路管理者から道路の開通に関する情報（道路開通情報）を収集することが必要である．しかし，道路開通情報を公表している道路管理者は少なく，ゆえに情報の収集に多大な労力が費やされている．他方，道路を建設する公共工事においては，公正な契約や品質の確保を目的として，工事に関するデータの管理や公表が進められている．本研究では，道路開通の情報収集を効率化するために工事データから道路開通の予測を試みた．

2. 道路開通情報と工事データ

2.1 道路開通情報

道路法は，道路の区域の決定や供用の開始等の場合には，その旨を公示しなければならないと定めている．このため，規模の大小に依らず道路に関する事項が公示には含まれ，しかもそれらは収集・処理しやすい形態で提供されるとは限らないため，公示から地図更新に必要な橋梁やトンネルの開通に関する情報を収集することは簡単ではない．大きな道路の変化を道路開通情報として公表している自治体は一部に留まっている．

2.2 工事データ

道路を建設する公共工事では，公正な契約，品質の確保などを目的として，発注見通し，入札などの幾つかの段階で工事に関する情報が公表されている．工事発注見通しから道路更新情報の自動抽出の試みが行われている[関本 12]．これは工事データを1件ずつ取得し処理する試みである．本研究は，場所毎に累積される複数の工事データを対象としている．複数の工事データを対象とする手法 [小林 12]では，説明変数を有り，無しの 2 値の名義尺度で処理していたが，本研究では説明変数の出現回数という離散的な比例尺度を用いて予測を行う．

3. 予測方法

3.1 場所と時期

道路開通情報では開通場所が大字・町丁目のレベルで公表されており，道路工事における工事場所も同様である．このため，開通と工事の場所は，大字・町丁目のレベルで管理し照合する．ただし，工事が複数の大字・町丁目に及ぶ場合には，工事によってはその一部だけがデータに登録される場合があり，場所毎の工事データの累積に漏れが生ずる恐れがある．それを防ぐため，起点と終点が同一工事で出現し，それらが3キロメートル以内であったら，これらはひとまとまりの場所として工事が進められていると判断し，場所クラスタとしてまとめる．そして，他の工事で場所クラスタの一部だけが出現しても場所クラスタの工事として扱うことにより，データの累積の漏れを防ぐ．

道路開通日，工事の完了日はそれぞれ日のレベルで扱われている．しかし，開通日直近の工事完了日，道路法上の供用日，開通式の日時等は必ずしも日単位で一致していないため，開通と工事の時期は，月あるいは年度のレベルで管理し照合する．

場所，時期のいずれも，上に述べた精度では直ちに地図データとすることはできず，この情報をもとに詳細な調査が必要である．同様な調査対象箇所のリストアップ作業は，現在，道路管理者への質問，様々な WEB サイト，地方新聞などを利用して行われている．本手法は，これらとは異なる情報源を用いて，この作業を支援，補完，照査することができるため有用である．

3.2 予測の評価

工事のデータは，実際の工事の進行に伴って場所ごとに累積され，予測にはこれを用いる．したがって，開通の有無は，場所については大字・町丁目あるいはそれらが集まった場所クラスタの単位で予測されることとなる．開通が予測された場所の真偽は道路開通情報の場所と照合され，一致した場合は場所について「真」であると判定される．開通区間は数百メートルから数キロメートルであり，複数の場所を横断するものがある．詳細調査の対象箇所のリストアップにはその一部があれば良いので，区間単位で真偽を整理する場合には区間に含まれる場所の一部を開通と予測していれば「真」と判定する．時期について，後述の工事件数によって予測する方法では年度単位で予測を行ったこと，そして，正確な開通時期は詳細な調査で明らかにす

連絡先[*1]：東京電機大学研究推進社会連携センター，wkoba@mail.dendai.ac.jp
連絡先[*2]：東京大学空間情報科学研究センター
連絡先[*3]：東京大学生産技術研究所人間・社会系部門

るとして，予測した年度と道路開通情報の年度が一致したものを時期について「真」であると判定する．

開通（陽性）と予測されたデータの場所と時期の両方について「真」であるものを「真陽性（TP）」，開通と予測されたがいずれかあるいは両方が偽であるものを「偽陽性（FP）」，開通では無い（陰性）と予測され，実際には開通であったものを「偽陰性（FN）」，開通では無かったものを「真陰性（TN）」とする．予測結果の評価には開通区間の真陽性率（TPR=TP/(TP+FN)）と場所の偽陽性率（FPR=FP/(FP+TN)）を用いる．TPR は実際の開通区間のうち時期（年度）と少なくとも開通区間の一部の場所が正しく予測された割合を表し，1であれば全ての開通区間が予測されたこととなる．FPR は開通では無い場所と年度に対して誤って開通と予測された割合であり，0に近いほど不要な対象がリストアップされず効率的であったこととなる．

3.3 工事件数による予測

道路が開通される場合には他の場所に比べて多数の工事が行われると考えて，それぞれの場所での各年度に実施された工事件数を閾値として一定以上を開通と予測する方法である．

3.4 工事の特徴を用いた機械学習による予測

工事データの工法・型式が工事の特徴を表すと考え，これによって道路開通を予測する方法である．工法・型式は，例えば，アスファルト舗装工，コンクリート擁壁工など 137 種類が設定されている[日本建設情報総合センター 12]が，説明変数には 20 種類に圧縮したものを用いた．データレコードは場所毎に一定の期間毎に作成する．一定の期間を年度としたものは，場所毎に年度毎の説明変数を集めたものをレコードとし，N箇月としたものは，工事が観測された月毎にその月を含む過去 N 箇月間の説明変数を集めたものをレコードとする．教師データのうち正例は，開通の場所で道路開通月を含むレコードとし，負例は全ての場所における正例以外の同じ期間の長さのレコードとした．例えば，N＝5 の場合の正例は，開通月を含み開通月までの 5 箇月間，すなわち，4 箇月前から開通月までの説明変数を集めたものである．なお，正例を抽出する開通の場所は，開通でない場所よりも少ないため，教師データを構成する際には正例をオーバーサンプリング[He 09]し，負例に均衡させた．

学習モデルにより教師データと期間の長さの等しいテストデータに対して開通の予測を行う．一つの年度に複数のレコードがあったなら，その都度，予測を行い，1度でも開通と予測されたならその場所のその年度は開通と予測する．

表 1. 場所と期間毎に生成されるデータレコードの例

説明変数	出現回数	正規化
基礎工事	5	0.72
土工事	2	0.29
橋梁下部	0	0
橋梁上部	0	0
路盤路床	1	0.14
アスファルト舗装	1	0.14
隧道	0	0
セメント舗装	0	0
歩道	0	0
緑化工事	3	0.43
道路付属物	1	0.14
総合土木工事	2	0.29
専門土木工事	0	0
斜面法面	1	0.14
水路管路	1	0.14
防水工事	0	0
通信管路	0	0
撤去	1	0.14
塗装工事	0	0
維持保守	0	0
CLASS	開通	開通

場所と期間毎に生成されるデータレコードの例を表 1 に示す．期間内の説明変数の出現回数を合計したものに加えて，特徴ベクトルの大きさを等しく1と正規化したデータを準備した．説明変数の出現回数は，開通区間の規模や工区の数等によってばらつくため，回数の影響を減らして説明変数の割合を表す正規化ベクトルが予測精度に与える影響を確かめるためである．

4. 実験

4.1 実験に用いたデータ

表2に示す埼玉県，愛媛県，新潟市の道路開通情報[埼玉県 12，愛媛県 12，新潟市 12]と工事データを用いて実験を行った．それらの説明変数の出現状況を図 1から図3に示す．

表 2. 実験に用いたデータ（2006-2010 年度）

自治体	開通数（5 年度分合計）	開通以外の場所数（各年度計．延べ）	工事件数（500 万円以上で場所特定できたもの）
埼玉県	16 区間（30 場所）	2,422	4,698
愛媛県	11 区間（12 場所）	1,488	3,790
新潟市	14 区間（29 場所）	1,036	2,305

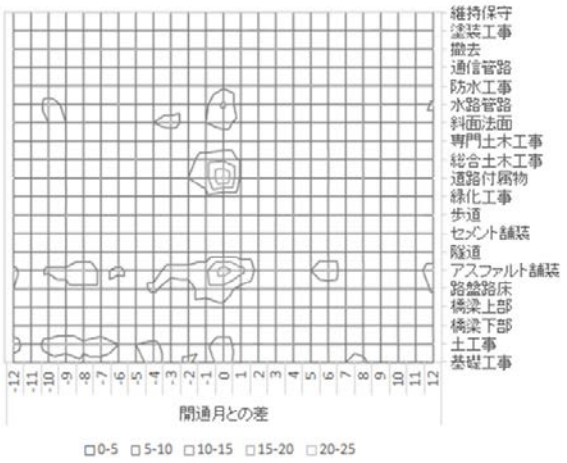

図 1. 説明変数の出現状況：埼玉県

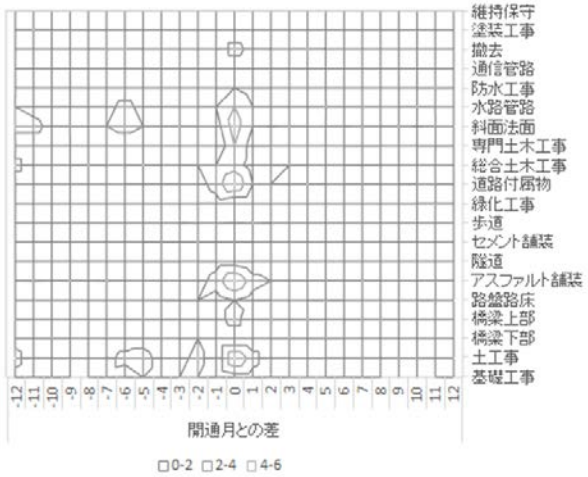

図 2. 説明変数の出現状況：愛媛県

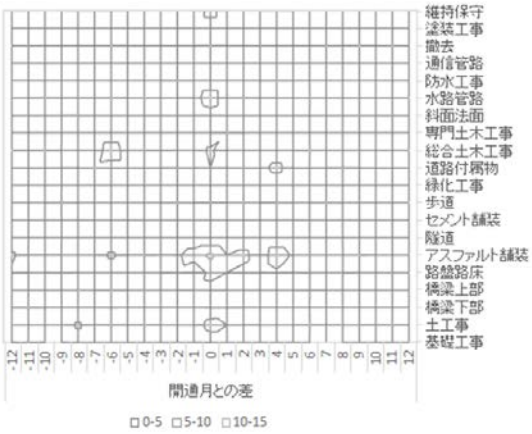

図 3. 説明変数の出現状況：新潟市

4.2 工事件数による予測結果の評価

3 自治体の各年度に行われた工事件数を閾値として開通を予測した場合の評価を図 4 に示す.

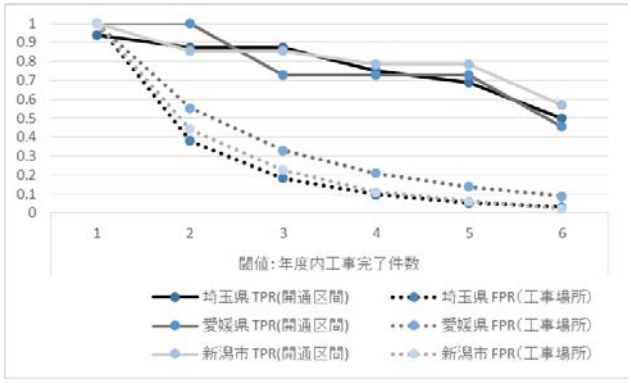

図 4. 年度内工事完了件数を閾値とする予測の評価

4.3 工法・型式を用いた予測結果の評価

教師あり機械学習による予測は表3に示す 5 種類のアルゴリズム[Quinlan 93, Breiman 01, John 95, Platt 98]を用いてその適合性を見ることとした. データを収集する期間は, 完了件数を収集したのと同じ年度単位, そして, データ収集期間の長さとの関係を調べるため 2 箇月から 12 箇月までの 1 箇月毎に変化させたものを準備した. 前述のようにこの期間内にそれぞれの場所で得られた説明変数の数をそのまま numeric 型の変数として用いたものとデータの大きさを正規化したものがある.

表 3. アルゴリズムと主なパラメータ

アルゴリズム名	パラメータ
決定木(C4. 5)	Confidence factor = 0.25
ランダムフォレスト(RandomForest)	木の数= 10 使用した説明変数= 5 木の深さ= 3
ナイーブベイズ分類器(NaiveBayes)	加算スムージング= 0.1
サポートベクターマシン(SVM-Poly)	線形カーネル: C=10
サポートベクターマシン(SVM-RBF)	RBF カーネル: C=50

評価結果を図 5 から図 9 に示す. これらは, 学習モデルを作成した自治体とは異なる自治体(いわゆる未知)のテストデータから得た結果を平均したものであり, 教師データと同じテストデータについて予測したものは含んでいない.

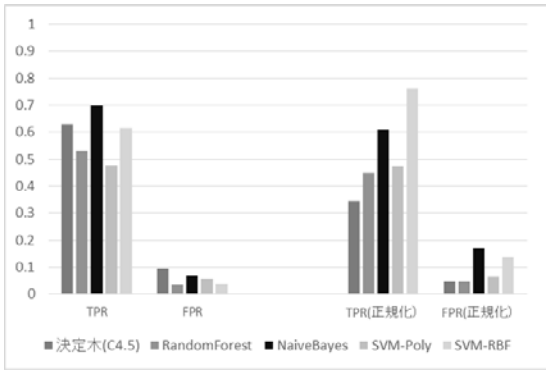

図 5. 年度単位の未知のテストデータに対する評価

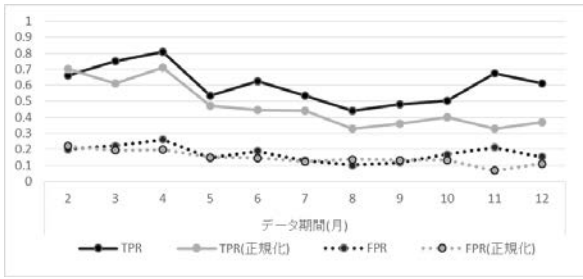

図 6.データ期間・正規化に対する評価:決定木(C4.5)

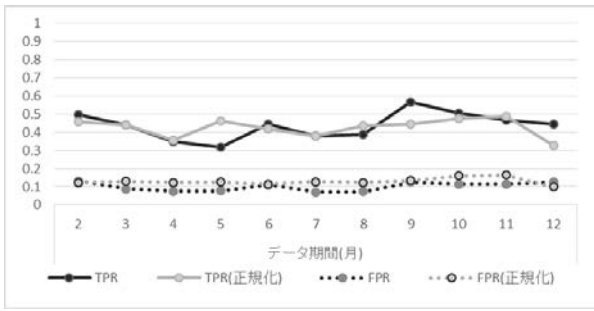

図 7.データ期間・正規化に対する評価:ランダムフォレスト

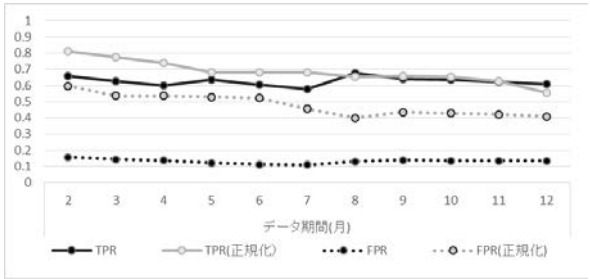

図 8.データ期間・正規化に対する評価:ナイーブベイズ分類器

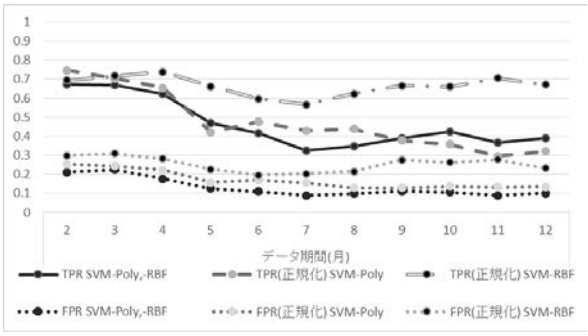

図 9.データ期間・正規化に対する評価:サポートベクターマシン

5. 考察

実験に選んだ自治体は，いずれも道路開通情報を公表しているものであるが，首都圏との距離，都道府県と政令市，規模，自然条件が異なり，その説明変数の出現状況も図 1 から 3 に示されるように異なり，厳しい条件となっている．

工事件数のみを用いて予測した結果を示した図 4 において，閾値を 1 としたものは，年度内に 1 度でも工事のあった場所全てを開通と予測するものである．その TPR は1であったため，全ての開通区間を予測できているが，FPR も 1 となるため，表 2 に示した開通以外の場所の数にあたる詳細調査箇所がリストアップされることとなる．FPR を低下させるために閾値を上昇させると TPR はそのトレードオフで低下し，閾値を 4，つまり，年度内工事回数 4 以上の場所だけを開通と予測した場合には，TPR が 0.7～0.8，FPR で 0.1～0.2 となる．FPR と表 2 から，これは開通場所でないものの予測数が年間数十箇所であることとなる．この傾向は図1に示されるようにどの道路管理者においてもほぼ一致しており安定していた．

次に年度単位で説明変数を収集して機械学習により予測を行った図 5 を見ると，①アルゴリズムによってその評価結果は数十％異なっていた．②工事件数による予測に匹敵する評価となったものがある一方それに及ばないものもあった．③出現回数によるデータではナイーブベイズ分類器が良い成績であったのに対して正規化したものではサポートベクターマシン RBF カーネルが良い成績であった．道路開通時期は必ずしも年度末とは限らないので，道路開通後に完了した工事データも教師データ，テストデータには含まれる場合がある．図1～3 のように開通後であっても当該場所で工事が行われていることは少なくなく，これらを含めて学習し，予測した結果と言える．

図 6から図 9は，データを収集する期間を変えて同様の実験を行ったものである．濃い実線は出現回数を用いた場合の TPR であり，濃い点線はそのときの FPR である．薄い実線，点線はそれぞれ正規化したデータによる TPR，FPR である．決定木及びそのメタ学習モデルであるランダムフォレストによる結果を示した図 6 と 7 では時期による変化が大きいことが示されている．決定木で分割統治する際には多様性を減ずる一つの説明変数が順に選ばれるため，説明変数の出現状況のばらつきにより結果が大きな影響を受けたためと考えられる．図 8 からナイーブベイズ分類器は出現回数を用いた場合にはデータ収集期間の長さの影響を他のアルゴリズムほど受けていないことが示された．正規化したデータを用いた場合には期間が短いほど高い FPR を示した．ナイーブベイズ分類器においては出現回数がゼロである説明変数を処理するために加算スムーズジングを行っており等しく 0.1 を加算している．正規化したデータにおいては一つ一つのデータの数値が小さく，しかも期間が短いほどデータ量が少ないためにこの値の影響を受けてこのような結果を示したものと考えられる．サポートベクターマシンによる結果を示した図 9 では正規化の有無，使用したカーネルによって異なる傾向が示された．

工事件数から単純に予測した方法が比較的安定した成績であったのに対して，工法・型式と機械学習による方法では，工事データを利用する期間，データの正規化の有無，教師つき機械学習アルゴリズムを変えて，異なる道路管理者の未知のテストデータに対して実験を行った結果，ばらつきが大きかった．その中では正規化しないデータに対してナイーブベイズ分類器を適用したものがデータの期間を変えても他の方法に比べて安定していた．

6. むすび

工事データによって開通の場所と時期を予測できれば，道路管理者，地図製作者の負担を増やさずに地図が効率的かつ速やかに更新され道路利用者の利便性を高めることができる．工事件数，工法・型式によって 3 つの自治体に対して道路開通を予測し，いずれの方法でもある程度の割合で開通を予測することができた．今後，精度の向上のためにより小規模の道路管理者のデータを用いた実験を行いたいと考えている．（本研究はデジタル道路地図協会からの研究助成による支援を受けた．）

参考文献

[Breiman 01] Breiman, L.: Random forests, Machine learning, 45(1), 5-32 (2001)

[He 09] He, H., & Garcia, E. A.: Learning from imbalanced data, Knowledge and Data Engineering, IEEE Transactions on, 21(9), 1263-1284 (2009)

[John 95] George h. John, Pat Langley: Estimating Continuous Distributions in Bayesian classifiers, in Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers (1995)

[Platt 98] Platt, J.: Sequential minimal optimization, A fast algorithm for training support vector machines. ftp://ftp.research.microsoft.com/pub/tr/.../TR-98-14.pdf (1998)

[Quinlan 93] Quinlan, J. R.: C4.5: programs for machine learning (Vol. 1). Morgan Kaufmann (1993)

[愛媛県 12] 愛媛県: 道路開通情報, http://www.pref.ehime.jp/h40400/5744/kaituu/kaituu.html (2012)

[小林 12] 小林亘,柴崎亮介,関本義秀: 工事実績情報を用いた道路供用の予測に関する研究，土木学会論文集 F3(土木情報学), 68(2),150-161 (2012)

[埼玉県 12] 埼玉県: 道路開通情報, http://www.pref.saitama.lg.jp/site/kaitsuzyouhou/, http://www.pref.saitama.lg.jp/site/dousei1006/905-20091224-51.html, http://www.pref.saitama.lg.jp/site/dousei1006/dousei035.html, http://www.pref.saitama.lg.jp/site/dousei1006/dousei059.html (2012)

[関本 12] 関本義秀，中條覚，ほか: 工事発注見通し情報を用いた全国における道路更新情報の自動抽出に向けた試み，土木学会論文集 D3（土木計画学），Vol.68,No.3,117-128 (2012)

[新潟市 12] 新潟市: 道路開通情報, http://www.city.niigata.lg.jp/kurashi/doro/road/doroseibi/dorokaitsuinfo/index.html (2012)

[日本建設情報総合センター 12] 一般財団法人日本建設情報総合センター: コリンズ XML 定義書, http://ct.jacic.or.jp/corporation/know/xml/file/c02x01.xls (2012)