高齢化社会における交通安全対策立案へのSVMの適用可能性

Application of Support Vector Machine to the Traffic Safety Plan in Aging Society

室蘭工業大学	○学生員	菊地健元	(Takeyuki KIKUCHI)
室蘭工業大学	学生員	長谷川裕修	(Hironobu HASEGAWA)
株式会社ドーコン	正　員	有村幹治	(Mikiharu ARIMURA)
室蘭工業大学	フェロー	田村　亨	(Tohru TAMURA)

１．はじめに

我が国の交通事故による死者数は、道路整備状況の改善や車両環境の向上、法制度の充実、ドライバーの交通安全意識の向上などによって昭和 45 年から減少し続けている。

道内での交通事故死亡者数は平成 16 年まで 13 年連続全国ワーストワンであったが、17 年には全国第 4 位にまで改善した。この間、死亡者数は 13 年から 17 年まで 5 年連続で減少した。これは、統計データのある昭和 22 年以降初めてのことである。さらに死者 302 人は昭和 29 年以降最も少ない数字となった。一方、交通事故死亡者発生の原因となる交通事故の発生件数は昭和 52 年から毎年増加している。したがって、交通事故死者数をより一層削減するには、交通事故の発生を抑制することが重要である。

ところで、我が国は高齢者数が他の先進国に比べても急速に増加している。それに比例して、高齢者が運転する車両が原因となる交通事故の発生件数も増加している。今後ますます高齢化が進む中で、道路施策の方向性を変えていかなければならない。

そこで本研究の目的は、高齢者と非高齢者それぞれの死傷率の高い事故を抽出することによって、両者の特徴を分析し、高齢化社会における効果的かつ効率的な交通安全対策の立案への応用可能性を検討することにある。具体的には、パターン認識手法として最近注目を集めているサポートベクターマシン(Support Vector Machine、以下 SVM と記す)を交通事故分析に適用し、高齢者と非高齢者でどのように交通事故要因が変化するかを分析し、評価することである。

２．SVMの概要

SVM は、1960 年代に米国の数理学者である Vapnik 等が考案した Optimal Separating Hyperplane(OSH)を起源としている。この手法では、線形分離可能な場合には高い識別能力を示したが、線形分離不可能な場合に関しては有効な手段ではなかった。1980 年代には、多層パーセプトロンという学習モデルが代表的であった。この手法はこれまで多方面に応用されてきたが、望ましくない局所最適解の選択などいくつかの問題点もあった。しかし、1990 年代になって Vapnik 自身により、カーネル関数を組み合わせた非線形の識別手法モデルができる。これが SVM であり、この拡張によって SVM はパターン認識の能力において、最も優秀な学習モデルの 1 つであるといわれている。

SVM は、線形しきい素子を用いて 2 クラスのパターン識別器を構成する手法である。SVM の概念を説明するために図 1 のような 2 つのデータ群(●のデータと■のデータ)の分類を考える。2 つのデータ群を分離するということは識別関数 $f(x)$を求める問題になるが、図 1 に示すように，分離平面(図 1 の直線)と，2 種類のデータとの間の距離(マージン)ができるだけ大きくなるような識別関数 $f(x)$を求めることを意味する。それによって求められる平面が最適な分離平面である。このとき、データを 2 クラスに完全分離できる場合をハードマージン、一部分離できない場合をソフトマージンという。以下、後者について説明する。

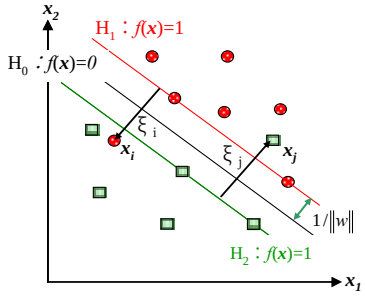

図 1　SVM の概念図

まず、$x_i(i=1\sim\ell)$で表される N 個の成分(入力)と、クラスラベル y_i {-1 あるいは 1} から成るトレーニングデータが与えられているとする。

このとき、ソフトマージン最適化の問題は(1)式で定義される。

$$\begin{aligned} &\min_{w,\xi} imize \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{\ell}\xi_i \ , w \in R^N \ , b \in R \ , \xi \in R^{\ell} \\ &subject \ \ to \quad y_i(w \cdot x_i + b) \geq 1 - \xi_i \ , i=1,\cdots,\ell \\ &\qquad\qquad \xi_i \geq 0 \ , i=1,\cdots,\ell \end{aligned} \tag{1}$$

$\boldsymbol{w}$ は入力に対する重みベクトルである。

この最適化問題に Lagrange 乗数 α_i、および ν_i を導入することで(2)式で表すことができる。

$$\begin{aligned} &\underset{\alpha}{Maximize} \ L_D(\alpha) = \sum_{i=1}^{\ell}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{\ell}\alpha_i\alpha_j y_i y_j x_i^T x_j \\ &subject \ to \ \sum_{i=1}^{\ell}\alpha_i y_i = 0 \\ &\qquad 0 \leq \alpha_i \leq C, \quad i=1,\cdots,\ell \end{aligned} \tag{2}$$

しかし、平面による識別ができる現象は世の中に殆ど存在しない。したがって、より複雑な識別を可能とするために、曲面による分離を考える。

SVM の基本的な構造は, 図 1 に示すような線形しきい素子である。しかし、これでは線形分離不可能なデータに適用することができない。そこで SVM によって非線形な分類を可能にする方法として高次元化が挙げられる(図 2)。これは非線形写像 Φによって、元の入力データを高次元特徴空間に写像し、特徴空間の中でもって線形分離を行う方法である。これにより、元の入力空間においては非線形な分類を行っていることになる。

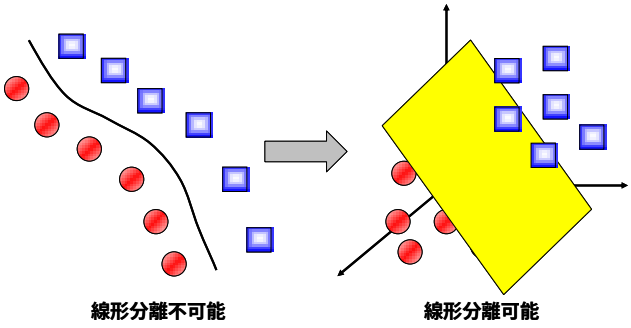

図 2　特徴空間への射像概念

これを実装する際には Φの計算は行わず、カーネル関数の計算に置き換えている(カーネルトリック)。カーネルトリックによって、SVM は Φを直接計算することをせず、計算上の困難を克服している。これを(2)式に適用すると以下の最適化問題が得られる。

$$\max_{\alpha} imize\ W(\alpha) = \sum_{i=1}^{\ell}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{\ell}\alpha_i\alpha_j y_i y_j K(x_i,x_j), \alpha \in R^{\ell} \quad (3)$$

$$subject\ to \sum_{i=1}^{\ell}\alpha_i y_i = 0 \qquad ,0 \le \alpha_i \le C,\ i=1,\cdots,\ell$$

このとき、バイアス b は(4)式で与えられる。

$$b^* = y_i - \sum_{j\in Sv}\alpha_j^* y_j K(x_j, x_i) \quad (4)$$

ここで、Sv はサポートベクターの集合、j は任意のサポートベクターを表す。

結局、識別関数 $f(x)$は(5)式となる。

$$f(x) = w^{*T}x_i + b^* = \sum_{j\in Sv}\alpha_j^* y_j K(x_j, x) + b^* \quad (5)$$

(5)式のバイアス b を固定値とする条件では(3)式の等号条件は消え、問題は 2 次関数の最大化問題となる。最急降下法を用いて次式で α を更新すれば最適解を求めることが出来る。

$$\alpha_i \leftarrow \min\left(C, \max(0, \alpha_i + \eta\frac{\partial W(\alpha)}{\partial \alpha_i})\right) \quad (6)$$

$$\eta_i = \frac{\omega}{K(x_i, x_i)} \qquad (0 \le \omega \le 2) \qquad , i = 1,\cdots,\ell$$

ここで、ω は収束比を表す。この最適化問題の収束判定は、(7)式に示すように主問題と双対問題の目的関数値の比較により行う。

$$\Pr oportion = \frac{\text{主目的関数値} - \text{双対目的関数値}}{\text{主目的関数値} + 1} = \frac{\sum_{i=1}^{\ell}\alpha_i - 2W(\alpha) + C\sum_{i=1}^{\ell}\xi_i}{\sum_{i=1}^{\ell}\alpha_i - W(\alpha) + C\sum_{i=1}^{\ell}\xi_i + 1} \le \varepsilon \quad (7)$$

$$\xi_i = \max\left\{0, 1 - y_i\left(\sum_{j=1}^{\ell}\alpha_j y_j K(x_i, x_j) + b\right)\right\}, (i = 1,\cdots,\ell)$$

3．SVM による事故分析

3．1　分析方法の概要

本研究での事故分析手順を図 3 に示す。まず始めに各データセットの作成を行い、次にそれぞれの SVM のパラメータを設定する。設定したパラメータを基に分析を行い、重大事故を判別して各データセットで比較、検討する。

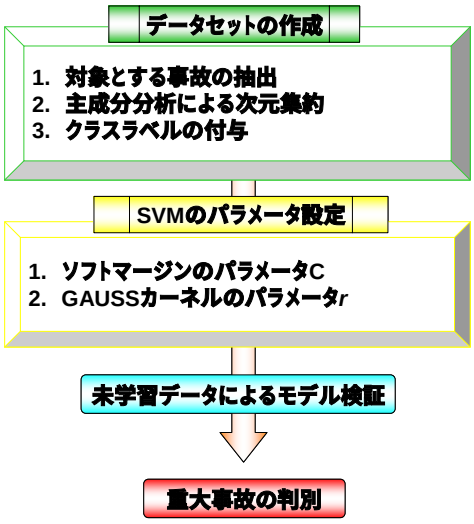

図 3　事故分析フロー

3．2　データの前処理

交通事故データとして(財)交通事故総合分析センターが提供する交通事故統計データ(ITARDA データ)を利用した。ITARDA データでは、事故・行動類型、被害の程度、当事者情報、事故発生地点の道路形状や路面状態、地域特性、沿道状況、などが集計されている。

本研究では、平成 11 年度から 16 年度に北海道渡島・檜山支庁で発生した交通事故約 7500 件の中からデータセット 1 として、

- DID 地区の交差点部で発生
- 加害者年齢 65 歳以上

を条件として抽出した。データセット 2 として、

- DID 地区の交差点部で発生
- 加害者年齢 65 歳未満

を条件として抽出した。これらを判別問題の例題として設定した。各事故に関連付けられた項目すべてが事故発生要因となる訳ではないため、要因となる 10 項目を選択した。さらに主成分分析によって、それらの要因を直感的に把握しやすいように縮約した。分析の結果、両データセットとも第 3 主成分までで累積寄与率が 60%を超えたので、要因数は 3 とした(表 1)。

表１　主成分分析の結果

高齢者 (65歳以上)			
要因および主成分解釈	主成分 1	主成分 2	主成分 3
主成分1:　道路構造			
車線数(車線)	-0.47471	0.146117	0.183716
幅員(車道幅員(m))	-0.51817	0.100027	0.080493
歩道(代表幅員(m))	-0.50831	-0.00066	0.00233
主成分2:　区間特性			
歩道(設置延長(km))	0.241563	-0.55366	0.108051
中央帯設置延長(km)	-0.37465	-0.43906	-0.14536
指定最高速度(km/h)	-0.11692	-0.60548	-0.16111
主成分3:　走行環境			
発生年(月)	-0.01161	0.127266	-0.44388
路面状態	-0.07482	0.018037	0.533179
平日昼夜率	0.072673	0.293575	-0.31511
平日昼間12時間大型車混入率	-0.15609	-0.0052	-0.56604
65歳未満			
要因および主成分解釈	主成分 1	主成分 2	主成分 3
主成分1:　道路構造			
車線数(車線)	-0.45154	-0.2959	-0.11561
幅員(車道幅員(m))	-0.49613	-0.20112	0.000838
歩道(代表幅員(m))	-0.49475	-0.04502	0.050743
主成分2:　区間特性			
歩道(設置延長(km))	0.093779	0.581879	-0.4032
中央帯設置延長(km)	-0.39931	0.420446	0.043211
指定最高速度(km/h)	-0.21333	0.565294	0.137243
主成分3:　走行環境			
発生年(月)	0.016026	-0.0368	0.372706
路面状態	-0.00959	-0.03047	-0.29372
平日昼夜率	0.261822	0.021119	0.509154
平日昼間12時間大型車混入率	-0.15012	0.179777	0.562461

高齢者、非高齢者とも、主成分 1 は車線数、幅員、歩道(代表幅員)に相関関係が見られたため道路構造とした。主成分 2 は、歩道(設置延長)、中央帯設置延長、指定最高速度に相関関係が見られたため区間特性とした。主成分 3 は、発生月、路面状態、平日昼夜率、平日昼間 12 時間大型車混入率に相関関係が見られたため走行環境とした。

識別クラスは両データセットとも死亡・重傷事故と軽傷事故の 2 クラスとした。主成分分析によって得られた主成分得点(x_1, x_2, x_3)を入力 x_i として、各データにクラス分けの指標 y_i\{-1,1\}を付与した(図 4、図 5)。

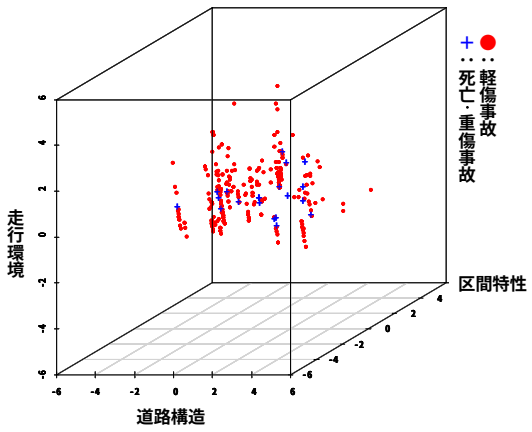

図４　データの分布(高齢者)

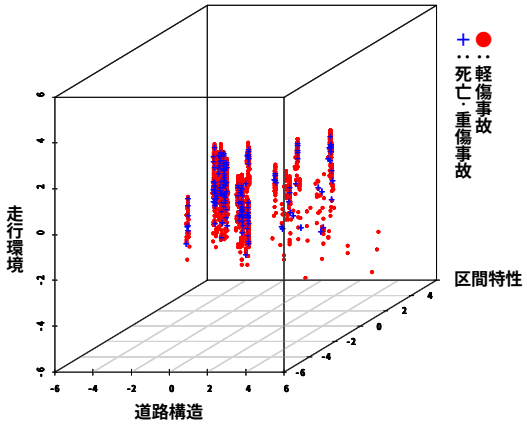

図５　データの分布(非高齢者)

３．３　モデルの構築

前節で作成した各データセットの中から、データセット 1 は y=-1 のデータ 15 個、y=1 のデータ 20 個を抽出しトレーニングデータとした。

データセット 2 は y=-1 のデータ 190 個、y=1 のデータ 360 個を抽出しトレーニングデータとした。

カーネル関数 $K(x_i,x_j)$は(8)式で表される RBF(Gauss)カーネルを用いる。

$$K(x_i, x_j) = \exp\left[\frac{-\left\| x_i - x_j \right\|^2}{2r^2}\right] \quad , \quad \mathrm{i,j} = 1,\cdots,\ell \qquad (8)$$

SVM の識別能力にはソフトマージンのパラメータ C の値と、Gauss カーネルのパラメータ r の値が重要になってくる。これは繰り返し計算によって適切な値を求めなくてはいけない。本研究では、SVM の 1 手法である 10-fold cross validation を用いてソフトマージンのパラメータ C、Gauss カーネルのパラメータ r を決定する。

10-fold cross validation とは、データをランダムに 10 等分割し、その中の 9 セットをトレーニングサンプル、残りの 1 セットを分析用サンプルとし、分析用サンプルの組み合わせを順に替えながら 10 通り行い、予測精度の平均をとる方法である。

試行した中での最も良い結果は、データセット 1 では C=60、r=5、データセット 2 では C=40、r=40、であったのでこれらの値を用いた。

３．４　未学習データによるモデルの検証

データセット 1 では、トレーニングデータを含めた 257 データを分析用データとし、データセット 2 では、トレーニングデータを含めた 2391 データを分析用データとした。作成した SVM を用いてそれぞれを識別した。

未学習データに対する SVM の判別率は、データセット 1 では全 257 データ中 79%の 204 データに関して正しく識別できた(図 6)。データセット 2 では全 2391 データ中 73%の 1757 データに関して正しく識別できた(図 7)。

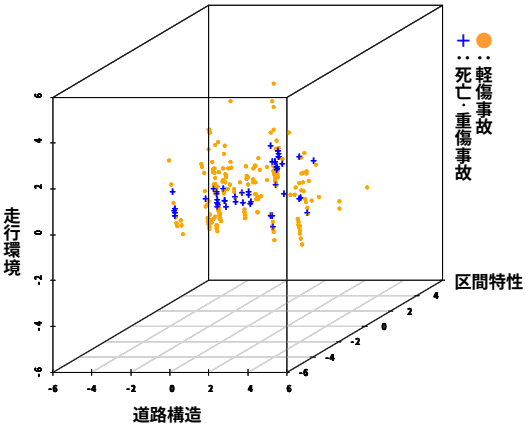

図 6　判別結果(高齢者)

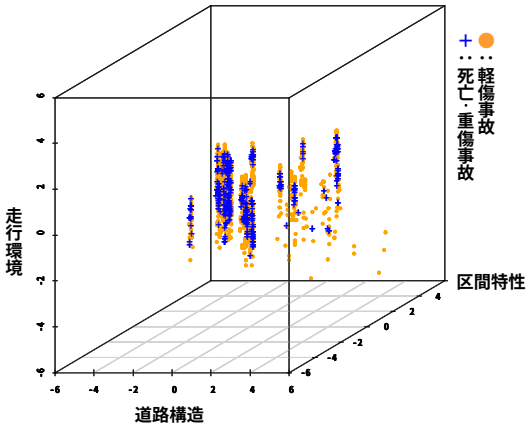

図 7　判別結果(非高齢者)

両者ともに非的中サンプルが、局所的ではなく全体的にあるので、汎化性能は良いモデルといえる。

３．５　分離曲面の推定

事故の際に、死者・重傷者が生じる可能性が高い危険領域を把握するためにグリッドデータを用いて分離曲面を推定した。なお、グリッドデータは三次元座標データであり、互いに直交する道路構造軸、区間特性軸、走行環境軸の各軸方向において、最小値-5、最大値 5 のデータ領域内を 100 等分したときの分割線の交点を座標として持つ。

推定方法は、3.3 で構築したモデルを用いてグリッドデータの識別関数値((5)式の $f(x)$)を求め、$-0.1< f(x)<0.1$ の値の集合を分離曲面とした。

図 8 に高齢者事故の分離曲面と判別結果を示す。分離曲面をより立体的に見せるために、陰影をつけて描画した。△は死亡・重傷事故データで、○は軽傷事故データである。

図 8 より、SVM によって判別された重大事故データは分離曲面の中に存在しており、これらは区間特性軸方向に偏りが見られ、道路構造軸、走行環境軸方向ではばらついていることが読み取れる。つまり、区間特性が高齢者の重大な事故が発生する要因になっている。区間特性は歩道(設置延長)、中央帯設置延長、指定最高速度からなり、これらを総合的に改良できる対策により、高齢者が運転する車両が原因となる重大事故の発生確率を減らすことができると考えられる。

また、構築した SVM に既存道路の特徴量を入力すれば、事故発生時に死者・重傷者が発生しやすい道路区間が特定でき、優先的に事故対策を行うべき区間を抽出することができると思われる。

非高齢者の危険領域の推定結果、および高齢者との比較・検討は発表時に行う。

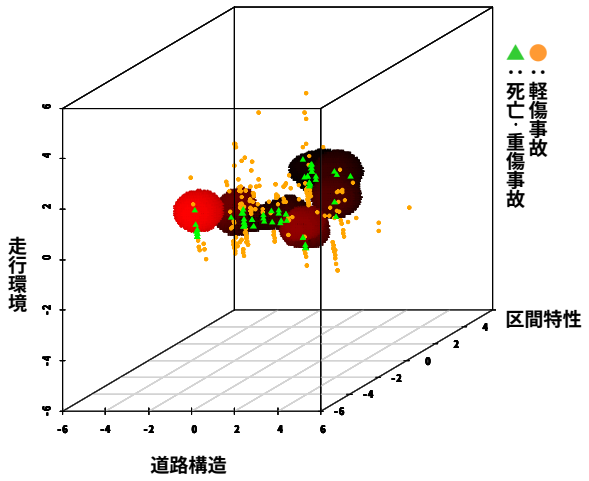

図 8　分離曲面と判別結果(高齢者)

４．おわりに

本研究では、ITARDA データを用いてデータセットを 2 つ（高齢者、非高齢者）作成し、死亡・重傷事故と軽傷事故に SVM で識別し、両データセットを比較した。

死亡・重傷事故データ数が軽傷事故データ数よりも圧倒的に少なかったが（約 10 分の 1)、結果的に両データセットとも 7 割程度の判別率でもって識別することができたので、データが少数であっても同程度の判別関数は得ることができるといえる。しかし、主成分分析で要因数を 3 つに縮約したため、マクロな施策立案でしか使えなくなってしまうという欠点がある。

謝辞：本研究を進めるにあたり、北海道開発局函館開発建設部からデータの一部を提供して頂いた。ここに記して謝意を表します。

参考文献

1) 北海道警察本部交通部交通企画課：平成 17 年中の交通事故概況、財団法人北海道交通安全協会 Web ページ、http://www.safety110.jp/h17nennkann/h171231.htm
2) N. Cristianini & J. Shawe-Taylor : An Introduction to Support Vector Machines and other kernel-based learning methods, Cambridge University Press, 2000. （邦訳:大北剛:サポートベクターマシン入門,共立出版, 2005.）
3) 木村浩、古田一雄：意思決定問題に見る地域性と知識の役割－原子力政策に対する賛否の意思決定を例題として－、社会情報学研究、Vol. 9、No. 1、pp. 41-53、日本社会情報学会(2005)