



サポートベクターマシンの土木計画学への適用に関する研究

メタデータ	言語: jpn 出版者: 土木学会 公開日: 2013-02-28 キーワード (Ja): サポートベクターマシン, 機械学習, データマイニング キーワード (En): 作成者: 長谷川, 裕修, 藤井, 勝, 有村, 幹治, 田村, 亨 メールアドレス: 所属:
URL	http://hdl.handle.net/10258/1768

サポートベクターマシンの土木計画学への適用に関する研究

その他（別言語等） のタイトル	Applying the Support Vector Machine to the Infrastructure Planning
著者	長谷川 裕修, 藤井 勝, 有村 幹治, 田村 亨
雑誌名	土木計画学研究・講演集
巻	35
号	102
ページ	1-4
発行年	2007-06
URL	http://hdl.handle.net/10258/1768

サポートベクターマシンの土木計画学への適用に関する研究*

Applying the Support Vector Machine to the Infrastructure Planning*

長谷川裕修**・藤井勝***・有村幹治****・田村亨*****

By Hironobu HASEGAWA**・Masaru FUJII***・Mikiharu ARIMURA****・Tohru TAMURA*****

1. はじめに

近年、ITS 技術の向上によりプローブデータに代表される連続的かつ多面的なデータが入手可能となっている。多くの研究者によりこれらのデータ活用方法の開発が進められているが、十分に活用できているとは言い難い。これらのデータの利用方法の一つとして、データマイニングを経て、データを分類する手法（判別分析）が土木計画学でも多用されてきた。しかし、これまで土木計画分野で一般に用いられてきた判別分析や数量化Ⅱ類は線形性と正規性を前提としており、非線形性を有する実問題への応用が課題であった。

これに対して非線形の手法であるニューラルネットワークの適用が検討されたが、過学習や局所解の問題を解決することはできなかった。そこで本研究では、非線形分離が可能なパターン認識手法の中でもその識別性能の高さから近年注目されているサポートベクターマシン（Support Vector Machines：以下 SVM と記す）の土木計画学への適用を検討する。SVM は凸 2 次計画問題の形をとるため、大域的最適解を得ることが保証される。また、ソフトマージン最適化の導入によって過学習を避け高い汎化性能を得ている。

2. サポートベクターマシンについて^{1)・2)}

SVM は、1960 年代に Vapnik が提案した最適超平面識別器を基礎とする二値判別手法である。最適超平面識別器は線形分離可能な問題に対しては高い性能を示し

たが、非線形な問題に対応できないため利用は進まなかった。しかし、最適超平面識別器は 1990 年代に Vapnik 自身によってカーネル法と組み合わせられることで非線形判別が可能な SVM として拡張された。この拡張により SVM は最も認識性能の優れた手法の一つと言われ、多様な分野で利用が広がっている。

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。図-1 に SVM の概念図を示す。

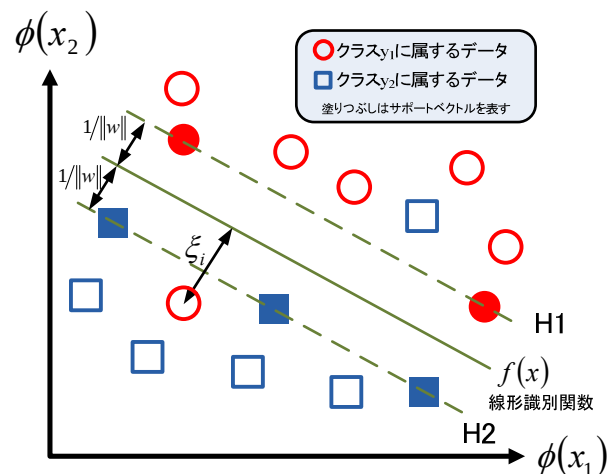


図-1 SVM 概念図

SVM を用いたパターン識別は、2つのクラス y_1 、 y_2 に属するデータ（図中の赤丸と青四角）を分離する識別関数 $f(x)$ を求める問題になる。分離可能な識別関数は無数にあるが、データの存在する領域の限界面 $H1$ 、 $H2$ とデータを分離する超平面間の距離 $1/\|w\|$ を最大化させるような識別関数 $f(x)$ を求める。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。ハードマージン SVM ではデータにノイズがある場合の過学習が問題となり実用には適さない。この問題を解決すべく提案されたソフトマージン SVM について以下に簡単に説明する。

まず、 $\mathbf{x}_i$ ($i=1\sim\ell$) 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

*キーワード：サポートベクターマシン，機械学習，データマイニング

**学生員，工修，室蘭工業大学大学院 工学研究科 博士後期課程 建設工学専攻（北海道室蘭市水元町27番1号、TEL0143-46-5289、FAX0143-46-5289）

***正員，工修，室蘭工業大学大学院 工学研究科 博士後期課程 建設工学専攻

****正員，工博，（株）ドーコン 交通部

*****フェロー，工博，室蘭工業大学工学部 建設システム工学科

定義される。

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

ここで、 w は入力に対する重みベクトルである。

この最適化問題（主問題）に Lagrange 乗数 α_i を導入することで (2) 式で表される双対問題が得られる。

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

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。

まず、入力データ x_i を高次元の特徴空間に写像する。

$$x \leftrightarrow \phi(x) = (\phi_1(x), \phi_2(x), \dots) \quad (3)$$

(2) 式内の特徴空間に対応する量はベクトルの内積で表される値であり、カーネル関数で置換することができる。この置換をカーネルトリックといい、これによって計算量の増大を防いでいる。カーネルトリックを

(2) 式に適用し、(4) 式に示す凸 2 次計画問題が得られる。なお、この最適化問題の局所的最適解は必ず大域的最適解となる。

$$\begin{aligned} & \underset{\alpha}{\text{maximize}} \quad W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j K(x_i, x_j), \alpha \in R^{\ell} \\ & \text{subject to} \quad \sum_{i=1}^{\ell} \alpha_i y_i = 0, 0 \leq \alpha_i \leq C, i=1, \dots, \ell \end{aligned} \quad (4)$$

このとき、バイアス b は (5) 式で得ることができる。

$$b^* = y_i - \sum_{j \in S_v} \alpha_j^* y_j K(x_j, x_i) \quad (5)$$

ここで、 S_v 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 (6) 式となる。

$$f(x) = w^{*T} x_i + b^* = \sum_{j \in S_v} \alpha_j^* y_j K(x_j, x) + b^* \quad (6)$$

さて、バイアス b を固定値とすると (4) 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_i を更新すれば大域的最適解を求めることが出来る。

$$\begin{aligned} & \alpha_i \leftarrow \min \left(C, \max(0, \alpha_i + \eta \frac{\partial W(\alpha)}{\partial \alpha_i}) \right) \\ & \eta_i = \frac{\omega}{K(x_i, x_i)} \quad (0 \leq \omega \leq 2), i=1, \dots, \ell \end{aligned} \quad (7)$$

この最適化問題の収束判定は、主問題と双対問題の目的関数の値の比較により行う。なお、収束条件 ε は

一般に 10^{-4} が用いられる。

$$\begin{aligned} \text{Proportion} &= \frac{\text{主目的関数値} - \text{双対目的関数値}}{\text{主目的関数値} + 1} \\ &= \frac{\sum_{i=1}^{\ell} \alpha_i - 2W(\alpha) + C \sum_{i=1}^{\ell} \xi_i}{\sum_{i=1}^{\ell} \alpha_i - W(\alpha) + C \sum_{i=1}^{\ell} \xi_i + 1} \leq \varepsilon \end{aligned} \quad (8)$$

ここに、

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

3. SVM適用事例

SVM はテキストのカテゴリ化、画像認識、手書き文字認識、バイオインフォマティクス等幅広い分野で応用されている³⁾。

我が国の土木計画分野における SVM 適用事例としては、庭田らによるアンケート自由回答の自動分類への適用⁴⁾と筆者らによる交通事故分類への適用⁵⁾がある。

庭田らは PI の過程で行われたアンケート調査で得られた自由記述式の回答の中からランダムに抽出された 215 文をトレーニングデータ、182 文を検証用データとし、SVM による 8 クラス（質問、疑い、確認、要求、不満、懸念、賛成、反対）の分類を試み、約 60% の正解率を得ている。これまで土木計画分野において多くのアンケート調査が行われてきたが、自由記述式の回答は「～という意見もあった」程度の取り扱いであり、それさえもアンケート実施者の恣意的な選択によるものであった。公共性の高い土木計画分野であるからこそ「顧客の声」を拾い上げる努力が重要であり、この研究の意義は大きい。

以下に筆者らによる交通事故分類への適用についてまとめる。

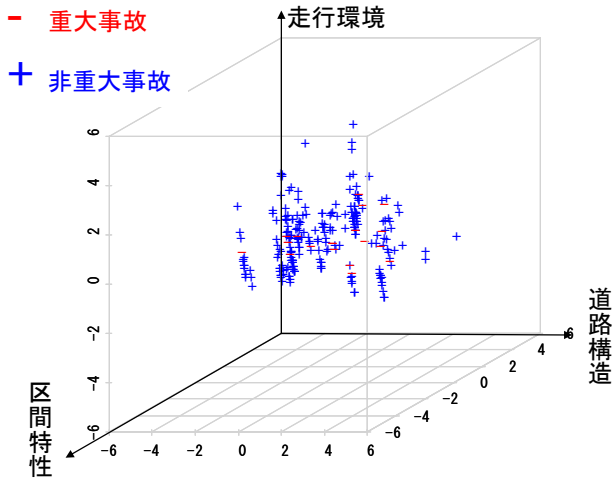
(1) 分析対象

(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) のうち、1) 平成 11 年度から 16 年度に北海道渡島・松山支庁で発生、2) 加害者年齢 65 歳以上、3) DID 地区の交差点部で発生した事故データ 257 件（エラーデータを除く）を対象に分析を行った。入力変数は車道幅員、歩道代表幅員、車線数、指定最高速度、歩道設置延長、中央帯設置延長、平日昼間 12 時間大型車混入率、路面状態、事故発生月、平日昼夜率の 10 変数、判別クラスは死亡・重傷者発生事故と軽傷事故の 2 クラスとした。

(2) データの下処理

SVM による分類を行う前に、データの下処理として主成分分析による次元縮約を行った。主成分分析の結果、変数は道路構造、区間特性、走行環境の 3 主成分に集約

された。図－2は分析対象の257データを道路構造軸、区間特性軸、走行環境軸からなる主成分軸空間にプロットしたものである。



図－2 データの分布

(3) モデルパラメータの設定

前節で作成したデータセットの中から、 $y=1$ のデータ 13 個、 $y=-1$ のデータ 20 個、合計 33 個を抽出しトレーニングデータとした。

なお、カーネル関数 $K(x_i, x_j)$ は (9) 式で表される Gauss カーネルを用いた。

$$K(x_i, x_j) = \exp\left[-\frac{\|x_i - x_j\|^2}{2\gamma^2}\right], \quad i, j = 1, \dots, \ell \quad (9)$$

SVM の識別能力にはソフトマージンのパラメータ C と Gauss カーネルのパラメータ γ の組み合わせが影響するが、これは繰り返し計算によって適当な値を求める必要がある。本研究では、10-fold cross validation 手法を用いてパラメータの組み合わせを決定した。10-fold cross validation 手法とは、ある C と γ の組み合わせに対して、1) 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 セットをテストサンプルとして識別を行い、予測精度を求める。2) テストサンプルを順に替え、それぞれの予測精度を求め、これらの予測精度の平均を当該組み合わせの予測精度とする手法である。この結果、本研究では $C=32$ 、 $\gamma=2$ の組み合わせを用いることとした。

(4) 未学習データによるモデルの検証

トレーニングデータを含む全 257 データを検証用データとし、前節で求めたパラメータを用いて構築した

SVM による判別を行った。判別結果を表－1 に示す。

検証用データ全体の誤識別率は約 28.8%、重大事故の誤識別率は 35.0%、非重大事故の誤識別率は 28.3%であった。これより、構築した SVM は未学習データに対して高い識別能力を持っていることが分かる。

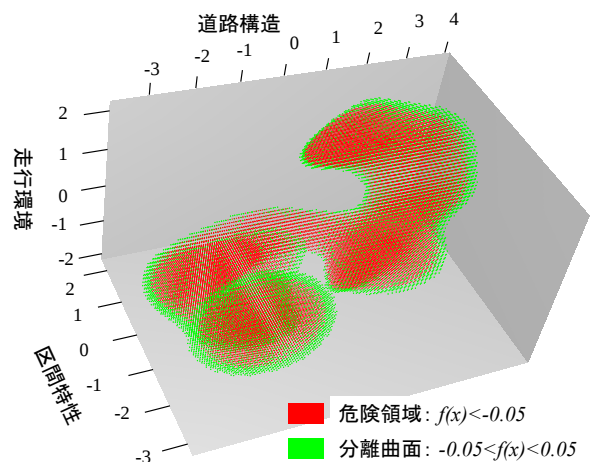
表－1

	正解数	不正解数	誤識別率
重大事故	13	7	35.0%
非重大事故	170	67	28.3%
TOTAL	183	74	28.8%

(5) 危険領域の推定

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、グリッドデータとは、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三次元座標データである。

構築したモデルを用いてグリッドデータの識別関数値((6)式を参照)を求め、 $-0.05 < f(x) < 0.05$ の値の集合を分離曲面とした。図－4に分離曲面の推定結果を示す。図中の緑色で示された分離曲面を境にして、内側の赤い領域が危険領域、外側が安全領域を表している。



図－3 分離曲面推定結果

この図より、次の2点に分かる。

- 1) 分離曲面は袋状になっており、その内側に危険領域が存在する（饅頭の皮とアンの関係）。
- 2) 危険領域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。すなわち、区間特性（歩道設置延長、中央帯設置延長、指定最高速度からなる）が高齢者の重大な事

故が発生する要因となっていることが明らかとなった。

以上より、走行速度などの区間特性値がある領域の中にある場合（例えば、速度が遅すぎず・早過ぎない状態で）、高齢者が運転する車両が重大事故を起こしやすいことが分かった。

また、SVM を用いることで、これまで判別できなかった要因の値域を識別でき、その値域に対して事故対策をすれば効果が上がる可能性が高いことが指摘できた。

4. おわりに

本研究は非線形の識別関数による分類を行う SVM の土木計画分野への適用可能性について論じたものである。本研究により、交通事故分析に SVM を適用することで、これまで判別できなかった要因の値域を識別でき、その値域に対して事故対策をすれば効果が上がる可能性が高いことが指摘できた。

本研究を通して、2007 年春大会で議論したいことの背景は以下の通りである。人々が多様な価値観・ライフスタイルを持ち、それらがインターネットを通じて容易に交換される現代において、過去に観測された現象・行動に一定の法則を仮定し、これを将来へ延長しようとする伝統的な予測モデルを用いることには大きな問題がある。データオリエンテッドに立ち返り、現象の記述・理解に努める必要があろう。

そして、ここで議論したい内容をまとめると次の通りである。土木計画分野には伝統的な PT 調査やアンケート調査結果の蓄積がある。また、近年の ITS 技術の進展によってプローブデータに代表される大量の時間的に連続したデータが蓄積されつつある。これらのデータから『隠された知識』を発見し、意思決定に必要な基準を発見することが土木計画学の次の研究方向として重要なテーマではなかろうか。

SVM を含む機械学習やデータマイニングなど、データから科学的な仮説や知識を発見するための理論と技術は「発見科学」として体系的な研究が進んでいる。発見科学における基本的な研究項目は以下の 5 つにまとめられよう。

1. 知識発見の論理

計算機による知識の発見においては、個々の仮説や理論はもとより仮説空間の論駁可能性が問題になる。また、発見に関する研究は、事例に学ぶ実証的な側面をもつ一方で、厳密な表現形式と演繹体系を用いた理論的な基礎づけを必要とする。

2. 推論による知識発見

科学的知識の発見には、上記の演繹推論のほか

に、帰納とアブダクション（発想）が必要である。アブダクションは、既存の知識では説明不可能な事象について仮説を立てて説明しようとする際に用いられるもので、科学的発見の契機を与えるものである。

3. 計算学習理論に基づく知識発見

実験・観測データからの知識発見には、計算機に例から規則を学習させる、いわゆる機械学習の手法が有効である。その基礎理論である計算学習理論に基づいた知識発見の方式について、計算量的に徹底した配慮をしながら研究をする必要がある。

4. 巨大データベースからの知識発見

以上のような知識発見に関する研究は、最終的には統合してシステム化し、それを自然科学および産業界における巨大なデータベースに対して実際に適用・評価して、次の研究にフィードバックすべきである。また、データマイニングの具体的方法論と手法を確立する必要がある。

5. ネットワーク環境における知識発見

知識発見においては、機械と人間、人間と人間、機械と機械の相補的共同作業が重視される。したがって、協調、競合、妥協、調停などの種々の様相を含むこうした相補的共同作業を効率よく遂行・支援するための計算機ネットワーク環境を研究開発する必要がある。

本研究で紹介した SVM による交通事故分析は 3 と 4 に関連するが、これに限らず、土木計画における知識発見の重要性について春大会で議論したい。

参考文献

- 1) N. Cristianini & J. Shawe-Taylor : An Introduction to Support Vector Machines and other kernel-based learning methods, Cambridge University Press, 2000.
- 2) 栗田多喜夫：サポートベクターマシン入門、産業技術総合研究所 脳神経情報研究部門 Web サイト、<http://staff.aist.go.jp/takio-kurita/index-j.html>
- 3) 例えば、Kernel-Machines Website : <http://www.kernel-machines.org/>
- 4) 庭田美穂、福田大輔、屋井鉄雄：自由回答の疑問型表現に着目した市民の関心の抽出方法に関する基礎的研究、第 33 回土木計画学研究発表会・講演集、vol.33、CD-ROM、2006.
- 5) 長谷川裕修、藤井勝、有村幹治、田村亨：交通事故分析へのサポートベクターマシンの適用に関する基礎的検討、第 34 回土木計画学研究発表会・講演集、vol.34、CD-ROM、2006.