研究レポート　No.5　～機械学習による判別分析～

2013 年 9 月 26 日　　株式会社アイズファクトリー　http://bodais.jp/company/

概要

企業の日々の営業活動において、多くの履歴が残されている。このデータをもとに、将来の営業活動として取るべき行動の示唆を得ることが、判別分析の利用方法の一例として挙げられる。近年、判別分析の手法として、サポートベクターマシン（SVM）が注目されている。本稿では、SVM の利用されるビジネスシーンとロジックについて解説し、決定木を使った画像解析の結果と比較して、SVM の判別能力の高さを示す。

1. はじめに

企業は、日々さまざまな営業活動を行っている。多くの企業は、営業に関する履歴を残している。いつ、どのような手段で、どこの部署に、どんな商品を売り込みに行ったのか。商談の結果はどうだったのか。この営業活動の履歴をもとに、どのような商品を次に売り込むべきかなど、将来の取るべきアクションを考えることもできる。

営業先	接触日時	接触手段	先方部署	売込商品	…	契約商品
A社	2013/5/11	訪問	商品部	商品X		商品X
B社	2013/6/7	電話	企画部	商品Y		商品Z
C社	2013/7/14	訪問	営業部	商品Z		商品Y
D社	2013/8/26	電話	商品部	商品Q		商品X

表 1. 営業履歴テーブル

将来取るべきアクションの判定を、データに基づいた分析をもとに行うことが、判別分析の一例として挙げられる。判別分析の手法には、決定木やナイーブベイズ判別機などがあるが、近年、判別精度の高い手法として、サポートベクターマシン（SVM）が注目されている。表 1 のように、過去の営業履歴を項目別に洗い出し、その商談の結果（契約商品など）を教師データとして、判別モデルを構築する。構築した判別モデルを利用して、現在進行中の案件や、将来的に想定される案件の結果を予測する。予測された結果をもとに、講じるべき策を考える。このような営業活動のサイクルに、SVM を用いた判別分析を利用することができる。なお、SVM の手法に関する詳細な解説として、文献 [1] ~ [4] を挙げておく。

2. マージン最大化と 2 次計画問題

判別モデルが扱う問題は、線形分離可能な問題と不可能な問題に分類される（図 1）。ビジネスシーンで遭遇する問題は、通常の場合、線形分離不可能な問題である。SVM は、線形分離不可能な問題に、情報を高次元空間に写像することにより、超平面で分離して、対処する（図 2）。また、SVM では、この超平面を、マージン最大化という学習則によって求めていく。

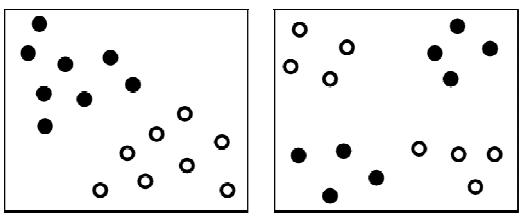

図 1. 線形分離可能な例（左）と線形分離不可能な例（右）（出所：文献 [2]）

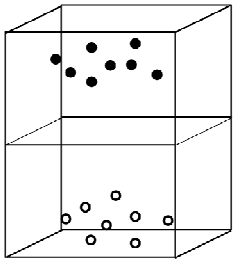

図 2. 高次元空間への写像（出所：文献 [2]）

クラス 1 とクラス 2（図 3 の C1, C2 に相当）の識別境界を考える。2 つの識別境界から等距離の地点には、実線が引いてある。破線から実線までの距離をマージンと呼ぶ。図 3 では、定義可能な 2 つのマージンが描いてある。識別境界の選択の仕方で、マージンが増減することが窺える。未知のデータに対する学習モデルの汎化性能を高める場合には、解が学習データに依存しすぎないように工夫する必要がある。図 3（左）では、線形超平面で完全なクラス分類が可能であるが、学習データの分布に合わせて非線形超平面を当てはめてしまうと、未知のデータに対しては予測力を大幅に失う可能性が高くなる。解がなるべく学習データに依存しないようにする方法として、マージンを最大化するような線形超平面の選択が考えられる。

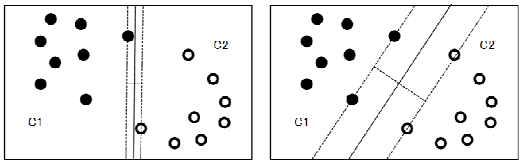

図 3. マージン最大でない識別境界（左）と最大である識別境界（右）（出所：文献 [2]）

マージン最大化のためには凸 2 次計画問題を解く必要があり、主問題の目的関数と不等式制約は以下のように表される。

目的関数：$\|\omega\|^2/2$，不等式制約：$y_i(\omega' x_i + b) \geq 1$

この問題の双対問題のラグランジアンは、

$$L(\alpha) = \sum \alpha_i - \sum\sum \alpha_i \alpha_{i*} y_i y_{i*} x'_i x_{i*} / 2$$

と表され、以下の制約を考慮して最大化することにより主問題である目的関数を最適化する。

制約条件：$\alpha_i \geq 0$，$\sum \alpha_i y_i = 0$

b は含まれていないが、次式で推定できる。

$$b = -(\omega' x_{SV1} + \omega' x_{SV2})/2$$

x_{SV1}，x_{SV2}は、マージン最大化後に、識別超平面から最短距離にある学習データであり、それぞれクラス 1、クラス 2 に属する。これらは、サポートベクターと呼ばれる。

3. 高次元化と非線形カーネル

SVM といえども、線形識別子として扱う限り、判別能力には限界がある。実際のデータ分布は複雑に込み入っており、単純な平面では上手く分割できない。実際には複雑に入り組んだ曲面による判別が必要となる。これを実現する方法として高次元化と非線形カーネルを説明する。

線形分離不可能問題に対しては、高次元空間への写像が効果的であることを冒頭で示した。したがって、x_iに関して高次元空間への写像$\Phi(x_i)$を考慮すると、次の双対問題を最大化することになる。

$$L(\alpha) = \sum \alpha_i - \sum\sum \alpha_i \alpha_{i*} y_i y_{i*} \Phi(x_i)'\Phi(x_{i*})/2$$

この式には、$\Phi(x_i)$の内積が表れている。$\Phi(x_i)$の内積をx_iの内積の関数として表現したものをカーネルといい$k(x'_i\ x_i)$と表記する。カーネルを利用することで、写像関数$\Phi(x_i)$の計算を回避することができる。これをカーネルトリックと呼ぶ。カーネルが存在するならば、双対問題のラグランジアンは、

$$L(\alpha) = \sum \alpha_i - \sum\sum \alpha_i \alpha_{i*} y_i y_{i*} k(x'_i\ x_{i*})/2$$

と表される。この式は、x_iの内積がカーネルで表されている以外、線形 SVM において定義された双対問題と同等である。つまり、線形分離不可能な問題に対しても、適切なカーネルが存在すれば、線形分離可能な場合と同様の学習則を利用することが可能となる。以下に、利用されることの多い、代表的なカーネルをあげておく。

線形カーネル：$k(x'_i\ x_j) = x'_i\ x_j$

多項式カーネル：$k(x'_i\ x_j) = (\gamma x'_i\ x_j + \delta)^d$

RBF カーネル：$k(x'_i\ x_j) = \exp(-\gamma \left\| x_i - x_j \right\|^2)$

シグモイドカーネル：$k(x'_i\ x_j) = \tanh(\gamma(x'_i\ x_j) - \delta)$

上述の 3 つのカーネルの未知のパラメタは、識別機の精度に大きく影響するため、適当な値にチューニングする必要がある。

4. 性能比較

SVM の判別能力の高さを示すため、画像解析の結果を一例として示す。図 4 はマレーシアのランカウイ島の衛星画像から、植生を判別分析した解析結果である（文献 [5]）。

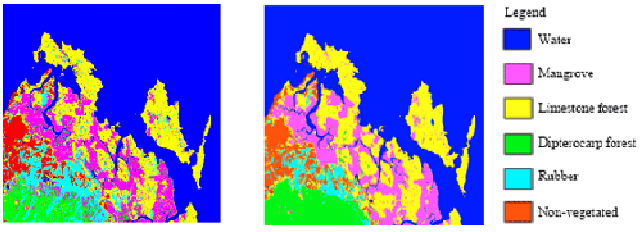

図 4. 決定木（左）と SVM（右）による衛星画像の判別分析（出所：文献 [5]）

Method	Kernel type	Overall accuracy (%)	Kappa coefficient
Support vector machine	Linear (Default)	73.7006	0.6630
	Linear (Optimum)	75.4903	0.6841
	Polynomial (Default)	74.2442	0.6694
	Polynomial (Optimum)	75.5405	0.6847
	RBF (Default)	74.7543	0.6753
	RBF (Optimum)	76.0004	0.6900
	Sigmoid (Default)	73.5919	0.6616
	Sigmoid (Optimum)	74.5034	0.6726
Decision tree	-	68.7846	0.6014

表 2. 植生の判別精度（出所：文献 [5]）

決定木を使った判別では、Dipterocarp forest（フタバガキ科の高木の森林）の植生が、Mangrove（マングローブ）や Rubber（ゴム）に誤って分類されたりしている（図 4）。また、表 2 で示すように、決定木（Decision tree）より SVM を用いた判別の方が、高い精度で判別できている。このように、衛星画像の解析からも SVM の高い判別能力を窺い知ることができる。

5. まとめ

本稿では、SVM のビジネスシーンにおける利用例と、ロジックについて解説し、その判別能力の高さを、画像解析の例を挙げて示した。判別分析のほかの利用例としては、以下のものなどがある。(1) 多数のメールの中から、スパムメールを除外する。(2) 手書き文字を認識して、コンピュータに処理させる。(3) 遺伝子の発現量を解析して、正常細胞と癌細胞を識別する。(4) 企業の倒産の可能性を予測する。これらに共通していることは、データを用いて判別を行っているということである。(1) の例では、メールの中の単語や文脈（テキストデータ）から、スパムメールであるか否かの判別が行われている。(2) の文字認識では、手書き文字の画像データから、その文字がどの文字に該当するかという判別が行われている。(3) の遺伝子の例では、遺伝子の発現量のデータから、正常細胞と癌細胞の判別が行われている。(4) の企業倒産の例では、財務データや景気動向のデータをもとに企業が倒産しそうか否かの判別が行われている。SVM を用いた判別分析は、今後、様々な場面で有効活用され、更なる広がりを見せていくことであろう。

6. 参考文献

[1] Cortes, Corinna and Vapnik, Vladimir N., "Support-Vector Networks", Machine Learning, 20, 273-297, 1995.

[2] 豊田秀樹　「データマイニング入門　—R で学ぶ最新データ解析—」，2008 年，東京図書.

[3] 小野田崇　「知の科学　サポートベクターマシン」，2007 年，オーム社.

[4] Nello Cristianini and John Shawe-Taylor,「サポートベクターマシン入門」、2005 年、共立出版.

[5] H.Z.M. Shafri and F.S.H. Ramle, “A Comparison of Support Vector Machine and Decision Tree Classifications Using Satellite Data of Langkawi Island.” Information Technology Journal, 8: 64-70, 2009.