2002年日本オペレーションズ・リサーチ学会
春季研究発表会

多目的計画法を用いたSVM

甲南大学　*浅田　武史　ASADA Takeshi
甲南大学　中山　弘隆　NAKAYAMA Hirotaka

1. 序

Support Vector Machines(SVMs) は近年、パターン分類問題で注目されている手法の1つである。一般的には2次計画問題 (Quadratic Programming;QP) で定式化されるが、距離関数を変えることで線形計画問題 (Linear Programming;LP) で定式化することができる。こうすることで計算時間を短縮する事ができる。ところで、SVMには完全分離を要求するHard margin問題があるが、場合によっては過学習 (over learning) になることがある。この問題を回避するため、Soft margin問題が導入される。しかし、Soft margin問題は重要なデータを不要なデータ（outlier）と見なす事がある。こうしたHard margin問題、Soft margin問題がかかえる問題を回避する方法として、本論文では多目的計画問題を用いたSVMを導入する。

2. Generalized Support Vector Machines

n次元実空間上に2つの集合A、Bがあり、このいずれかに属するようなデータ$\mathbf{x}_i(i = 1,\ldots,m)$があるとする。集合$A$の要素$\mathbf{x}_i$に対して$y_i = +1$を与え、集合$B$の要素$\mathbf{x}_j$に対して$y_j = -1$を与える。ここで集合$A$と$B$を線形な超平面で分離することを考える。線形分離不可能な場合は、原空間Xからある高次元特徴空間 (high dimensional feature space)Zへの非線形写像$\varphi : \mathbf{x}_i \longrightarrow \mathbf{z}_i \quad (\mathbf{x}_i \in X,\ \mathbf{z}_i \in Z)$を考え、線形分離可能な状態にする。このとき、$Z$における線形な分離超平面を$\mathbf{w}^{\mathrm{T}}\mathbf{z} + b = 0$とすれば、次のような問題を得る：

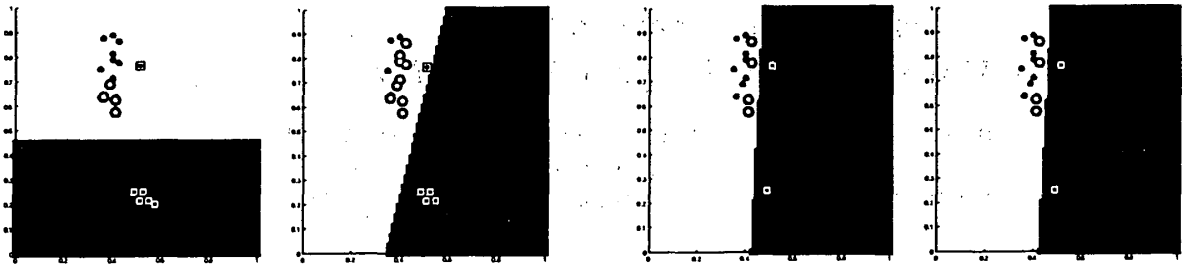
$$\begin{array}{lll} \text{[GSVM]} \quad \text{Minimize} & \|\mathbf{w}\|_q & \\ \text{s.t.} & y_i\left(\mathbf{w}^{\mathrm{T}}\mathbf{z}_i + b\right) \geq 1 & (i = 1,\ldots,m) \end{array} \tag{2.1}$$

ここで、距離を測るものとしてl_pノルムを用い、その共役ノルムをl_qノルムと定義した。

距離の測定においてl_2ノルムを用いれば問題 (2.1) は2次計画問題になる。一方、l_1ノルムやl_∞ノルムを用いれば問題 (2.1) は線形計画問題になる。

3. Hard margin問題とSoft margin問題

前章で述べられた問題は全て完全分離を要求するものであった。これをHard margin問題という。しかし、Hard margin問題では、場合によっては過学習になることがあるかもしれない。この問題を回避するため、Soft margin問題が導入される。Soft margin問題とは、slack変数ξを用いる事で、正しく分離できないデータを許容するものである。



$C = 1.0$(L)　　$C = 1.0$(M)　　$C = 10.0$(L)　　$C = 10.0$(M)

図 1: 多目的計画法による SVM(M で示す) と線形計画法による SVM(L で示す) との比較。点 (0.5125, 0.7625) は、(L) では不要なデータとして除外しているが、(M) ではそのようなことを行っていない。

4. 多目的計画法の利用

Soft margin 問題は過学習を回避するためのものであったが、Soft margin 問題が許容するエラーデータの中には、場合によっては重要なデータが混じっている可能性もある。こうした重要なデータを見落とすことを回避するため、多目的計画法を用いる。そのため、まず slack 変数 $\boldsymbol{\xi}$ を導入する。これは正しく分離されなかったデータと分離超平面との距離を表す。次に、余剰変数 (surplus variables) $\boldsymbol{\eta}$ を導入する。これは正しく分離されたデータと分離超平面との距離を表す。この 2 つの変数を用いて以下の問題が定式化される：

$$
\begin{aligned}
\text{[MOPSVM]} \qquad \text{Minimize} \quad & C\sum_{i=1}^{m}\xi_i + \sum_{i=1}^{m}\eta_i \\
\text{s.t.} \quad & y_i\left(\mathbf{w}^{\mathrm{T}}\mathbf{z}_i + b\right) \geq 1 - \xi_i \\
& y_i\left(\mathbf{w}^{\mathrm{T}}\mathbf{z}_i + b\right) \leq 1 + \eta_i \\
& \boldsymbol{\xi} \geq \mathbf{0},\ \boldsymbol{\eta} \geq \mathbf{0} \qquad (i = 1, \ldots, m)
\end{aligned}
\tag{4.1}
$$

人工的なデータによる従来の SVM と [MOPSVM] との比較は図 1 で示される。

5. 終わりに

Soft margin 問題、Hard margin 問題にある固有の問題を解決するために導入された多目的計画法は分離する 2 つの集合のうち、データ数の少ない方の領域を広めていることが分かる。ところで実際のデータはどちらかのカテゴリ値に偏っている傾向にある。従って、実際の問題においては有力な手法と考えられ、良い結果が得られると期待される。

参考文献

[1] O.L.Mangasarian, *Generalized Support Vector Machines*, In A.Smola, P.Bartlett, B.Shölkopf, and D.Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 135-146, Cambridge, MA ,2000, Mit Press

[2] K.P.Bennett and C.Campbell, *Support Vector Machines: Hype or Hallelujar?*, SIGKDD Explorations, 2, 2, 2000

[3] 中山弘隆・谷野哲三, 多目的計画法の理論と応用, 社団法人計測自動制御学会編, コロナ社,1994