数理解析研究所講究録 1461 巻 2005 年 154-160

サポートベクターマシンによる大規模データ分類

株式会社数理システム　山下浩 (Hiroshi Yamashita)
株式会社数理システム　雪島正敏 (Masatoshi Yukishima)

Mathematical Systems Inc.

1　はじめに

サポートベクターマシンは，2値分類のための学習機械として近年広く研究されているが，学習データの数が大規模な場合に計算時間が膨大になるという欠点がある．本発表では，この点についてクラスタリングによるデータスカッシングと，それに伴って新しい概念を入れたサポートベクター分類器について述べる．

与えられたトレーニングデータ集合 S を $(\mathbf{x}_i, y_i), i = 1, ..., \ell$ とする．$\mathbf{x}_i$ は $\mathbb{R}^N$ の点で，$y_i \in \{-1, +1\}$ を教師データとする．データ集合 S が $\mathbb{R}^N$ の超平面によって $y_i = 1$ のグループと $y_i = -1$ のグループに分離される場合を線形分離可能と言う．候補となる超平面を $(\mathbf{w}\cdot\mathbf{x}) + b = 0$ と表す．ここで，$\mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}$ で，$(\mathbf{w}\cdot\mathbf{x})$ は $\mathbf{w}$ と $\mathbf{x}$ の内積を表す．識別関数を $f(x) = \mathrm{sgn}\left((\mathbf{w}\cdot\mathbf{x}) + b\right)$ とする．超平面 $(\mathbf{w}, b)$ に対するサンプル $(\mathbf{x}_i, y_i)$ のマージンは $\delta_i = y_i\left(\frac{(\mathbf{w}\cdot\mathbf{x}_i)+b}{\|\mathbf{w}\|}\right)$ と定義される．($\|\cdot\|$ は2ノルム.) $\delta_i > 0$ ならばデータ $(\mathbf{x}_i, y_i)$ は正しく識別されていることになる．上記マージン $\delta_i, i = 1, ...\ell$ の最小値をサンプルデータ集合 S に対する超平面 $(\mathbf{w}, b)$ のマージンと言う．

(i) 線形分離可能な場合　まず線形分離可能なデータ集合を考えると，分離の条件は $y_i((\mathbf{w}\cdot\mathbf{x}_i) + b) > 0,\ i = 1, ...\ell$ となるので，超平面を正規化して $\min_i \{y_i((\mathbf{w}\cdot\mathbf{x}_i) + b)\} = 1$ とすることができる．このとき超平面のマージンは $\|\mathbf{w}\|^{-1}$ となる．マージンが最大になるような超平面を考えると，それは，2次計画問題

$$\begin{array}{ll} \text{最小化} & \frac{1}{2}\|\mathbf{w}\|^2, \qquad\qquad \mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R} \\ \text{条　件} & y_i\left((\mathbf{w}\cdot\mathbf{x}_i) + b\right) \geq 1, \quad i = 1, ...\ell \end{array}$$

の解によって与えられる．カーネルトリックのためには，以下の双対問題を考えると都合が良い：

$$\begin{array}{ll} \text{最大化} & -\frac{1}{2}\sum_{i,j=1}^{\ell} \alpha_i\alpha_j y_i y_j (\mathbf{x}_i\cdot\mathbf{x}_j) + \sum_{i=1}^{\ell}\alpha_i, \quad \boldsymbol{\alpha} \in \mathbb{R}^\ell \\ \text{条　件} & \sum_{i=1}^{\ell}\alpha_i y_i = 0, \boldsymbol{\alpha} \geq \mathbf{0} \end{array}$$

最適解を $(\mathbf{w}^*, b^*, \boldsymbol{\alpha}^*)$ とすると，$\mathbf{w}^*, b^*$ は $\boldsymbol{\alpha}^*$ によって表すことができるが，双対変数 $\alpha_i^*, i = 1, ...\ell$ の中で非零のものは不等式制約条件が有効となっているもの，すなわち $y_i\left((\mathbf{w}^*\cdot\mathbf{x}_i) + b\right) = 1$ となるもの（サポートベクター）なので，サンプル数に比べて数が少ないことが期待できる．このように，分離超平面は少数のサポートベクターで表現されることが多い．

(ii) 線形分離が不可能な場合　次に，線形分離が不可能な場合を考える．この場合は元の問題の条件にスラック変数を導入して，目的関数にペナルティを課すと，主問題と双対問題は（$C_i > 0$ はペナルティパラメータ）：

$$\begin{array}{ll} \text{最小化} & \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{\ell} C_i\xi_i, \qquad\qquad \mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^\ell \\ \text{条　件} & y_i\left((\mathbf{w}\cdot\mathbf{x}_i) + b\right) \geq 1 - \xi_i, \xi_i \geq 0, \quad i = 1, ...\ell \end{array}$$

$$\begin{array}{ll} \text{最大化} & -\frac{1}{2}\sum_{i,j=1}^{\ell} \alpha_i\alpha_j y_i y_j (\mathbf{x}_i\cdot\mathbf{x}_j) + \sum_{i=1}^{\ell}\alpha_i, \quad \boldsymbol{\alpha} \in \mathbb{R}^\ell \\ \text{条　件} & \sum_{i=1}^{\ell}\alpha_i y_i = 0, \mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{C} \end{array} \tag{1}$$

(iii) 曲面による識別　このような目的のために，高次元の空間への非線形変換とその空間でのカーネルトリックと言われる方法が知られている．まず，入力データをより高次元な特徴空間に（非線形）射像する．すなわち，$\mathbf{x} \mapsto \boldsymbol{\phi}(\mathbf{x}) = (\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), ...)$ として，特徴空間において線形分離を考える．その際に，内

積に対応する項をカーネル関数で置き換える：$K(\mathbf{x},\mathbf{y})=(\boldsymbol{\phi}(\mathbf{x})\cdot\boldsymbol{\phi}(\mathbf{y}))$．このとき，1ノルムソフトマージン最適化問題は：

$$\begin{array}{ll}\text{最大化} & -\frac{1}{2}\sum_{i,j=1}^{\ell}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i,\mathbf{x}_j)+\sum_{i=1}^{\ell}\alpha_i,\quad \boldsymbol{\alpha}\in\mathbb{R}^{\ell}\\ \text{条　件} & \sum_{i=1}^{\ell}\alpha_i y_i=0,\ \mathbf{0}\le\boldsymbol{\alpha}\le\mathbf{C}\end{array}\tag{2}$$

通常のサポートベクター分類器は，この(2)を解く．大規模な問題，すなわちデータの数ℓが多い場合には，上記2次計画問題のヘッセ行列が大規模密行列となり，通常の2次計画法のアルゴリズムはもとより，SMOあるいはSVM^{light}等の近似アルゴリズムによる求解でも多大な計算時間を要する．したがって，大規模データに対しては特別な工夫が必要とされる．本発表はそのための試みである．

上で述べたように，サポートベクター分類器は最終的にサポートベクターとなるデータの数は全データ数に比べてそれ程多くないことが期待でき，かつサポートベクター以外のデータは分離曲面の形状には関与しないことが特徴である．したがって，少数の一部データから元のデータのサポートベクターを精度よく決定できれば目的を達成することができる．

本研究では，元のデータをクラスタリング手法によってグループ分けして，その各クラスターを一つのデータとみなし，それらに対して何らかのサポートベクター分類法を適用することを考える．クラスター数は，元のデータの数よりも相当程度少なくなっていることが期待できるので，サポートベクター分類は比較的小規模なものとなる．この結果から，元データの中でサポートベクターの候補となるデータのみを選び出して，再度サポートベクター分類を適用して最終的な結果を得ることも可能である．このような観点からの研究は[1],[2]にも見られる．

2　大きさを持ったデータに対するサポートベクター分類法

クラスタリングによってまとめられたデータ群はクラスターの“位置”と“大きさ”という属性を持っていると仮定する．ここでは簡単のために，中心の座標と半径によって上記性質を表すことにする．したがって，有限の大きさを持つ“球”の集まりであるデータ群を分類する問題となる．本研究では，大規模データに対する分類の高速化という観点からこのような問題に到達したが，この問題はそれ自体興味深いものである．たとえば，データが誤差を含むときに多次元空間内の点として表すよりも，大きさをもったものとして考えた方が自然な場合がある．また，多次元の時系列データのように一つの実体が複数の期にまたがるデータを保持しているときに，別な点として表すよりも過去の履歴データによって生成される大きさを持った“点”として考える方が合理的である．このような観点からサポートベクターマシンのアルゴリズム（サポートボールマシン）を考えると，興味深い性質が分ってくる．

トレーニングデータが大きさを持っているということから，データ集合Sは$(\mathbf{x}_i,y_i,r_i),i=1,...,\ell$とする．$r_i$はデータの“半径”である．超平面$(\mathbf{w},b)$に対するサンプル$(\mathbf{x}_i,y_i)$のマージンは

$$\delta_i=y_i\left(\frac{(\mathbf{w}\cdot\mathbf{x}_i)+b}{\|\mathbf{w}\|}\right)-r_i$$

となる．

(i) 線形分離可能な場合　分離の条件は$y_i((\mathbf{w}\cdot\mathbf{x}_i)+b)-\|\mathbf{w}\|r_i>0,\ i=1,...\ell$となるので，$\min_i\{y_i((\mathbf{w}\cdot\mathbf{x}_i)+b)-\|\mathbf{w}\|R_i\}=1$と正規化する．マージン最大化問題は

$$\begin{array}{ll}\text{最小化} & \frac{1}{2}\|\mathbf{w}\|^2,\mathbf{w}\in\mathbb{R}^N,b\in\mathbb{R}\\ \text{条　件} & y_i((\mathbf{w}\cdot\mathbf{x}_i)+b)-\|\mathbf{w}\|r_i\ge 1,i=1,...\ell\end{array}$$

となる．

(ii) 線形分離が不可能な場合 この場合の問題はスラック変数と対応するペナルティパラメータを導入して

$$\begin{array}{ll} \text{最小化} & \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{\ell} C_i'\xi_i, \mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^\ell \\ \text{条　件} & y_i(\mathbf{w}\cdot\mathbf{x}_i)+b) - \|\mathbf{w}\| r_i \geq 1-\xi_i, \xi_i \geq 0, i = 1, ...\ell \end{array}$$

となる．補助変数 $p \in \mathbb{R}$ を導入すると，上の問題は

$$\begin{array}{ll} \text{最小化} & \frac{1}{2}p^2 + \sum_{i=1}^{\ell} C_i'\xi_i, \ \mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^\ell, p \in \mathbb{R} \\ \text{条　件} & y_i\left((\mathbf{w}\cdot\mathbf{x}_i)+b\right) - pr_i \geq 1-\xi_i, i = 1, ...\ell \\ & p = \|\mathbf{w}\|, \boldsymbol{\xi} \geq \mathbf{0} \end{array} \tag{3}$$

となる．この問題を緩和して

$$\begin{array}{ll} \text{最小化} & \frac{1}{2}p^2 + \sum_{i=1}^{\ell} C_i'\xi_i, \ \mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^\ell, p \in \mathbb{R} \\ \text{条　件} & y_i\left((\mathbf{w}\cdot\mathbf{x}_i)+b\right) - r_i p \geq 1-\xi_i, i = 1, ...\ell \\ & p \geq \|\mathbf{w}\|, \boldsymbol{\xi} \geq \mathbf{0} \end{array} \tag{4}$$

とすると，2 次錐計画問題となる．今後，最適解では $\mathbf{w} \neq \mathbf{0}$ と仮定する．

「緩和問題 (4) の最適解は (3) の最適解となる.」何故なら，(4) の最適解 $(\mathbf{w}, b, \boldsymbol{\xi}, p)$ が $p > \|\mathbf{w}\|$ となったとすると，不等式制約条件を犯すことなく目的関数を更に減少させることができるので矛盾である．緩和問題の最適解が $p = \|\mathbf{w}\|$ をみたすので，それは (3) の最適解となる

問題 (4) のラグランジュ関数は

$$L(p, \mathbf{w}, b, \boldsymbol{\alpha}', \boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{1}{2}p^2 + \sum_{i=1}^{\ell} C_i'\xi_i - \sum_i \alpha_i'(y_i\left((\mathbf{w}\cdot\mathbf{x}_i)+b\right) - r_i p - 1 + \xi_i) - (\beta_p p + \boldsymbol{\beta}_w \cdot \mathbf{w}) - \boldsymbol{\gamma}\cdot\boldsymbol{\xi}$$

となる．ここで，$\boldsymbol{\alpha}' \in \mathbb{R}^\ell, \beta_p \in \mathbb{R}, \boldsymbol{\beta}_w \in \mathbb{R}^N, \boldsymbol{\gamma} \in \mathbb{R}^\ell$ は双対変数である．双対問題は

$$\begin{array}{ll} \text{最大化} & -\frac{1}{2}\left(\beta_p - \sum_i \alpha_i' r_i\right)^2 + \sum_i \alpha_i', \ \boldsymbol{\alpha} \in \mathbb{R}^\ell, \beta_p \in \mathbb{R}, \boldsymbol{\beta}_w \in \mathbb{R}^N \\ \text{条　件} & \sum_i \alpha_i' y_i = 0, \boldsymbol{\beta}_w = -\sum_i \alpha_i y_i \mathbf{x}_i, \\ & \mathbf{0} \leq \boldsymbol{\alpha}' \leq \mathbf{C}', \beta_p \geq \|\boldsymbol{\beta}_w\| \end{array}$$

あるいは

$$\begin{array}{ll} \text{最大化} & -\frac{1}{2}\left(\beta_p - \sum_i \alpha_i' r_i\right)^2 + \sum_i \alpha_i', \ \boldsymbol{\alpha}' \in \mathbb{R}^\ell, \beta_p \in \mathbb{R}, \boldsymbol{\beta}_w \in \mathbb{R}^N \\ \text{条　件} & \sum_i \alpha_i' y_i = 0, \beta_p^2 \geq \sum_{i,j} \alpha_i'\alpha_j' y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j, \\ & \mathbf{0} \leq \boldsymbol{\alpha}' \leq \mathbf{C}', \beta_p \geq 0 \end{array} \tag{5}$$

となる．最適解では相補性条件

$$\begin{aligned} \alpha_i'(y_i\left((\mathbf{w}\cdot\mathbf{x}_i)+b\right) - r_i p - 1 + \xi_i) &= 0, \\ \alpha_i' \geq 0, y_i\left((\mathbf{w}\cdot\mathbf{x}_i)+b\right) - r_i p - 1 + \xi_i &\geq 0 \end{aligned}$$

$$\beta_p p + \boldsymbol{\beta}_w \cdot \mathbf{w} = 0 \tag{6}$$

$$p\boldsymbol{\beta}_w + \beta_p \mathbf{w} = \mathbf{0} \tag{7}$$

$$(C_i' - \alpha_i')\xi_i = 0, C_i' - \alpha_i' \geq 0, \xi_i \geq 0 \tag{8}$$

と制約条件（ラグランジュ関数の停留条件）

$$\begin{aligned} \beta_p &= p + \sum_i \alpha_i' r_i \\ \boldsymbol{\beta}_w &= -\sum_i \alpha_i y_i \mathbf{x}_i \\ \sum_i \alpha_i' y_i &= 0 \end{aligned}$$

が成立している．また，$p=\|\mathbf{w}\|$ と (7) より $\beta_p=\|\boldsymbol{\beta}_w\|=\|\sum_i \alpha_i' y_i \mathbf{x}_i\|$ (>0) が成立している．

$$\mathbf{w}=\frac{\beta_p-\sum_i \alpha_i' r_i}{\beta_p}\sum_i \alpha_i' y_i \mathbf{x}_i$$

正しく識別されたサポートベクター $(0<\alpha_i'<C_i', \xi_i=0)$ に対して

$$y_i\left((\mathbf{w}\cdot\mathbf{x}_i)+b\right)-r_i p-1=0$$

なので

$$b=y_i(r_i p+1)-\frac{\beta_p-\sum_j \alpha_j' r_j}{\beta_p}\sum_j \alpha_j' y_j \mathbf{x}_j\cdot\mathbf{x}_i$$

となる．識別関数は

$$f(x)=sgn\left(\frac{\beta_p-\sum_i \alpha_i' r_i}{\beta_p}\sum_i \alpha_i' y_i \mathbf{x}_i\cdot\mathbf{x}+b\right)$$

で与えられる．

データの半径がゼロとなるとき，上の問題は通常の SVM 最適化問題に一致することを示すことができる．

(iii) 曲面による識別　1 ノルムソフトマージン最適化問題は

$$\begin{array}{ll}\text{最大化} & -\frac{1}{2}\left(\beta_p-\sum_i \alpha_i' r_i\right)^2+\sum_i \alpha_i',\ \boldsymbol{\alpha}'\in\mathbb{R}^\ell, \beta_p\in\mathbb{R}\\ \text{条　件} & \sum_i \alpha_i' y_i=0, \beta_p^2\geq\sum_{i,j}\alpha_i'\alpha_j' y_i y_j K(\mathbf{x}_i,\mathbf{x}_j),\\ & \mathbf{0}\leq\boldsymbol{\alpha}'\leq\mathbf{C}', \beta_p\geq 0\end{array}\qquad (9)$$

このとき，識別関数は $(0<\alpha_i'<C_i', \xi_i=0)$

$$b=y_i(r_i p+1)-\frac{\beta_p-\sum_j \alpha_j' r_j}{\beta_p}\sum_j \alpha_j' y_j K(\mathbf{x}_j,\mathbf{x}_i)$$

$$f(x)=\mathrm{sgn}\left(\frac{\beta_p-\sum_i \alpha_i' r_i}{\beta_p}\sum_i \alpha_i' y_i K(\mathbf{x}_i,\mathbf{x})+b\right)$$

となる．

3　カーネルによる “半径” の表現

カーネルトリックを利用する場合には，元の空間での “半径” のデータも変換してカーネルによる表現を得る必要がある．任意の $\mathbf{x}\in\mathbb{R}^N, \mathbf{y}\in\mathbb{R}^N$ に対して

$$\|\mathbf{x}-\mathbf{y}\|^2=\|\mathbf{x}\|^2+\|\mathbf{y}\|^2-2(\mathbf{x}\cdot\mathbf{y})$$

となり，元のデータ空間での距離 r は

$$r=\|(r\mathbf{e}+\mathbf{x})-\mathbf{x}\|=\left(\|(r\mathbf{e}+\mathbf{x})\|^2+\|\mathbf{x}\|^2-2\left((r\mathbf{e}+\mathbf{x})\cdot\mathbf{x}\right)\right)^{1/2}$$

と表せるので（$\mathbf{e}$ は任意の単位ベクトル）

$$r\longmapsto\left(K((r\mathbf{e}+\mathbf{x}),(r\mathbf{e}+\mathbf{x}))+K(\mathbf{x},\mathbf{x})-2K((r\mathbf{e}+\mathbf{x}),\mathbf{x})\right)^{1/2}$$

と置きなおす．RBF カーネルでは

$$r\longmapsto\left(2-2\exp(-r^2/\sigma^2)\right)^{1/2}$$

となり，単位ベクトル $\mathbf{e}$ の取り方に依存しない．多項式カーネルやシグモイドカーネルでは $\mathbf{e}$ の取り方に依存するので，何らかの工夫が必要となる．以下の実験では RBF カーネルを使用した．

4 近似問題とその解法

上記2次錐計画問題の解法は通常の制約付き非線形最適化のアルゴリズムで解く事はできるが，データスカッシングを実行した後でも，大規模問題となることが予想でき現実的ではない．そこで，SMOのような近似アルゴリズムを利用することを考える．問題(3)の制約条件に現れる $\|\mathbf{w}\|$ を定数 $\bar{p} \geq 0$ で置き換えた問題：

$$\begin{array}{ll} \text{最小化} & \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{\ell} C_i\xi_i, \mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^\ell \\ \text{条　件} & y_i(\mathbf{w}\cdot\mathbf{x}_i)+b) - \bar{p}r_i \geq 1-\xi_i, \xi_i \geq 0, i = 1, \ldots \ell \end{array}$$

を考える．この問題の双対問題は

$$\begin{array}{ll} \text{最大化} & -\frac{1}{2}\sum_{i,j=1}^{\ell} \alpha_i\alpha_j y_i y_j(\mathbf{x}_i\cdot\mathbf{x}_j) + \sum_{i=1}^{\ell}(1+\bar{p}r_i)\alpha_i, \boldsymbol{\alpha} \in \mathbb{R}^\ell \\ \text{条　件} & \sum_{i=1}^{\ell} \alpha_i y_i = 0, \mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{C} \end{array} \tag{10}$$

となり，問題(1)を少しだけ変更した問題になっている．したがって，SMOなどの既存の近似アルゴリズムを少し変更するだけで解く事ができる．カーネルトリックを利用した問題に対しても同様である．

$\bar{p} > 0$ を動かして上記問題をSMOアルゴリズムで解き，$\bar{p} = p = \|\mathbf{w}\|$ となる最適解を求める．問題(10)の解で，$p = \bar{p}$ が成立していて，解 α_i が得られたとする．このとき

$$\frac{p}{\beta_p}\alpha_i' = \alpha_i,\ \beta_p = p + \sum_i \alpha_i' r_i$$

が成立すれば，α_i' は問題(5)において $C_i' = \frac{\beta_p}{p}C_i$ と置いた問題の解になっている．

$$\beta_p = p + \frac{\beta_p}{p}\sum_i \alpha_i r_i$$

より $p > \sum_i \alpha_i r_i$ ならば

$$\beta_p = \frac{p^2}{p - \sum_i \alpha_i r_i}$$

と置けばよい．$p > \sum_i \alpha_i r_i$ が成立しない場合は，問題(5)の解は得られない．

逆に，問題(5)の解 α_i' が得られたとする．このとき，α_i を

$$\frac{p}{\beta_p}\alpha_i' = \alpha_i$$

とおけば，α_i は $C_i = \frac{p}{\beta_p}C_i', \bar{p} = p$ とおいた問題(10)の解となっている．

また，問題(10)の最適解において $\bar{p} \neq p$ のときは，$\bar{p} \to p, r_i \to r_i\bar{p}/p$ とおけば，上記議論が適用できるので，“半径”を変化させた問題の解として対応付けが可能となる．

5 数値実験

上記アルゴリズムを実際のデータを用いて評価した．実験にはWindowsXP Pentium4 3GHz CPU 2GByteメモリの計算機を使用した．データはUCI KDD repositoriesで公開されているForest cover typeを使用した([3])．このうち，主要な2つのcover typeであるSpruce/FirとLodgepole Pineを教師情報（上記 y_i の値）として実験を行った．Forest cover typeのデータは10個の数値変数と44個のバイナリ変数からなる．このうちバイナリの5変数の内容はcover typeをSpruce/FirとLodgepole Pineに制限した場合には同一の値をとるので分析には使用しない．数値データとバイナリデータで範囲が異なる為，数値データは $[0,1]$ の範囲に規格化して分析した．使用したデータは表1．の通りである．

クラスタリングには，CF ベクトルを用いた BIRCH アルゴリズムを採用した ([4])．このアルゴリズムではクラスタの半径 (分散) が大きくならないように閾値を設けて行う．クラスタリングの結果，980 のクラスタが作成された．

サポートベクター分類器のカーネル関数は RBF カーネルを使用し，分類器の作成時には，上で述べたようにクラスタの半径を考慮した．また，クラスタに含まれるデータ数をペナルティパラメータの重みとして考慮した．サポートベクター分類器は SMO アルゴリズムで実装した．

分類器の性能は，誤判別率と学習時間を用いて評価した．誤判別率の判定には，学習データと検証データを分けて，検証データの誤判別率で計算した．比較のために，全データを用いた場合と，ランダムサンプリングを行ったデータに対しても分類器を作成し評価を行った．ランダムサンプリング後のデータ数はクラスタリングの結果のデータ数と等しくなるようにした．

内容	説明
説明変数	49
目的変数	カテゴリ
学習データ	
Type Spruce/Fir	46159
Type Lodgepole Pine	52869
検証データ	
Type Spruce/Fir	165681
Type Lodgepole Pine	230432

表 1．データ内容

クラスタリングによる方法とランダムサンプリングによる方法で分類器を作成する場合は，誤判別率が小さくなるようなペナルティパラメータ（$C_i = C \times w_i$，w_i はクラスタ内のデータ数に比例した重みで $\sum w_i = 1$ ）を求めた．全データの場合は，この係数は 1 に固定した．

結果は表 2. の通りである．Baseline Error Rate は検証データで件数の多いクラスであると予測した場合の誤判別率である．計算時間は 1 ケースあたりのものである．

モデル（データ数）	誤判別率	計算時間 (秒)
ランダムゲス/Baseline Error Rate	41.8%	
ランダムサンプリング (980)	31.2%	6
本方法 (980)	22.4%	9
ランダムサンプリング (1928)	27.9%	20
本方法 (1928)	19.1%	24
ランダムサンプリング (4512)	23.3%	132
本方法 (4512)	17.2%	124
ランダムサンプリング (11613)	18.5%	692
本方法 (11613)	14.5%	1035
全件 (99028)	11.3%	30832

表 2．計算結果

実験結果から明らかなように，クラスタリングによりデータ件数を減らすことで学習が大幅に高速化されることが判った．本発表で提案したデータが" 半径" を持つというモデルに関しては，モデルの解析・計算実験とも今後行うべき事は多い．

参考文献

[1] H.Yu,J.Yang, and J.Han, Classifying large data sets using SVM with hierarchical clusters, In Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003.

[2] D.Boley and Dongwei Cao, Training support vector machine using adaptive clustering, Technical report, University of Minnesota, 2004.

[3] UCI KDD archive http://kdd.ics.uci.edu/

[4] Tian Zhang, Raghu Ramakrishnan and Miron Livny,"BIRCH: an efficient data clustering method for very large databases", Proceedings of the 1996 ACM SIGMOD international conference on Management of data, pp. 103 - 114, 1996.