G1001-06
0915-1060/10/¥500/論文/JCLS

スパース正則化およびマルチカーネル学習のための最適化アルゴリズムと画像認識への応用

東京大学　冨岡 亮太・鈴木 大慈／東京工業大学　杉山 将

スパース正則化は凸最適化を通して変数選択や多様な情報源の統合を実現するための系統的な枠組みとして近年注目されている。多くのスパース正則化法は滑らかでない最適化問題として定式化されるため、このような問題を効率的に解く方法が求められている。本解説論文では、一般の凸な損失関数と広いクラスの正則化関数に対する最適化手法をproximal minimizationの枠組みから導出し、理論的な性質を議論し、既存の手法との関係を解説する。また、スパース正則化の特別な場合としてマルチカーネル学習を取り上げ、これを画像認識に応用する。

はじめに

スパース正則化は、いかに観測データを少ない数の説明変数や基底関数で説明するか、という問題（変数選択）に対するひとつの近似解法として統計科学[1]や信号処理[2)、3)]の分野で研究されてきた。一方、近年、パターン認識や機械学習の分野ではSVMに代表されるカーネル法[4]の成功によって非常に多数の説明変数を用いることが日常的になってきたため、スパース正則化を含む新しい正則化の方法に注目が集まっている。例えばマルチカーネル学習[5)、6)]と呼ばれる手法は（特別な場合に）ひとつのスパース正則化法とみなせることが明らかになり、統計、信号処理、機械学習の問題を個別に扱うのではなく包含するような枠組みやアルゴリズムが求められている。

問題設定

本解説ではスパース正則化学習あるいはスパース正則化信号復元を次の最適化問題の解として定義する：

$$\underset{w\in R^n}{\text{minimize}} \quad L(Aw)+\phi_\lambda(w). \cdots (1)$$

ここでwはn次元のベクトルで求めたい判別関数の係数ベクトルであったり、復元されるべき信号であるとする。Lは損失関数、$A(\in R^{m\times n})$はデザイン行列で、通常は観測の数mが未知変数の数nより小さい。ϕ_λは正則化項である。ここでは損失関数は下に凸で2階微分可能とする。微分可能でない損失関数（例えばマルチカーネル学習で用いられるヒンジ関数）は別に扱う必要がある。正則化関数ϕ_λは下に凸であるが、微分不可能であってもよい関数で、任意の正の定数$\eta>1$について、$\eta\phi_\lambda=\phi_{\eta\lambda}$と仮定する。また、$\lambda>1$を正則化定数と呼ぶ。

［例］例えば，$L(z)=\frac{1}{2}\|z-b\|_2^2$, $\phi_\lambda(w)=\lambda\|w\|_1$とおくと、以下の最適化問題を得る。

$$\underset{w\in R^n}{\text{minimize}} \frac{1}{2}\|Aw-b\|_2^2+\lambda\|w\|_1. \cdots (2)$$

ただし$\|\cdot\|_2$は通常のユークリッドL_2ノルム、$\|w\|_1=\sum_{j=1}^n|w_j|$は$L_2$ノルムを表す。これは不良設定な最少2乗あてはめ問題をL_1ノルムを用いて正則化したものであるが、L_1ノルムの原点での微分不能性のために有限の$\lambda>1$でスパースな、すなわち係数ベクトルwが多くのゼロ要素を含むような最適解を持つことが知られていて、統計科学ではlasso(least absolute shrinkage and selection operator)[1]と呼ばれている。

この他にも損失関数は解きたい問題に応じて尤度関数の対数などから様々なものが得られる。例えばロジスティック関数[7]：

$$f_\ell(z)=\sum_{i=1}^n\log(1+e^{-y_iz_i})$$

が挙げられる。

最適化アルゴリズム

本解説では式(1)の最適化問題を効率よく解くアルゴリズムを扱う。これまでに多くの方法が提案されてきた。

難しさの由来

最適化問題(1)はなぜ解くのが難しいのだろうか？そのひとつの理由は微分不可能であるということであ

る。そこで1つの方法として(1)を拘束条件を持つ微分可能な問題に書き直し、内点法[8]を適用することが考えられる。内点法は特別な損失／正則化関数の場合には多項式時間で解けることが知られているなど理論的には魅力的であるものの、学習や信号復元の問題ではデザイン行列Aが密であるため、実際の計算時間では多くの既存手法に遅れを取っている。

微分不可能性が最適化問題(1)を解く際の唯一の障害だろうか？ここでデザイン行列Aが単位行列であると仮定し、簡単のため(2)の損失関数と正則化関数の組を考える。

$$\min_{w\in\mathrm{R}^n}\left(\frac{1}{2}\|w-b\|_2^2+\lambda\|w\|_1\right)= \sum_{j=1}^{n}\min_{w_j\in\mathrm{R}}\left(\frac{1}{2}(w_j-b_j)^2+\lambda|w_j|\right) \quad \cdots (3)$$

であるので、上の式を最小化するw_j^*は：

$$w_j^*=\mathrm{ST}_\lambda(b_j)=\begin{cases} b_j-\lambda & (\lambda\le b_j),\\ 0 & (-\lambda\le b_j\le\lambda),\\ b_j+\lambda & (b_j\le-\lambda).\end{cases} \cdots (4)$$

これをb_jからw_jへの関数とみてSoft-Thresholding関数[9][10][11][12][13]と呼び、その概形を図1(b)に示した。この関数の形からL_1正則化がいかにスパースな解をもたらすかということを端的に見ることができる。すなわち、－λからλ間の値はゼロに切り捨てられ、絶対値λ以上の値は原点方向にλだけ縮小されている。これは一種の非線形ノイズ除去処理と見てもよい。結局、$A=I_n$の場合、損失項が変数ごとに分離しているため、最適化問題(2)は解析的に解くことができることがわかった。

Proximity operator

上で見たSoft-Thresholding関数はMreau[14]によってproximity operatorと名付けられ研究された(Combettes & wajs[11]も参照)。一般に関数fのMoreau's envelope Fは以下で定義される：

$$F(z)=\inf_{x\in\mathrm{R}^n}\left(\frac{1}{2}\|x-z\|_2^2+f(x)\right)\cdots (5)$$

さらに、zから最小化を達成する点への対応をproximity operatorと呼び、以下のように定義する：

$$\mathrm{prox}_f(z)=\arg\min_{x\in\mathrm{R}^n}\left(\frac{1}{2}\|x-z\|_2^2+f(x)\right) \quad \cdots (6)$$

[事実1] 一般に下に凸で下半連続な関数fとその凸共役[注1] (Legendre変換)f^*に関して以下が成り立つ：

$$\mathrm{prox}_f(z)+\mathrm{prox}_{f^*}(z)=z. \quad \cdots (7)$$

この式はzの一種の直交分解とみなすことができて、Combettes & Wajs[11]ではこの式をzのfに関するMoreau's decompositionと呼んでいる。

[事実2] 一般に下半連続な下に凸な関数fのMoreau's envelope Fは微分可能で、その微分は$\nabla_w F(w)=\mathrm{Prox}_{f^*}$で与えられる。(第1図参照)ただし$\mathrm{Prox}_{f^*}$は$f$の凸共役に関するproximal operatorである。

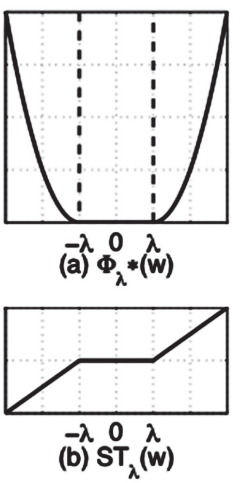
Iterative Shrinkage/Thresholding (IST)法

上で見たように最適化問題(1)は損失関数の項が変数ごとに分離していれば、効率的に解くことができる。そこで以下のように(1)の損失関数項を点w^tの周りで局所的に分離可能な形に近似することを考える[13]：

$$L(Aw)\cong L(Aw^t)+\nabla f_1(Aw^t)^{\mathrm{T}}A(w-w^t) +\frac{1}{2\eta_t}\|w-w^t\|_2^2=Q_{\eta_t}(w;w^t). \quad \cdots (8)$$

Lの2階微分可能性より、少なくとも局所的にはη_tを十分小さく選ぶ事で$L(Aw)\le Q_{\eta_t}(w;w^t)$となるようにすることができる。従って、次のような反復アルゴリズムを考えることができる[9][10][11][12][13]：

1. w^1を適当に初期化。
2. 停止基準が満たされるまで反復：



第1図　(a) 1次元箱関数 $\phi_\lambda^*(w)=\begin{cases}0 & (|w|\le\lambda)\\ +\infty & (|w|>\lambda)\end{cases}$ **のMoreu's envelopeの概形。箱関数は1次元L_1正則化関数** $\phi_\lambda(w)=\lambda|w|$ **の凸共役である。(b) Soft-Thresholding関数(式(4))。この関数は(a)のMoreau's envelopeの微分になっている（事実2）。**

注1；一般に、関数fの凸共役(convex conjugate)関数f^*を$f^*(y)=\sup_{w\in\mathrm{R}^n}\left(y^{\mathrm{T}}w-f(w)\right)$と定義する。

$$w^{t+1} = \arg\min_{w\in\mathrm{R}^n}\left(Q_{\eta_t}(w;w^t) + \phi_\lambda(w)\right) \quad \cdots (9)$$

（ただし $\eta_1, \eta_2, \ldots$ は適当に選ばれているとする。）

ここで、式 (9) の最小化は以下のように書き直せることに注意する：

$$w^{t+1} = \arg\min_{w\in\mathrm{R}^n}\left(\phi_{\eta_t\lambda}(w) + \frac{1}{2}\left\|w-(w^t-\eta_t A^{\mathrm{T}}\nabla L(Aw^t)\right\|_2^2\right). \quad \cdots (10)$$

ただし、仮定より $\eta\phi_\lambda = \phi_{\eta\lambda}$ を用いた。式 (10) は式 (5) の形をしているため、式 (6) の proximity operator を用いると以下のように書き直すことができる：

$$w^{t+1} = \mathrm{prox}_{\phi_{\eta_t\lambda}}\left(w^t - \eta_t A^{\mathrm{T}}\nabla L(Aw^t)\right) \quad \cdots (11)$$

式 (11) は次のように解釈することができる。まず、正則化項を無視してステップ t での解 w^t からステップサイズ η_t の勾配ステップを取る。その後、正則化項の効果を proximity operator を用いて取り込む。例えば $\phi_\lambda(w) = \lambda\|w\|_1$ の場合、$\mathrm{prox}_{\phi_\lambda(w)} = \mathrm{ST}_\lambda$ である（式 (4) の Soft-Thresholding 関数）。ステップサイズ η_t が適切に（例えば w^t の近傍で $L(Aw) \le Q_{\eta_t}(w;w^t)$ となるように）取られていればこのアルゴリズムは大域的最適解に収束する。この方法は Iterative Shrinkage/Thresholding (IST)[15] と呼ばれている。IST の利点は 1 反復あたりの計算量（勾配の計算コスト＋proximity operator の計算コスト）が小さいことである。一方 IST はステップサイズ η_t の選択方法が難しい（これに関しては最近 Wright ら[13] で扱われている）ことや、等方的な 2 次関数で近似してしまっているため、スケーリングに弱いなどの問題点がある。

提案手法—DAL

Proximal minimization に基づく導出

この節では dual augmentd Lagrangin (DAL) と呼ぶ最適化法を提案する。DAL は初め双対問題に対する拡張ラグランジュ法として導出[16]（「拡張ラグランジュ法としての提案手法」節を参照）されたが、ここでは上の IST 法との関連に注目して導出する。前の節では式 (1) の損失項を式 (8) の形に線形近似することで proximity operator に基づくアルゴリズム (IST) が導出されることを見た。ここでは線形近似ではなく双対問題を通して proximity operator を導出する。まず、以下のような反復アルゴリズムを考える：

1. w^1 を適当に初期化。
2. 停止基準が満たされるまで反復：

$$w^{t+1} = \arg\min_{w\in\mathrm{R}^n}\left(f(w) + \frac{1}{2\eta_t}\left\|w-w^t\right\|_2^2\right). \quad \cdots (12)$$

ただし関数 f は式 (1) のように定義した。ここで式 (12) は式 (9) と異なり $L(Aw)$ の線形近似でなく $L(Aw)$ そのものが最小化の中に入っていることに注意する。式 (12) は proximity operator を用いて、$w^{t+1} = \mathrm{prox}_{\eta_t f}(w^t)$ のように書き直すことができる。このような反復アルゴリズムは proximal minimization algorithm と呼ばれ、おおよそ $1/\eta_t$ のレートで収束することが知られている（$\eta_t \to \infty$ のとき超 1 次収束[17]）。式 (12) は本来の目的関数に 2 次関数を加えただけであり、一見最適化問題として式 (1) を解くのと同程度に困難であるように見えるが、実は式 (12) の proximal minimization は微分可能な補助目的関数を最小化することで得られることが示せて、結局、以下の繰り返しアルゴリズムを得る[18]：

1. w^1 を適当に初期化。
2. 停止基準が満たされるまで反復：

$$w^{t+1} = \mathrm{prox}_{\phi_{\eta_t\lambda}}\left(w^t + \eta_t A^{\mathrm{T}}\alpha^t\right) \quad \cdots (13)$$

ただし、α^t は以下の微分可能な最小化問題の解である：

$$\alpha^t = \arg\min_{\alpha\in\mathrm{R}^m}\left(L^*(-\alpha) + \frac{1}{\eta_t}\Phi^*_{\eta_t\lambda}\left(w^t + \eta_t A^{\mathrm{T}}\alpha\right)\right). \quad \cdots (14)$$

ここで、L^*, ϕ^*_λ は L, ϕ_λ の凸共役関数とし、Φ^*_λ は ϕ^*_λ の Moreau's envelope である。（式 (5) を参照）

式 (13) で $\alpha^t = -\nabla L(Aw^t)$ とおくと式 (11) の更新式 (IST) を得ることに注意する。

陰勾配法としての提案手法

式 (12) に陰勾配法 (implicit gradient method) としての解釈を与える[17]。w^{t+1} は式 (12) を最小化するので、$0 \in \partial f(w^{t+1}) + \frac{1}{\eta_t}(w^{t+1}-w^t)$. ここで ∂f は f の劣微分である。従って、

$$w^{t+1} - w^t \in \eta_t \partial f(w^{t+1}). \quad \cdots (15)$$

この式は劣勾配法 (subgradient method)[19] と非常に類似しているが、(劣) 勾配がステップ前の点 w^t ではなく、ステップ後の点 w^{t+1} で評価されていることに注意を要する。すなわち、式 (15) はステップ後の点 w^{t+1} に関する方程式であり、これを解く（式 (12) の最小化をする）ことで結果的に勾配方向へ進むことが実現される。また、ナイーブな (劣) 勾配法はステップサイズ η_t や劣勾配の選び方によっては目的関数 f が増加してしまうこともあるが、更新式 (12) からは $f(w^{t+1}) + \frac{1}{2\eta_t}\left\|w^{t+1}-w^t\right\|_2^2 \le f(w^t)$ より、明らかに $f(w^{t+1}) \le f(w^t)$ であり、等号成立は $w^{t+1}=w^t$ のときに限られることがわかる。さらに前節の IST とは異なり（理論的には）ステップサイズ η_t を

どのように取ってもこの単調減少性が成り立つことは注目に値する。

拡張ラグランジュ法としての提案手法

式 (13)(14) は式 (1) の双対問題に対する拡張ラグランジュ法 (augmented Lagrangian method)[17) 20)] と見ることができる。この際対象となるのは線形等式制約を持つ以下の双対問題：

$$\underset{\alpha,\, v}{\text{maximize}} \quad -L^*(-\alpha)-\phi_\lambda^*(v),$$
$$\text{subject to} \quad v = A^{\mathrm{T}}\alpha$$

であり、主変数 w^t はラグランジュ乗数の役割を果たす[16) 18)]。

下界の最小化としての提案手法

式 (12) で得られる Moreau´s envelope を以下のように $f_\eta(w)$ と書く：

$$f_\eta(w)=\min_{w'\in\mathrm{R}^n}\left(f(w')+\tfrac{1}{2\eta}\|w'-w\|_2^2\right).$$

定義より、$f_\eta(w)\le f(w)$。さらに $w^*=\underset{w\in\mathrm{R}^n}{\arg\min} f(w)$ とすると、

$$\min_{w\in\mathrm{R}^n} f_\eta(w)=\min_{w\in\mathrm{R}^n}\min_{w'\in\mathrm{R}^n}\left(f(w')+\tfrac{1}{2\eta}\|w'-w\|_2^2\right)$$
$$=\min_{w'\in\mathrm{R}^n} f(w')=f(w^*). \quad \cdots (16)$$

ここで式 (16) の被最小化関数が w と w' に関して同時凸であることを用いた。また、この同時凸性から $f_\eta(w)$ も下に凸な関数である。従って、$f_\eta(w)$ は任意の正の η に関して $f_\eta(w)$ の下界であり、最小を達成する w^* において $f_\eta(w)$ と一致する凸関数である。

具体例 ― マルチカーネル学習

マルチカーネル学習[5)] (multiple kernel learning; MKL) は複数のカーネルの線形結合によって得られるカーネル関数を用いた一種のカーネル学習法であり、結合カーネルに基づいて判別関数を学習するだけでなく線形結合の係数も同時に最適化する方法である。Bach ら[6) 21)] によって結合係数が非負の場合、MKL は式 (1) の特別な場合になることが示された。具体的に目的関数を書くと、

$$\underset{\beta_j\in\mathrm{R}^m,\, b\in\mathrm{R}}{\text{minimize}} \quad L\left(\sum_{j=1}^n K_j\beta_j+b\mathbf{1}\right)+\lambda\sum_{j=1}^n\|\beta_j\|_{K_j}. \quad \cdots (17)$$

ここで、m はサンプル数、n はカーネルの数とし、K_j ($\in R^{m\times m}$) は j 番目のカーネル行列とする。β_j は j 番目のカーネルに対する判別関数の重みベクトルであり、b はバイアス項、また、$\mathbf{1}$ はすべて 1 の m 次元ベクトルとする。さらに $\|\beta_j\|_{K_j}=\sqrt{\beta_j^{\mathrm{T}}K_j\beta_j}$ とする。ここで、式 (17) の正則化項は K_j で定義されるノルムの線形和である (2 乗和ではないことに注意。) 従って、ノルムがゼロとなる点で式 (2) と同様に微分不可能性がありスパースな解を生じうることがわかる。ただし、ここでのスパース性は式 (2) と異なり、カーネルごと (β_j ごと) に生じる。B を β_j を列方向に並べた $m\times n$ 行列とすると、上の最適化問題の最適解では B はランク 1 となることが知られている[6)]。

従って、$\theta_j=\frac{\|\beta_j\|}{\sum_j\|\beta_j\|}$ ($\|\cdot\|$は任意のノルム) とおき、$\beta_j=\theta_j\beta^*$と書くと、θ_j が線形結合の係数、β^* が結合カーネルに基づく判別関数の重みベクトルであることがわかる。

最適化問題 (17) は正則化されないバイアス項 b を含むため厳密には式 (1) とは微妙に異なる。その違いも式 (12) に基づくと系統的に取り込むことができる。すなわち、式 (12) の代わりに以下の更新式を考える：

$$(B^{t+1},b^{t+1})=\underset{B\in\mathrm{R}^{m\times n},\, b\in\mathrm{R}}{\arg\min}\left(f(B,b)+\sum_{j=1}^n\frac{\|\beta_j-\beta_j^t\|_{K_j}^2}{2\eta_j^t}+\frac{(b-b^t)^2}{2\eta_b^t}\right).$$

ただし、は式 (17) の目的関数とする。前節と同様に上の更新式も微分可能な補助目的関数の最小化を用いて書き直すことができて (導出は Suzuki & Tomioka[22)] を参照されたい)、以下の更新式を得る：

$$\beta_j^{t+1}=\mathrm{STK}_{\eta_j^t\lambda}^j\left(\beta_j^t+\eta_j^t\alpha^t\right),$$
$$b^{t+1}=b^t+\eta_b^t(\mathbf{1}^{\mathrm{T}}\alpha^t).$$

ただし、α^t は以下で定義される関数 φt (d) を最小化する：

$$\varphi_t(\alpha)=L^*(-\alpha)+\sum_{j=1}^n\frac{1}{2\eta_j^t}\left\|\mathrm{STK}_{\eta_j^t\lambda}^j(\beta_j^t+\eta_j^t\alpha)\right\|_{K_j}^2+\frac{1}{2\eta_b^t}\left(b^t+\eta_b^t(\mathbf{1}^{\mathrm{T}}\alpha)\right)^2. \quad \cdots (18)$$

また、Soft-Thresholding 関数 STK_λ^j を式 (4) を拡張して以下のように定義する：

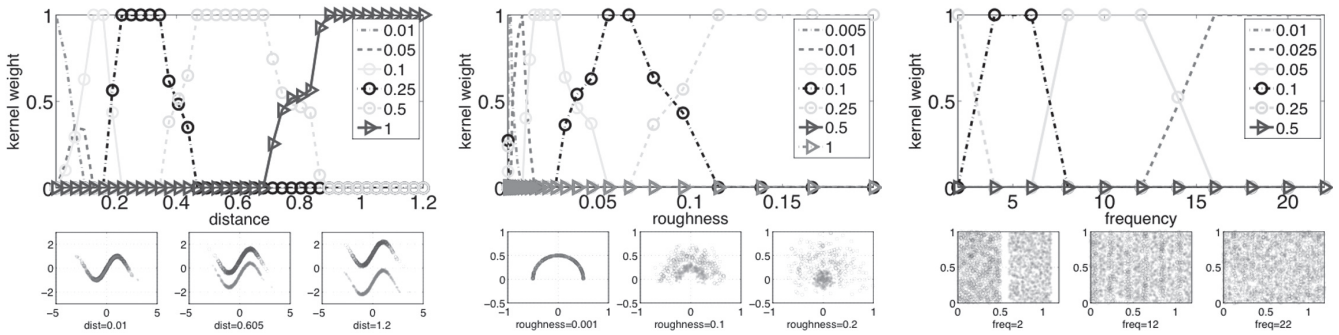
$$\mathrm{STK}_\lambda^j(\beta_j)=\max\left(\|\beta_j\|_{K_j}-\lambda,0\right)\frac{\beta_j}{\|\beta_j\|_{K_j}} \quad (j=1,\ldots,n).$$

式 (18) は一見複雑だが、右辺第 1 項は損失関数 L の凸共役であり、f_1 がサンプルの和に分解される場合、同様に n 項の和に分解されているので扱いやすい。また、第 2 項は Soft-Thresholding 関数 STK_λ^j の 2 乗であり、解がスパースな場合、n 項の和の中のアクティブな成分のみを考えればよいので、スパースさを生かした効率的な計算ができる。我々はこのアルゴリズムを SpicyMKL(Sparse Iterative MKL) と名付け、提案している[22)]。

数値実験

デモンストレーション

第 2 図に MKL を用いると問題の複雑さに合わせて適切なカーネルが選択できることを人工データを用いて示している。ここでは学習モデルとしていくつかの異なる幅を持つガウスカーネルの線形結合を考えている



第2図　人工データを用いたMKLのデモンストレーション。問題の複雑さにあわせたカーネルが自動的に選択できる。各図において横軸は何らかの複雑さの指標を表す。図の中にキャプションで示されたいくつかの幅を持つガウスカーネルを用いてMKLを行った。縦軸はそれぞれのカーネルのカーネル重み $\theta_j = \frac{\|\beta_j\|}{\sum_j \|\beta_j\|}$ を表し、それぞれの図の下側には、いくつかの複雑さの値での訓練データの分布する様子を示す。(a) クラス間分離性を変化させた場合。(b) クラス内局在性を変化させた場合。(c) 分離平面の複雑さを変化させた場合。

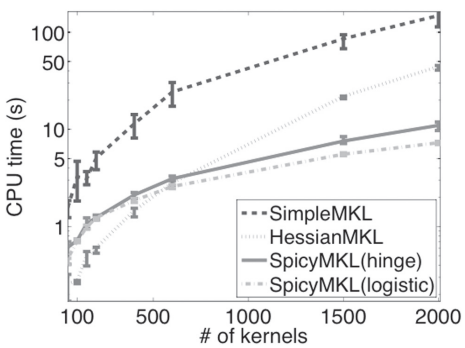
(式 (17) を参照)。損失関数はロジスティック損失関数とした[7]。

計算速度

第3図にSpicyMKLと既存の最適化アルゴリズム(SimpleMKL[21]とHessianMKL[23])の計算速度の比較を示す。SpicyMKLは常にSimpleMKLより速く、またカーネルの数nに対する計算時間の依存性では最も優れている。

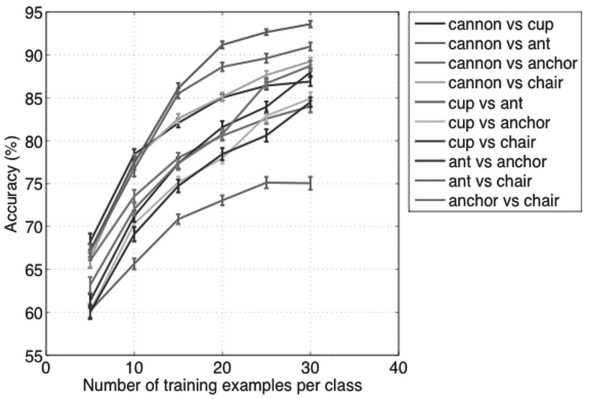
画像認識

画像認識のベンチマークデータCaltech101[24]を用いてマルチカーネル学習を用いた画像認識の性能を検証する。このデータセットには101種類のカテゴリの画像が多数集められているが、我々はanchor, ant, cannon, chair, cup の5つのクラスを利用し、10通りの2値分類問題に対する性能を報告する。我々の用いた特徴抽出法は3段階に分かれていて、その組み合わせで1760個のカーネル関数を生成する。第1段階は特徴点の抽出と量子化である。特徴点はすべての画像に対して規則的にグリッド状に収集した。これらの点に対して、hsvsift、sift（スケール自動)、sift（スケール4px)、sift（スケール8px）の4通りの特徴選択を行った[25]。さらに、すべての特徴点の中からランダムに200点を選んで代表点とし、すべての特徴点をこれらの代表点の最も近いものに量子化した。この時点で各画像は多数のvisual wordsの集まりとして表現されている。第2段階は領域分割である。我々は各画像全体、4分割、16分割したものおよびこれらの分割をspatial pyramid[26]を用いて統合したものの22通りの領域についてvisual wordsのヒストグラムを計算した。第3段階は上記のヒストグラム特徴の間の類似度の計算である。我々はガウシアンカーネルと*カイ二乗カーネル*の2種類のカーネル関数をそれぞれ10通りのバンド幅で計算した。結果的に1760(=4×22×20)個のカーネル関数が計算された。我々は認識対象となる2クラスそれぞれから各クラス5-30枚の画像をランダムに選んで訓練データとして、上述の1760個のカーネルを用いたマルチカーネル学習を行い、これを残りの画像でテストした。ここでは100回の平均テスト精度を報告する。



第3図　MKL最適化問題式(17)に対する計算時間の比較。縦軸は計算時間（秒）、横軸はカーネルの数(n)。

第4図に認識精度を示す。マルチカーネル学習を用いると各クラス30枚程度の画像からほとんどの場合で



第４図　Caltech101データセットに対する1760個のカーネルを用いたマルチカーネル学習の性能。縦軸は認識精度（%）、横軸は1クラスあたりの訓練画像の数。

80パーセント以上の精度で認識できることがわかった。また、スパース正則化のために非常にたくさんのカーネル関数の中から認識に有用なものが自動的に選択できることがわかった。実際、この実験では各クラス30枚の画像を使った場合、実際に用いられているカーネルの数は2~12個程度であった。

おわりに

この解説では proximal minimization[17] の枠組みに基づいて効率的なスパース正則化のためのアルゴリズム (SpicyMKL[22]) を導出し、既存のアルゴリズムとの関係やこのアルゴリズムの様々な解釈を解説した。また数値例では人工データおよび画像認識の実データを用いてSpicyMKLアルゴリズムの、とくにカーネルの数に対する効率の良さとマルチカーネル学習のカーネル選択に関する有用性を示した。

参考文献

1) R. Tibshirani, Regression shrinkage and selection via the lasso. J. Roy. Stat. Soc. B, 58(1):267-288, 1996.
2) S. Chen, D. Donoho and M. Saunders, Atomic decomposition by basis pursuit. SIAM J. Sci. Comput., 20(1):33-61, 1998.
3) I. F. Gorodnitsky and B. D. Rao, Sparse Signal Reconstruction from Limited Data Using FOCUSS: A Re-weighted Minimum Norm Algorithm. IEEE Trans Signal Process., 45(3):600-616, 1997.
4) B. Schölkopf and A. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond. MIT Press, Cambridge, MA, 2002.
5) G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui and M. I. Jordan, Learning the Kernel Matrix with Semidefinite Programming. Journal of Machine Learning Research, 5:27-72, 2004.
6) F. R. Bach, G. R. G. Lanckriet and M. I. Jordan, Multiple kernel learning, conic duality, and the SMO algorithm. In Proc. ICML 2004, page 6, New York, NY, USA, 2004. ACM.
7) C. M. Bishop, Pattern Recognition and Machine Learning. Springer, 2007.
8) S. Boyd and L. Vandenberghe, Convex Optimization. Cambridge University Press, 2004.
9) M. A. T. Figueiredo and R. D. Nowak, An EM algorithm for wavelet-based image restoration. IEEE Trans. Image Process., 12:906-916, 2003.
10) I. Daubechies, M. Defrise and C. D. Mol, An Iterative Thresholding Algorithm for Linear Inverse Problems with a Sparsity Constraint. Communications on Pure and Applied Mathematics, LVII:1413-1457, 2004.
11) P. L. Combettes and V. R. Wajs, Signal recovery by proximal forward-backward splitting. Multiscale Modeling and Simulation, 4(4):1168-1200, 2005.
12) W. Yin, S. Osher, D. Goldfarb and J. Darbon, Bregman Iterative Algorithms for L1-Minimization with Applications to Compressed Sensing. SIAM J. Imaging Sciences, 1(1):143-168, 2008.
13) S. J. Wright, R. D. Nowak and M. A. T. Figueiredo, Sparse Reconstruction by Separable Approximation. IEEE Trans Signal Process, 57(7):2479-2493, 2009.
14) J. J. Moreau, Proximité et dualité dans un espace hilbertien. Bulletin de la S. M. F., 93:273-299, 1965.
15) M. A. T. Figueiredo, J. M. Bioucas-Dias and R. D. Nowak, Majorization-Minimization Algorithm for Wavelet-Based Image Restoration. IEEE Trans. Image Process., 16(12):2980-2991, 2007.
16) R. Tomioka and M. Sugiyama, Dual Augmented Lagrangian Method for Efficient Sparse Reconstruction, IEEE Signal Process. Lett., 16(12): 1067-1070, 2009.
17) R. T. Rockafellar, Augmented Lagrangians and applications of the proximal point algorithm in convex programming. Math. of Oper. Res., 1:97-116, 1976.
18) R. Tomioka, T. Suzuki and M. Sugiyama, Super-Linear Convergence of Dual Augmented Lagrangian Algorithm for Sparse Learning. Arxiv:0911.4046, 2009.
19) D. P. Bertsekas, Nonlinear Programming. Athena Scientific, 1999. 2nd edition.
20) D. P. Bertsekas, Constrained Optimization and Lagrange Multiplier Methods. Academic Press, 1982.
21) A. Rakotomamonjy, F. R. Bach, S. e. Canu and Y. Grandvalet, SimpleMKL. Journal of Machine Learning Research, 9:2491-2521, 2008.
22) T. Suzuki and R. Tomioka, SpicyMKL. Arxiv: 0909.5026, 2009.
23) O. Chapelle and A. Rakotomamonjy, Second order optimization of kernel parameters. In NIPS 2008 Workshop on Kernel Learning, 2008.
24) L. Fei-Fei, R. Fergus and P. Perona, Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In Proc. IIEEE CVPR 2004, Workshop on Generative-Model Based Vision, 2004.
25) K. E. A. van de Sande, T. Gevers and C. G. M. Snoek, Evaluating Color Descriptors for Object and Scene Recognition, IEEE PAMI. 2010. In press.
26) S. Lazebnik, C. Schmid and J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In Proc. IEEE CVPR, 2006.

【筆者紹介】

冨岡亮太
東京大学　大学院
情報理工学系研究科 数理情報学専攻
助教
〒113-8656 東京都文京区本郷 7-3-1
TEL:03-5841-6898
FAX:03-5841-6897
E-mail: tomioka@mist.i.u-tokyo.ac.jp

鈴木大慈
東京大学　大学院
情報理工学系研究科 数理情報学専攻
助教
〒113-8656 東京都文京区本郷 7-3-1
TEL:03-5841-6909　FAX:03-5841-6909
E-mail: s-taiji@stat.t.u-tokyo.ac.jp

杉山将
東京工業大学 大学院
情報理工学研究科 計算工学専攻
准教授
〒152-8552 東京都目黒区大岡山 2-12-1-W8-74
TEL:03-5734-2699　FAX:03-5734-2699
E-mail: sugi@cs.titech.ac.jp

Keyword キーワード　機械学習，カーネル法，マルチカーネル学習，画像認識，正則化，スパース，特徴選択，L1