Vol.2015-AVM-88 No.1
2015/2/27

視線情報を用いたユーザ所望画像の識別に対するスパースコーディングの適用に関する検討

西口侑希[1,a)]　菅沼睦[2,b)]　亀山渉[3,c)]

概要：近年，ディジタル画像の増加に伴って画像検索技術の発展が期待されている．セマンティックギャップの解消に向けて，ユーザのフィードバックから学習を行い，出力結果をユーザが求める結果に近づけていく適合性フィードバックが注目されている．視線を用いたフィードバックが可能になれば，ユーザは画像を閲覧するだけでフィードバックを返すことができるだけでなく，ユーザ自身が意識しなかったような画像もフィードバックに用いることができる可能性がある．視線によるフィードバックを実現するためには，視線からユーザが所望する画像を識別する必要がある．これまで筆者らは画像への視線や瞳孔径から抽出された特徴量で SVM を構築して識別を試みてきた．本報告では，スパースコーディングを用いて識別することで，従来の SVM による識別よりも識別性能が向上したことを報告する．

On Applying Sparse Coding to Classification of Users' Desired Images with Eye Movements

Yuki Nishiguchi[1,a)]　Mutsumi Suganuma[2,b)]　Wataru Kameyama[3,c)]

Abstract: Digital image consumption has grown so rapidly, and image retrieval is expected to be more improved. To bridge the semantic gap which is an underlying problem of image retrieval, relevance feedback can be one of the solutions. Relevance feedback with eye-movements allows users to give their feedback just by watching the images and, moreover may make it possible to select the images even though they don't recognize what they want. Feedback with eye-movements demands the reasonable classification of users' desired images with eye-movements. We have already reported the result of a SVM trained with the features extracted from eye-movements and pupil size toward each image. In this report, it is shown that we improve the classification performance with sparse coding.

1. はじめに

近年，大量のディジタル画像の取得や蓄積が容易になり，画像検索技術のさらなる発展が期待されている．しかしながら，人間が画像の意味内容を理解できるのに対して，コンピュータは色や形状といった低次元の特徴量でしか画像を扱うことができない．この画像に対する理解の差はセマンティックギャップと呼ばれ，ディジタル画像を扱う際の問題となっている [1].

セマンティックギャップを解消する方法の 1 つとして適合性フィードバック [2] が注目を集めている．これは，システムの検索結果に対してユーザがフィードバックを与えることで，システムは学習を行い，出力結果をユーザが求める結果に近づけていく手法である．適合性フィードバックを用いた画像検索システムを実現するためには，何らかの形態でユーザからシステムへ検索結果についてフィードバックを与える必要がある．しかしキーボードやマウス操作で画像を選択する方法では，ユーザは本来の結果画像の閲覧

[1] 早稲田大学大学院国際情報通信研究科
GITS, Waseda University
[2] 早稲田大学国際情報通信研究センター
GITI, Waseda University
[3] 早稲田大学基幹理工学部情報通信学科
Department of Communication and Computer Engineering, School of Fundamental Science and Engineering, Waseda University
[a)] y_nishiguchi@toki.waseda.jp
[b)] mutsumi@aoni.waseda.jp
[c)] wataru@waseda.jp

には不必要な行為を要求されてしまう．加えて，意識して選択した画像のみしかフィードバックに用いることができない．したがって画像閲覧中の視線からの適合性フィードバックでは，ユーザは画像を閲覧するだけでフィードバックを返すことができるだけでなく，ユーザ自身が意識しなかったような画像もフィードバックに用いることができる可能性がある．視線によるフィードバックを実現するためには，視線からユーザが所望する画像を識別する必要がある．

これまで筆者らは，所望画像と非所望画像に向けられる視線および瞳孔径は異なると考え，各画像への視線停留時間や停留回数などを特徴量として抽出し，SVM（サポートベクターマシン）による識別器を構築することで所望画像の識別を試みてきた [3]．しかしながら，実際の画像検索システムにおいて満足なフィードバックとするにはさらなる性能向上が必要である．本報告では，スパースコーティングによって抽出された特徴を SVM によって識別することで，従来の特徴抽出から構築された SVM による識別よりも識別性能が向上したことを報告する．

2. スパースコーディング

スパースコーディング [4] とは，式 1 のように，入力信号 $\mathbf{x}$ を基底ベクトル (辞書)$\mathbf{D}$ の少ない基底の線形和で近似表現する手法である．ここで $\boldsymbol{\alpha}$ はスパース係数と呼ばれ，基底ベクトルに対する重み値を表す．また，辞書 $\mathbf{D} \in \mathbb{R}^{m \times k}$ は，辞書サイズ k が信号の次元数 m より大きい，過完備基底とする．

$$\mathbf{D}\boldsymbol{\alpha} = \mathbf{x} \tag{1}$$

スパースコーディングでは，基底中の多くの係数が 0 となるため，信号を効率よく表現できる．また，あらかじめ規定された基底ではなく，学習した基底を用いるため，データにより適した柔軟な表現が可能である．さまざまな分野で応用されており，画像のノイズ除去 [5]，顔画像認識 [6]，音声信号処理 [7]，あるいは異常検知 [8], [9] などで有用性が示されている．

辞書学習にはさまざまなアルゴリズムが提案されているが，ここでは文献 [4] に基づいて滑らかな非凸目的関数を最適化する問題として考える．学習事例 $\mathbf{X} = (\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_n}) \in \mathbb{R}^{m \times n}$ に対して，コスト関数

$$f_n(\mathbf{D}) = \frac{1}{n} \sum_{i=1}^{n} l(\mathbf{x_i}, \mathbf{D}) \tag{2}$$

を最小化することを考える．ここで，l は損失関数で，$\mathbf{D}$ が $\mathbf{x}$ をスパースな形で表現できるのであれば，$l(\mathbf{x}, \mathbf{D})$ は小さくなる．通常，信号の次元数 m に対して，信号のサンプル数 n は大きくなる．$l(\mathbf{x}, \mathbf{D})$ をノルム l_1 のスパースコーディング問題の最適解と定義することで，式 3 を得る．

$$l(\mathbf{x}, \mathbf{D}) = \min \frac{1}{2} \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_1 \tag{3}$$

ここで λ は正則化パラメータであり，スパース性と誤差を調整している．この問題は Basic Pursuit[10] あるいは Lasso[11] などと呼ばれ，最終的に式 4 の最適化問題として定義できる．

$$\min \sum_{i=1}^{n} (\frac{1}{2} \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda \|\boldsymbol{\alpha}_i\|_1) \tag{4}$$

辞書 $\boldsymbol{D}$ の要素が極めて大きな値を持つ，すなわちスパース係数 $\boldsymbol{\alpha}$ が極めて小さな値を持つことがないよう，一般的に辞書 $\mathbf{D}$ の各列 $\mathbf{d}_1, \cdots, \mathbf{d}_k$ の l_2 ノルムを 1 以下にするなどの制約を加える (式 5)．

$$\mathbf{d}_j^t \mathbf{d}_j \leq 1 (j = 1, \cdots, k) \tag{5}$$

本研究では，辞書のスパース性を高めることができるとされる，Elastic-Net 制約 (式 6) を用いた．

$$\|\mathbf{d}_j\|_2^2 + \gamma \|\mathbf{d}_j\|_1 \leq 1 (j = 1, \cdots, k) \tag{6}$$

ここで γ は辞書 $\mathbf{D}$ のスパース性を調整するパラメータである．

3. 提案手法

3.1 入力信号

各画像に対して，入力信号 $\mathbf{S} = (\mathbf{s_1}, \mathbf{s_2}, \cdots, \mathbf{s_i}, \cdots, \mathbf{s_n})$ を算出した．ここで，n は各画像に視線が落ちた時間長に相当する．「画像に視線が落ちる」とは 60[Hz] で計測された視線の位置 (視点) が対象画像内に含まれている状態を指す．

画像に視線が落ちた時刻 t_i に対して，t_i 時の視点と，t_i から前後 0.5[s] 間の各視点とのユークリッド距離を算出し，信号 $\mathbf{s_i}$ とした．すなわち，時刻 t_i での視点の座標を (x_i, y_i) と表し，時刻 t_i から前後 $\frac{1}{60}$[s] ごとの時刻をそれぞれ $t_{i-1}, t_{i-2}, \cdots$ および $t_{i+1}, t_{i+2}, \cdots$ で表すと，時刻 t_i に対する入力信号 $\mathbf{s_i}$ は

$$\begin{pmatrix} \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2} \\ \sqrt{(x_i - x_{i-2})^2 + (y_i - y_{i-2})^2} \\ \vdots \\ \sqrt{(x_i - x_{i-30})^2 + (y_i - y_{i-30})^2} \\ \sqrt{(x_i - x_{i+1})^2 + (y_i - y_{i+1})^2} \\ \sqrt{(x_i - x_{i+2})^2 + (y_i - y_{i+2})^2} \\ \vdots \\ \sqrt{(x_i - x_{i+30})^2 + (y_i - y_{i+30})^2} \end{pmatrix} \tag{7}$$

で表される 60 次元の信号となる．この 60 次元の信号を各画像に視線が落ちたすべての時刻に対して算出し，信号 $\mathbf{S}$ とした．

なお，ページの切り替えの直前や直後で，前後それぞれ 0.5[s] の範囲内に同一ページ内で視線が計測できない場合，視点間の距離は 0 とした．

3.2 所望画像の識別

図 1 に識別フローを示す．視線が画像範囲内に観測されなかった画像については，学習および識別に用いない．

3.2.1 学習

まず，学習データ内の正解画像から得られた入力信号を用いて辞書 **D**(基底ベクトル) を学習させる．つぎに，Lasso[11] を用いて，学習データ内のすべての画像から得られた入力信号の，辞書 **D** に対するスパース係数を算出する．各画像に対する入力信号およびスパース係数の数は，その画像に視線が落ちた時間に比例する．そこで $\frac{1}{60}$[s] ごとの入力信号に対する各スパース係数を SVM で学習する．

3.2.2 識別

学習された辞書 **D** を用いて検証データ内のすべての入力信号に対するスパース係数を算出する．スパース係数を SVM で識別し，その識別結果を画像ごとに多数決することで，画像の識別結果とする．

4. 評価実験

提示された複数枚の画像から所望画像を選択するときの視線を取得するために実験を行った．

4.1 実験装置

実験装置を図 2 に示す．画像を表示するためにナナオ社 23 インチ LCD ディスプレイ (FS2332) を用いた．ディスプレイの解像度は 1080×1920[px] である．視線は Tobii Technology 社アイトラッカ Tobii X60 を用いて 60[Hz] で計測した．なお，被験者は画面からおよそ 60[cm] の位置

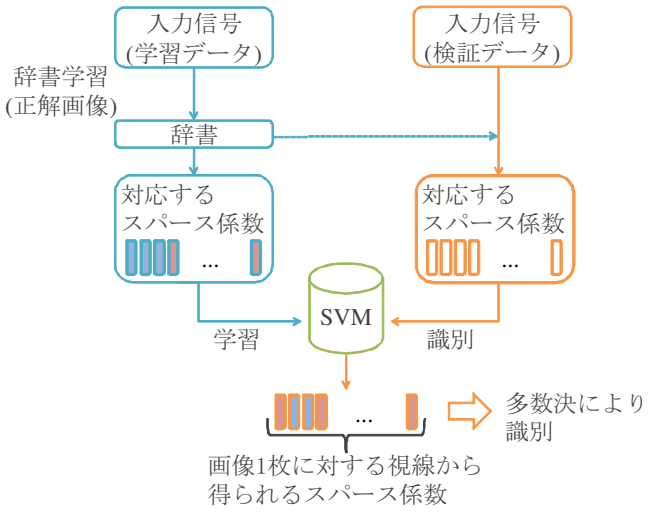

図 1: スパースコーディングによる識別フロー

図 2: 実験装置

に着席した．

4.2 実験内容

実験手順を図 3 に示す．まず，被験者にクエリ画像 1 枚を画面の中央に 2 秒間提示した．続いて提示される 8 枚のタスク画像群からクエリ画像に最も類似していると思う画像を 1 枚選択させた．次に，選択した画像をキーボードのテンキーで対応する位置の番号を押下することで回答させた．ここでクエリ画像を疑似的な所望画像とみなしている．

画面内での各タスク画像の表示位置は固定され，いずれの被験者に対しても同じように提示したが，各試行画面の提示の順序は被験者ごとにランダムに決定した．

なお，実験のタスクに積極的に参加することを促し，一般的な画像検索時の状況を再現するため被験者にキーの押下を要求したが，識別に被験者の回答は使用していない．

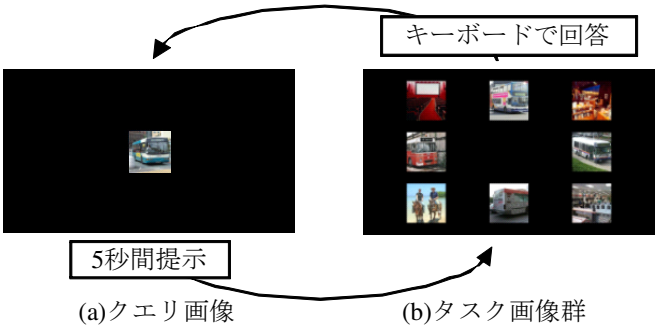

図 3: 実験手順

4.3 使用画像

実験には，Pascal VOC data sets (The VOC2012 Challenge)[12] および SUN Database: Scene Categorization Benchmark[13] 内の画像を使用した．

各データセットの画像のサイズや縦横比は統一されていなかったが，サイズや縦横比によって画像の注目されやすさが生じないように，すべての画像サイズを 256 × 256[px] に統一した．本来の縦横比が 1:1 でなかったものについては，縦横比を変化させることで画像の印象が異なってしまうことのないよう，縦横比が 1:1 になるようにトリミングを行ってから，画像サイズを変化させた．トリミングの範

囲は，画像のカテゴリ内容が変わることがないように，筆者が判断して決定した．なお，実験時に 1 枚誤って縮尺が他の画像よりも小さいまま表示されてしまったが，実験結果に大きく影響を与えるものではないと判断した．

4.4 実験構成

実験は 6 つのタスクで構成した．6 種類のクエリ画像を用意し，1 つのタスク内で 1 種類のクエリ画像を提示した．クエリ画像のカテゴリは，“bus”，“canal urban”，“igloo”，“phoning”，“riding horse”，“running” であった．各タスクにおいて，クエリ画像と同一カテゴリの画像および同一カテゴリの画像とみなした画像を正解画像 (Ground Truth) とした．試行回数は各タスク 50 回とした．また，正解画像の枚数は，各クエリに対して 129 枚，合計 774 枚であった．各タスクは被験者ごとにランダムに実施した．

“running” をクエリ画像としたときの視線を検証データとし，その他の 5 つのカテゴリをクエリ画像としたときの視線を学習データとして識別器の構築 (辞書学習) に用いた．

4.5 被験者

被験者は，男性 3 人，女性 2 人の合計 5 人 (平均年齢 23.80 才，標準偏差 2.71) であった．

4.6 実装およびパラメータ

SVM およびスパースコーディングの実装環境には，それぞれ R[14] パッケージ kernlab[15] および SPAMS[16] を用いた．SVM のカーネルにはラジアル基底関数カーネル (RBF カーネル) を用いた．

各パラメータは，辞書サイズ $k = 80$，Elastic-Net 制約のパラメータ $\gamma = 0.3$ とした．正則化パラメータ λ は被験者 1,2,4,5 では $\lambda = 10.0$，被験者 3 では $\lambda = 5.0$ とした．これは，被験者 3 は $\lambda = 10.0$ では辞書 $\mathbf{D}$ の要素がすべて 0 となり，辞書学習ができなかったためである．

5. 結果と考察

図 4 から図 8 に，被験者 5 人の従来の SVM を用いたとき，ならびに，スパースコーディングを用いたときの ROC 曲線を示す．なお，SVM の特徴量には [3] で述べた 14 次元特徴量を利用した．視線停留が検出されなかった画像は学習および識別には用いなかった．また，図 9 および図 10 に従来の SVM を用いたとき，ならびに，スパースコーディングを用いたときの適合率，再現率および F-値の 5 人の平均を示す．横軸は，識別関数の値に対して，正解・非正解の予測クラスを分離するしきい値である．

図 4 から図 8 より，被験者 4 を除いて，SVM を用いた識別よりも識別性能の向上を確認した．特に，被験者 1 から 3 では ROC 曲線が著しく改善された．

図 10 および図 9 より，スパースコーディングを用いた場合の方が高い F-値を得られたことがわかる．さらに，最も高い F-値が得られるときの適合率および再現率の差を比較すると，SVM を用いた場合よりもスパースコーディングを用いた場合の方が小さく，適合率と再現率の偏りが少ない識別ができた．

スパースコーディングを適用することで，SVM の学習に利用した特徴量よりも特徴的なパターンを学習できたと考えられる．しかし，本報告で設定した辞書サイズなどの各パラメータの値が最適である保証はない．また，被験者 3 から得られたデータにおいて辞書学習ができない正則化パラメータが存在した理由は明らかになっていない．

被験者 4 については SVM を用いた場合と同じく，スパースコーディングでも識別できていない．これは，被験者 4 の視線が観測された画像は限定的であったため，被験者 4 の視線からは正解画像と非正解画像を識別するための特徴的なパターンが抽出できなかったと考えられる．また，観測された視線が少なく，学習に必要なデータが十分ではなかったことも原因と思われる．

6. まとめと今後の課題

本報告では，スパースコーディングを用いて，所望画像の識別を試みた．スパースコーディングを適用することで，従来の SVM を用いた手法より識別性能を向上させることができた．

筆者らは，瞳孔径変動を特徴量として用いることを提案している [3]．本報告の提案手法の入力信号は時系列で観測された視点間の距離に基づいており，反応遅延を考慮する必要がある瞳孔径は用いなかった．そのため，瞳孔径変動の持つ情報は反映することができなかった．今後は，瞳孔径変動を入力信号として扱うことができるか検討したい．また，実際の画像検索への適用に向けて，処理時間や必要な精度を検討する必要がある．

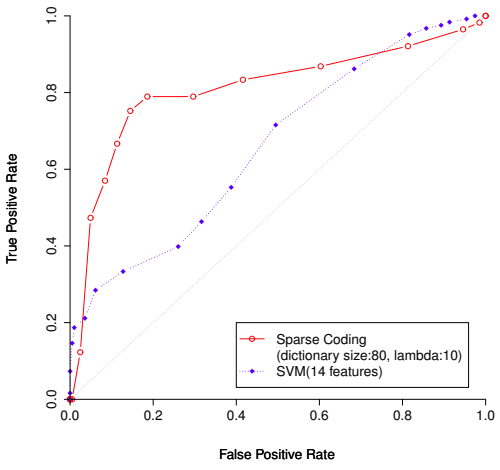

図 4: ROC 曲線 (被験者 1)

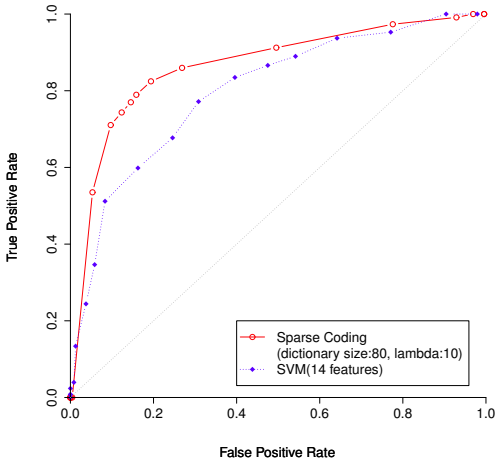

図 5: ROC 曲線 (被験者 2)

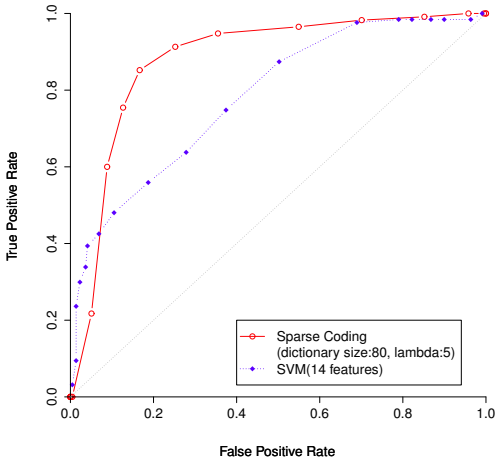

図 6: ROC 曲線 (被験者 3)

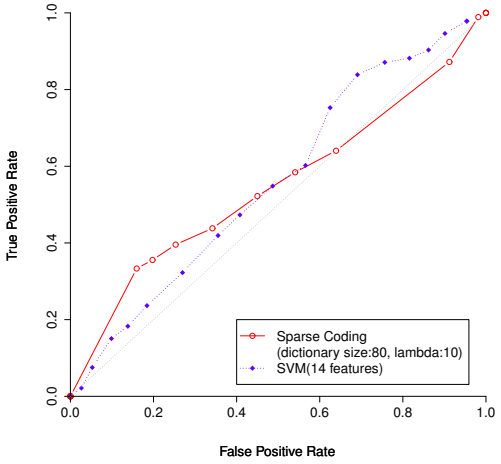

図 7: ROC 曲線 (被験者 4)

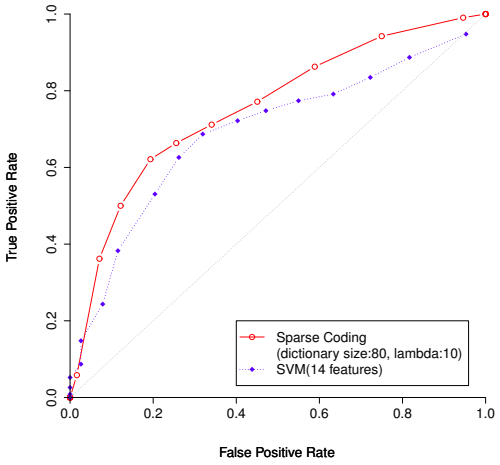

図 8: ROC 曲線 (被験者 5)

謝辞

本研究の一部は，JSPS 科研費 25330136 の助成による．

参考文献

1) Alan Hanjalic, Rainer Lienhart, Wei-Ying Ma and Jonh R. Smith, “The Holy Grail of Multimedia Information Retrieval: So Close or Yet So Far Away?”, Proceedings of the IEEE, Vol.96, Iss.4, 2008.
2) Xiang Sean Zhou and Thomas S. Huang, “Relevance feedback in image retrieval: A comprehensive review“, Multimedia Systems, Vol.8, Iss.6, pp.536-544, 2003.
3) 西口侑希, 菅沼睦, 亀山渉, “視線停留発生時刻の分散を用いたユーザ所望画像識別の検討”, 第 13 回情報科学技術フォーラム, H-035, 2014.
4) Julien Mairal, Francis Bach, Jean Ponce and Guillermo Sapiro, “Online Learning for Matrix Factorization and Spase Coding”, Journal of Machine Learning Research, Vol.11, pp.19-60, 2010.
5) Julien Mairal, Michael Elad, and Guillermo Sapiro, “Sparse Representation for Color Image Restoration”, IEEE Transactions on Image Processing, Vol.17, Iss.1, pp.53-69, 2008.
6) John Wright, Allen Y. Yang, Arvind Ganesh, S. Shankar Sastry and Yi Ma, “Robust Face Recognition via Sparse Recognition”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.31, Iss.2, pp.210-227, 2009.
7) Cédric Févotte, Nancy Bertin and Jean-Louis Durrieu, “Nonnegative Matrix Factorization with the Itakura-Saito Divergence: With Application to Music Analysis”, Neural Computation, Vol.21, No.3, pp.793-830, 2009.
8) Bin Zhao, Li Fei-Fei and Eric P. Xing, “Online Detection of Unusual Events in Videos via Dynamic Sparse Coding”, IEEE Conference on Computer Vision and Pattern Recognition, pp.3313-3320, 2011.
9) Amir Adler, Michael Elad, Yacov Hel-Or and Ehud Rivlin, “Sparse coding with anomaly detection”, IEEE International Workshop on Machine Learning for Signal Processing, pp.1-6, 2013.
10) Scott Shaobing Chen, David L. Donoho and Michael A. Saunders, “Atomic Decomposition by Basis Pursuit”, SIAM Journal on Scientific Computing, Vol.20, Iss.1, pp.33-61, 1998.

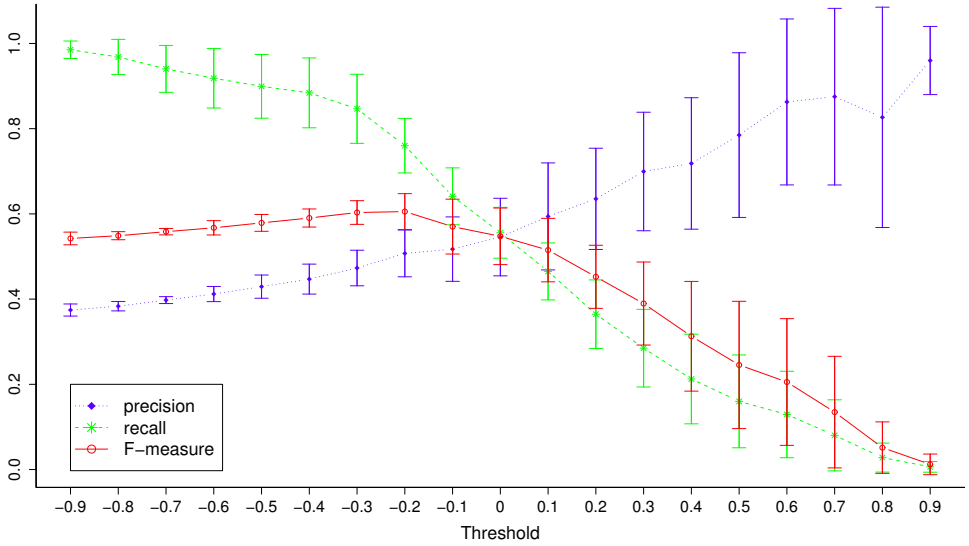

図 9: 適合率・再現率・F-値 (SVM)

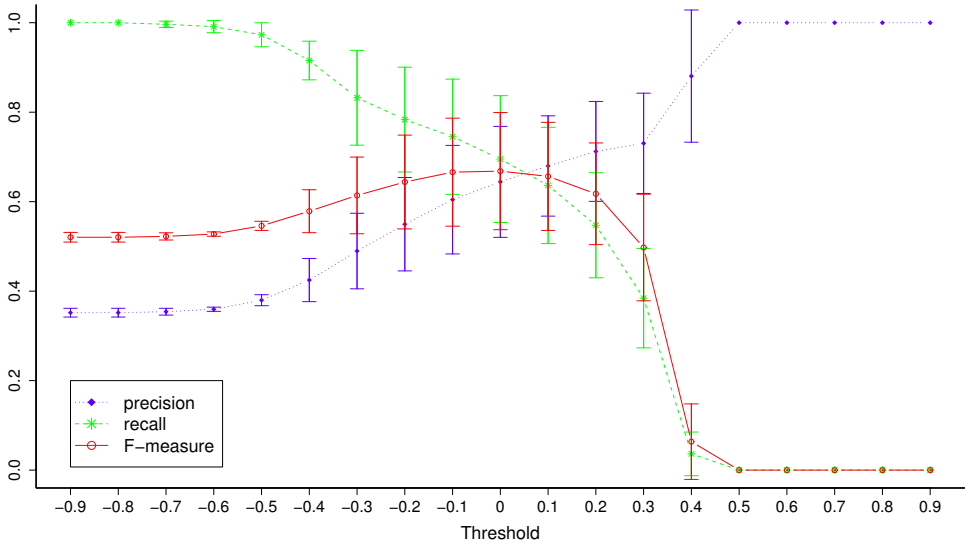

図 10: 適合率・再現率・F-値 (スパースコーディング)

11) R. Tibshirani, “Regression shrinkage and selection via the lasso”, Journal of the Royal Statistical Society, Series B(Methodological), Vol.58, No.1, pp.267-288, 1996.
12) http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2012/index.html (2014 年 12 月 21 日最終確認)
13) J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torrlba, “SUN Database: Large-scale Scene Recognition from Abbey to Zoo”, IEEE conference on Computer Vision and Pattern Recognition, pp.3485-3492, 2010.
14) http://www.r-project.org/(2015 年 1 月 16 日最終確認)
15) cran.r-project.org/web/packages/kernlab/kernlab.pdf (2015 年 1 月 6 日最終確認)
16) http://spams-devel.gforge.inria.fr/doc-R/html/index.html(2015 年 1 月 8 日最終確認)