6R-3

SVM学習のための データサンプリング法の検討

高橋洸 *　松本一則 †　橋本和夫 ‡

* 東北大学工学部情報知能システム総合学科　†KDDI 研究所
‡ 東北大学大学院情報科学研究科

1　はじめに

サポートベクターマシン (SVM) は，各事例にクラスが割り当てられているようなデータ集合を構造的損失最小化の視点からクラスごとに分類する学習機械である．カーネル法を適用することにより非線形分類問題も解くことが可能であり，その優れた分類性能と汎化性能から画像検出や手書き文字認識などパターン認識の分野でよく用いられる．しかしSVM学習を行うには二次計画問題を解く必要があるため，大規模データを適用させる際には計算量が膨大になってしまうという欠点が知られている．この問題を解決するアプローチとしては，主に二次計画問題を解くアルゴリズムを改善する手法と，データを能動的にサンプリングし学習を進めていく手法の二種類がある．前者のアルゴリズム改善手法としては，Platt による SMO(Sequential Minimal Optimization) 法 [2] が特に有名で実装にもよく用いられている．本論文では，後者の能動的なデータサンプリング戦略について検討を行うため，直観的な検証手法を用いていくつかのデータセットに対し分類実験を行う．

2　検証手法

本論文で検証を行うデータサンプリング戦略は **Algorithm1** のように表される．この手法は，データセットからランダムサンプリングを行って得た複数個のデータセットをそれぞれSVMで学習させ，最もテストデータを正しく分類できたデータセットを逐次選択していくというものであり，ノイズの混入を抑え，正解率の安定性を高めることを目的としている．

Algorithm 1 検証用データサンプリング戦略

1: 元データセット O から n_1 個をランダムにサンプリングしたデータセットを m 個生成する．この各データセットを $A_{1i}(i = 1, 2, ..., m)$ とする．
2: $A_{1i}(i = 1, 2, ..., m)$ でSVM学習を行い，モデル M_{1i} を生成する．
3: $M_{1i}(i = 1, 2, ..., m)$ にテスト用データセット T を適用し正解率を求め，最も正解率の高いモデル M_{1max} を生成したデータセットを A_{1max} とする．
4: データセット $O \setminus A_{1max}$ から n_2 個をランダムにサンプリングしたデータセットを m 個生成する．この各データセットを $A_{2i}(i = 1, 2, ..., m)$ とする．
5: $A_{1max} \cup A_{2i}(i = 1, 2, ..., m)$ でSVM学習を行い，モデル M_{2i} を生成する．
6: 3: から 5: を同様に繰り返す．

3　実験

いくつかのベンチマークデータセットを用いて検証実験を行った．使用したデータセットは表1の通りである．

Dataset	Train	Test	Dims
2-moon	2000	10000	2
text	1896	50	7510
MNIST ‘3’vs‘8’	11982	1984	780

表 1: 実験に用いたデータセット

本実験では，それぞれのデータセットからランダムにサンプリングしたとき (random) の正解率と，検証手法を用いたとき (verification) の正解率を比較した．SVMの実装は Sinz らによる UniverSVM[3] を使用した．結果は図 1〜3 に示す．横軸はデータ数 (samples)，縦軸は正解率 (accuracy) であり，正解率は各データ数で 5 つのデータセットで求めた（すなわち **Algorithm 1** で $m = 5$ とした）ものの平均をとってある．また横軸は対数目盛としている．

A Study of Data Sampling Method for SVM Learning
Takeru TAKAHASHI*, Kazunori MATSUMOTO†and Kazuo HASHIMOTO‡
* Department of Information and Intelligent Systems, School of Engineering, Tohoku University
†KDDI Institute
‡Graduate School of Information Sciences, Tohoku University

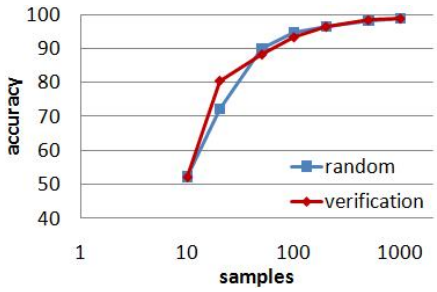

図 1: 2-moon を用いた分類実験

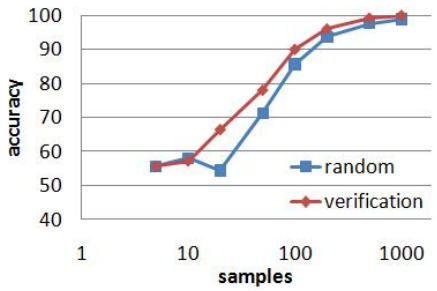

図 2: text を用いた分類実験

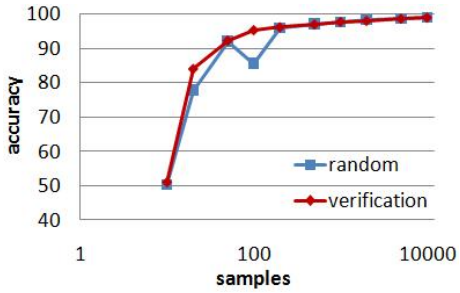

図 3: MNIS '3'vs'8' を用いた分類実験

図 1 の 2-moon での実験では，全体的にはほぼ同じ正解率となっているがデータ数 20 で改善がみられる．データ数が少なく正解率が安定していないときに，テストデータを良く分類する点をサンプリングできていると言える．しかし正解率が安定してくると，その中から最も良いデータセットを選んでも正解率は改善されないことがわかる．図 2 の text での実験では全体的に正解率が改善されている．ここでは 2-moon での実験と比べると正解率が安定してくるのが遅いため，学習の初期で良いデータセットを選べたことが学習後期まで良い影響を与えていると考えられる．図 3 の MNISI '3'vs'8' での実験では，正解率は部分的に改善されている．データ数 100 にて単純なランダムサンプリングでの正解率が 50 のときよりも落ちているが，データ数の増加に従って生じたノイズの影響を受けていると考えられ，検証手法ではノイズに対する耐性が強化されているといえる．

4 考察

第 3 章の実験結果から，検証手法が単純なランダムサンプリングよりも良い性能を示すことがわかった．しかし学習の初期では正解率の改善がみられたが，学習がある程度進むと正解率のゆらぎが少なくなるため改善はほとんど見られなくなった．本手法のように学習の初期に有効となるように学習済みのデータに加えデータセット全体から満遍なくサンプリングを行う手法としては，学習済みのデータから離れたデータからサンプリングしていく KFF(Kernel Farthest Fast) 法 [1] などがある．また，学習がある程度進んでから正解率を高めていく手法としては，生成した判別面により近いデータからサンプリングしていく Simple 法 [4] などが知られる．これらの手法と比較し本検証手法は学習結果のフィードバックが次の学習にあまり生かされておらず，安定性に欠けるといえる．これを踏まえ本検証手法を改善するには，正解率の不安定さに応じてサンプリングするデータ数 n やサンプリングを行う回数 m を変化させるといったことが有効だと考えられる．

5 まとめ

本研究では，直観的な能動データサンプリング戦略を用いて SVM によるデータセット分類実験を行った．本検証手法は単純なランダムサンプリングよりも良い性能を示したが，その改善は非常に僅かなものに留まった．これは検証手法が学習状況のフィードバックが不足していたためであり，今後はより効率の良いサンプリング戦略を構築するため，学習状況をいかに有効活用していくか研究を進めていきたい．

参考文献

[1] Y. Baram, et al., "Online Choice of Active Learning Algorithms", Journal of Machine Learning Research, vol.5, pp.255-291, 2004.

[2] J. Platt, "Fast Training of Support Vector Machines Using Sequential Minimal Optimization", pp.185-208, MIT Press, 1999.

[3] F. Sinz, et al., UniverSVM, `http://www.kyb.tuebingen.mpg.de/bs/people/fabee/universvm.html`

[4] S. Tong and D. Koller, "Support Vector Machine Active Learning with Applications to Text Classification", Journal of Machine Learning Research, pp.999-1006, 2000.