E-9 SVMとクラスタリングを用いた文書分類のための能動学習
Active Learning for Text Classification using SVMs and Clustering

佐々木 寛† 高村 大也† 松本 裕治†
Hiroshi Sasaki Hiroya Takamura Yuji Matsumoto

1. まえがき

近年，機械学習に基づく文書分類方法が盛んに研究されている．しかし，特に大規模な文書データベースに対して文書分類を機械学習法で行う際には，大量のラベル付けを行わなければ，高い精度を達成することは困難である．ところが大量のラベル付データを人手により作成することはコストが高い作業であり，このコストを軽減するためにラベル付データが少ない場合においても高い文書分類精度を実現することが望まれる．このような問題を解決する方法として本稿ではサポートベクターマシン (SVM) を用いた能動学習を行っている．さらにより効果的な学習データを選択する新しい方法として，SVMの分類結果に対してクラスタリングを適用した．

2. SVMを用いた能動学習

能動学習とはランダムに学習データを選択するのではなく，能動的にその時々に応じてできる限り有効性の高いデータを選択して学習を行う手法である．

一般にランダムに事例を選択した場合，選択された複数個の事例には冗長性がある場合が考えられる．しかし，能動的に事例を選択することにより，たとえ同じ数の学習データであっても情報量的には多くを学習したことになり，学習器の判定能力に改善をもたらす．

実際に，いかにしてこのようなデータを選択するかには，さまざまな議論がある [1]．しかし，SVMを用いる際にはそれまでの学習データによって構成された分離平面との距離が近い事例をそのような事例として扱うというヒューリスティクスを用いることができる．

Schohnら [3] はこのヒューリスティクスを用いて能動学習を行った．こうした方法においては理想的には一つずつ事例を追加して学習，分類を繰り返すのが望ましい．なぜなら，一度に複数個の学習データを選択するとそれらの中には部分的に同様の情報を含むことが考えられるからである．しかし実用を考慮した際，1事例ずつSVMにより学習，分類を行うのは時間的な制約から困難である．そのため，我々はできる限り多様性に富んだ複数個の学習データを選択的に学習データとして追加することが重要であると考えた．

本稿ではこのような目的のもとSVMの分類結果に対しクラスタリングを適用した．

3. クラスタリングを用いた学習データの選択

後の実験結果でも示されるが，ランダムに事例を追加していく手法は上記のヒューリスティクスを用いた場合に比べ学習のスピードが遅い．そのため，我々はこのヒューリスティクスの利点を持ち合わせたまま，なるべく多様性に富んだデータを学習データとして追加していくことを目指す．

このため，従来手法と同様に分離平面からの距離が近い順に複数事例選択し，それを特定数のクラスタにクラスタリングし，各クラスタからそのクラスタの中で最も分離平面から近いものを学習データとして追加した．

もし，最初に選択する個数が非常に大きな場合はその中に分離平面からの距離が大きい事例が含まれてくるため，ヒューリスティクスによる利点が生かせない．しかし，小さすぎる場合は学習データとして選択される事例の多様性が期待できない．そこで，本来は各カテゴリ中の文書集合中における文書間の差異に着目して最適な個数を求めるべきであると考えられるが，これは今後の課題として，今回は分離平面から近い順に50事例選択しそれを10クラスタに分類し，上記の方法により10事例を学習データとして追加した．

クラスタリングには Hierarchical Bayesian Clustering[2] を用いた．以下に Text Classification の手順を示す．

(初期設定)

1. 学習データの候補データセット P(全8815事例) からランダムに10事例 (正例を1つ負例を9つに固定) を選択し，それを学習用データセット I とする．
2. I を用いて学習を行う．$iteration = 0$ とする．

(繰返し)

3. $++iteration$ とする．
4. 現在の学習モデルにより P を分類する．
5. 4の分類結果において分離平面からの距離が最も小さい50事例を選択する．
6. 5の50事例を10クラスタにクラスタリングし，各クラスタの中で最も分離平面からの距離が小さい1事例，計10事例を選択しラベル付けを行った後 I に追加する．
7. I を用いてSVM学習を行う．
8. 7の学習モデルにより評価用データセット Q(全3023事例) を分類する．
9. $iteration \leq 20$ ならば3に戻る．$iteration > 20$ ならば終了．

4. 実験条件

実験に用いたデータは Reuters-21578 データセットである．そのうちの8815事例を上にあげたアルゴリズムに従って学習データの選択に用い，これらとは独立に3023

†奈良先端科学技術大学院大学 情報科学研究科

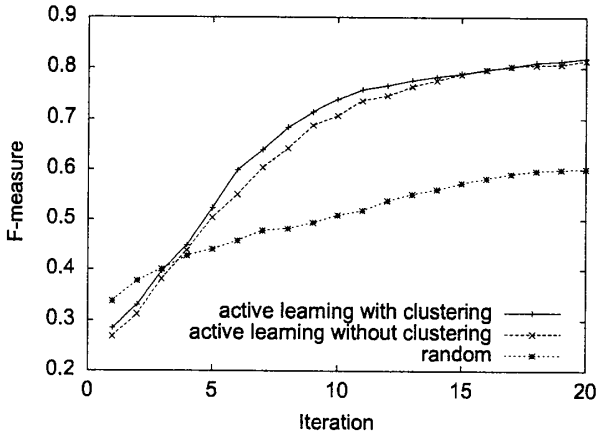

図 1: average of 10 categories.

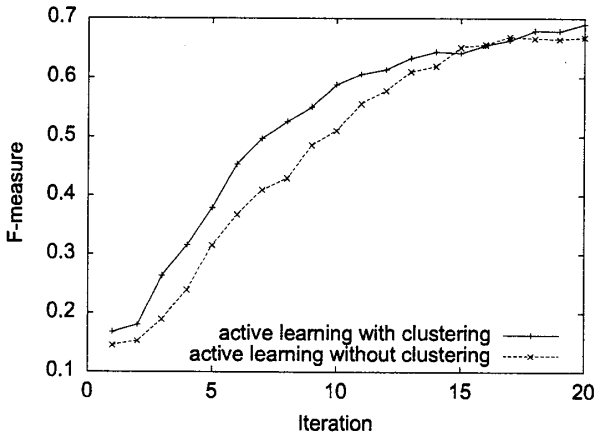

図 2: money-fx.

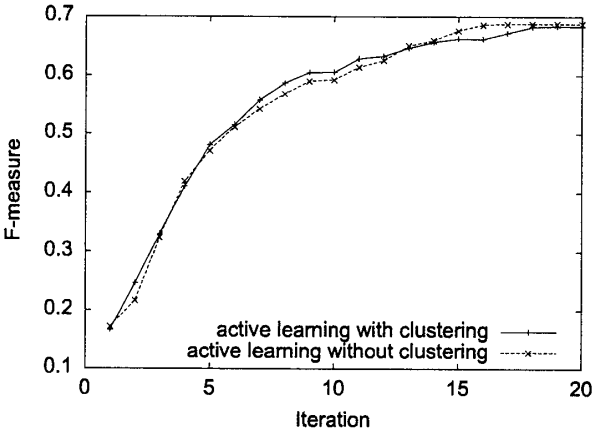

図 3: trade.

事例を評価に用いた.

データの前処理としては stemming を行った. この中から事例数の最も多い 10 のカテゴリ (earn, acq, money-fx, grain, crude, trade, interest, ship, wheat, corn) を対象として分類実験を行った. 各々のカテゴリに対し, 能動学習を用いずランダムに学習データをサンプリングした方法, クラスタリングを用いない能動学習手法, クラスタリングを適用した能動学習手法の三手法を比較した.

各学習ごとの追加学習データ数を 10 とし, それを 20 回繰返し各段階における F-measure の変化の様子を比較する. さらに, それぞれの実験を 20 回繰返し各々の実験で得られた各段階での F-measure の平均値を算出した. なお能動学習においても最初の 10 事例の予備学習データはランダムに収集されたものであるが, 二つの能動学習においては予備学習データは同一にした. SVM の学習には TinySVM[‡]を用いた. 用いたカーネル関数は線形カーネルである.

5. 実験結果

図 1, 図 2, 図 3 に実験結果を示す. 図 1 は 10 カテゴリの各段階における F-measure の平均値を, 上記三手法について求めたグラフである. 一回目の iteration 結果は既に一回, 能動学習が終了したことを表している. この結果から, 能動学習はいずれもランダムに学習データを追加した場合よりも F-measure が全体として良いことがわかる. しかし, 始めの 3 回の iteration までは逆にランダムに事例を追加する手法が最も高い値を示している. これは, 能動学習においてはクラスタリングを用いていない場合と用いた場合とを比較した結果, クラスタリングを用いることにより学習データ数が少ない段階において F-measure を改善することが確認された. このことから, 多様性に富んだ事例を学習データに追加するべきであるという考えが正しいことがわかる. 図 2 はクラスタリングの有無による差が顕著なカテゴリである"money-fx", 図 3 は差がほとんどない"trade"における結果を示したものである. これら二つの結果を比較すると我々の提案手法は必ずしも全ての場合において有効とは限らないが, 少なくとも幾つかのカテゴリに対しては有効であることがわかった. 今後の課題としてはこれら二つのケースにおいてどのような原因のため上記の結果をもたらしたかを究明し, クラスタリングが有効でない場合にこの方法以外の有効な手法を考えていきたい.

[‡]http://cl.aist-nara.ac.jp/~taku-ku/

6. おわりに

SVM を用いた能動学習のための学習データの選択方法に関してクラスタリングを適用する手法を提案し, その有効性について検討した. Reuters-21578 データセットを用いて実験を行いクラスタリングを用いていない場合と比較した結果, 学習データ数が少ない段階において提案手法が F-measure を改善することが確認された.

参考文献

[1] Campbell, C., Cristianini, N. and Smola, A. (2000). Query Learning with Large Margin Classifiers. *ICML2000*. pp. 111–118.

[2] Iwayama, M. and Tokunaga, T. (1995). Cluster-Based Text Categorization: A Comparison of Category Search Strategies. *SIGIR95*. pp. 273–280.

[3] Schohn, G. and Cohn, D. (2000). Less is More: Active Learning with Support Vector Machines, *ICML2000*. pp. 839–846.