クラスタリングによる文書分類用コーパスの構築手法の提案

柳原 正　　服部 元　　松本 一則　　小野 智弘
株式会社KDDI研究所
{td-yanagihara, gen, matsu, ono}@kddilabs.jp

1　はじめに

近年，大量に存在するウェブページを自動分類し，活用するサービスが増えている．たとえば，商品に関する意見を取り出す評判抽出 [1] や，有害文書の文書分類 [2] などのサービスが挙げられる．前者のサービスではユーザにとって有益と思われる情報を提示したり，後者のサービスではユーザにとって有害と思われる情報を取り除く機能を実現する．

上記で示したサービスでは，機械学習に基づく文書分類の技術を用いている．たとえば，ナイーブベイズ識別器 [3] やサポートベクターマシン (SVM)[4] のような識別器が著名である．このとき，高い分類精度を実現するためには，大量の学習データから構築されたコーパスを用意しなければならない．このようなコーパスは大量のラベル無しデータに対し，人手によるラベル付与 (アノテーション) を行う作業を通して実現される．このとき，アノテーションの量が多くなるに連れ，人的コストと時間が増大することが課題となる．

そこで，上記で述べたサービスを対象とした文書分類用コーパスを構築する際に，アノテーションの量を減らす手法として，クラスタリングに基づく能動学習を用いた文書分類用コーパスの構築手法を提案する．

2　関連研究

能動学習を用いることで，高精度の識別器を作成するために必要となるアノテーションの量を減らせることが従来より知られている．特に，SVMを使った能動学習 [5][6] が著名であり，これらの手法では，ラベル無しデータに対し，識別器によってラベルと識別面からの距離を求め，識別面に最も距離が近い事例をアノテーションの対象として選択する．これは，識別面付近の事例ほど，判定結果に対する信頼性が低く，このような曖昧な事例を正しく学習することによって得られる情報が多いという考え [7] に基づく．選択された事例はアノテーションが行われ，識別器を作成するために使用した初期の学習データと組み合わせ，新たな学習データを作成する．作成された学習データから識別器を作成し，ラベル無しデータを再度判定する．この過程を繰り返すことにより，高精度の識別器を構築するために必要となるアノテーションの量を減らすことが可能となる．

しかし，文献 [5] と文献 [6] の手法をウェブページの文書分類に適用した場合，精度向上の効果が小さくなる可能性がある．その理由として，ウェブページには類似する文書が数多く存在し，これによって識別面付近に多数の類似する事例が発生しやすくなるためである．文献 [8] では，この問題を解決するために，SVMにおける正例と負例のソフトマージンの間に存在する事例に対し，k-means 法 [9] によるクラスタリングを行い，抽出したクラスタの重心点に最も距離が近い事例をアノテーションの対象となる代表点として取り出す手法を提案している．

しかし，文献 [8] の手法には，次の2点の課題が存在する．

課題1　識別面をまたがるクラスタが形成される可能性があり，このときクラスタ内の事例が正例または負例に偏っているときに，クラスタを代表しない事例が代表点として選択されてしまう場合がある．

課題2　重心点に最も近い事例が識別面から離れている可能性があり，これによって得られる情報量が少ない事例が代表点として選択されてしまう場合がある．

3　提案手法

本稿では，2章で述べた課題1と課題2を解決する能動学習の手法を提案する．課題1に対し，SVMにおける正例のソフトマージンと識別面の間に存在する事例と，負例のソフトマージンと識別面の間に存在する事例のそれぞれに対し，クラスタリングを行うことで，クラスタを代表しない事例が代表点として選択されないようにする．課題2に対し，クラスタリングを

行う際に各事例に対し，識別面に近い事例ほど重みが高くなる重み係数をかけあわせることで，クラスタの重心点を識別面に近づけることを可能とする．これにより，得られる情報の量が多い事例が代表点として選択されやすくなる．

本手法の動作手順を以下に示す．提案手法が文献 [8] と比べて異なる点として，SVM の正例・負例におけるそれぞれのソフトマージンと識別面との間にある空間に存在する事例に対し，類似文書のクラスタリングを実施する際に，得られたクラスタの重心点の算出に重み係数を利用する点が挙げられる．

図 1 に提案手法の処理フローを示す．

Step 1 アノテーションされた学習データ L を用意する．
Step 2 識別器 C に対し，L を学習データとして用いる．
Step 3 C を用いて，ラベル無しデータ U に機械ラベルを付与する．
Step 4 機械ラベルが付与された U のそれぞれに対し，識別面からの距離を求め，正例と負例のそれぞれに対する帰属確率に変換し，事例ごとの重み係数を求める．
Step 5 SVM におけるソフトマージンと識別面の間に位置する正例／負例と判定された事例に対し，クラスタリングを行う．
Step 6 得られたクラスタの重心点に最も近い事例をアノテーション対象データ s とする．
Step 7 s に対し，アノテーションを行う．
Step 8 L と s を組み合わせ，識別器 C の新たな学習データとする．
Step 9 Step 3 に戻る．以降，終了条件を満たすまで繰り返す．(e.g. 繰り返しを一定回数行う)

図 1: 提案手法のフロー図

以降、Step 4 から Step 7 の詳細を述べる。

Step 4 の詳細説明

U のうちで機械ラベルが付与された事例のうち，任意の事例 X_i に対する識別面からの距離 d_i の求め方について説明する．

識別器 C は，識別関数 $f(X)$ により，事例 X ($X = \{X_1, X_2, ...X_N\}$) に対して判定結果であるラベル Y ($Y = \{-1, +1\}$) を決定する．識別器 C の判定結果を決定する識別関数 $f(X)$ は式 1 のように表せる．識別器 C が線型 SVM で作られる場合では，関数 $g(X)$ は式 2 として表せる．このとき，W は重みベクトルであり，b はバイアスを表す．これにより，任意の事例 X_i を与えたとき，式 3 によって d_i を求める．

$$f(X) = sign(g(X)) \tag{1}$$

$$g(X) = \langle W \cdot X \rangle + b \tag{2}$$

$$d_i = \frac{g(X_i)}{\|W\|} \tag{3}$$

Step 5 の詳細説明

事例 X のそれぞれに対する識別面からの距離 D をもとに，正例と負例のそれぞれへの帰属確率を求める．本稿では，シグモイド分布を仮定し，帰属確率への変換を行う手法 [10] を適用することで，識別面からの距離を正例に対する帰属確率 P^+ と負例に対する P^- に変換する．

$$P^+ = P(Y = +1|f(X)) \tag{4}$$

$$= \frac{\exp(Af(X) - B)}{1 + \exp(Af(X) - B)} \tag{5}$$

$$P^- = P(Y = -1|f(X)) \tag{6}$$

$$= \frac{1}{1 + \exp(Af(X) - B)} \tag{7}$$

事例 X_i における P^+ と P^- を求めたあと，X_i に対する重み係数 ω_i を求める．

$$\omega_i = \frac{1}{1 + \mathrm{argmax}\,|P(Y_i|f(X_i)) - 0.5|} \tag{8}$$

Step 6 の詳細説明

本稿では，アノテーションの対象データを選び出すために，SVM のソフトマージンと識別面の間に存在する事例に対してクラスタリングを行う．はじめに，$0 \leq g(x) \leq 1$ を満たす事例を取り出し，正例に帰属する曖昧な事例の集合を S^+ とする．次に，$-1 \leq g(x) \leq 0$ を満たす事例を取り出し，負例に帰属する曖昧な事例の集合を S^- とする．S^+ と S^- のそれぞれに対し，k 個のクラスタを抽出する．このとき，クラスタごとに含まれる事例 X に対し，事例に割り当てられた重み係数 ω_x を正規化した上で，各事例のベクトルにかけあわせる．以下に例を示す．

1. k 個のクラスタは $c_1, c_2, ...c_k$ とする．これらのうち，任意のクラスタ c_i を選択する．このとき，クラスタ c_i 内に含まれる事例は $X_1, X_2, ...X_{n_i}$ とし，それぞれの事例には重み係数 $\omega_1, \omega_2, ...\omega_{n_i}$ とする．

2. 重み係数 $\omega_1, \omega_2, ...\omega_{n_i}$ を正規化し，正規化された重み係数 $\omega'_1, \omega'_2, ...\omega'_{n_i}$ を求める．以下に任意の重み係数 ω_j から正規化された ω'_j を求めるための式を示す．

$$\omega'_j = \frac{\omega_j}{\sum_{l=1}^{n_i} \omega_l} \tag{9}$$

3. 求めた 各ω' と事例 x をかけあわせ，クラスタの重心点を求め，クラスタを更新する．以降，クラスタが収束するまで繰り返す．

Step 7 の詳細説明

クラスタを抽出した後に，重心点となるベクトルに対して最も近い事例を代表点とみなし，アノテーション対象データ s として取り出す．

取り出された s にアノテーションを行った後に，初期の学習データ L と組み合わせることで新たな学習データを作成する．新たな学習データを使い，識別器を生成し，ラベル無しデータ U を判定する．以降，手順を一定回数を満たすまで繰り返す．

4 評価実験

提案手法の精度を検証するため，有害文書の文書分類で精度評価を行う．以下に実施した評価実験の内容を示す．

4.1 比較対象

本稿では，比較手法として文献 [8] と同様に，正例のソフトマージンと負例の間のソフトマージンの間に存在する事例に対し，クラスタリングを行う手法 (以降，従来手法) との比較を行う．この他の実験条件は提案手法と同一の内容に揃える．

4.2 実験条件

実験データ

インターネットより約 300 万件のウェブページを自動収集し，文書に含まれる内容に基づき,「有害」「無害」のいずれかのラベルのアノテーションを行う．フィルタリング事業者が提供するカテゴリ表 [11] のうち，以下に示すカテゴリに該当する場合は「有害」とし，それ以外の場合は「無害」とした．このとき,「有害」ラベルの文書を正例とし,「無害」ラベルの文書を負例とした．

- 不法 (違法と思われる行為，違法と思われる薬物，不適切な薬物利用)
- 主張 (テロリズム・過激派，武器・兵器，告発・中傷，自殺・家出，主張一般)
- アダルト (性行為，ヌード画像，性風俗，アダルト検索・リンク集)

識別器

識別器としては，LIBLINEAR[12] を使用した．LIBLINEAR が提供する識別モードのうち，L2-regularized L2-loss support vector classification (dual) を使用した．なお，LIBLINEAR では標準のままでは，各事例における識別面からの距離が出力されないため，出力できるように修正した．

素性

素性は学習データ内に含まれる単語とし，これにより文書から単語を素性としたベクトルを作成した．単語を抽出するためには，形態素解析器 MeCab[13] の ver. 0.98 を用いた．形態素解析器の辞書としては，IPA 辞書-2.7.0 (20080701) に加え，学習データに出現する単語を含む 40 万語が登録された辞書を追加した．素性の選択方法として，赤池情報量基準に基づく特徴選択手法 [14] を用いた．これにより，正例のデータに関連する単語と負例のデータに関連する単語をそれぞれ約 4000 件ずつ抽出できた．

クラスタリング手法

クラスタリング手法として，repeated bisection[15] を用いた．これは，主要なクラスタリング手法である k-means 法と比べ，クラスタを抽出するための収束が早く終わるためである．クラスタ数は計 1000(正例と負例はそれぞれ 500 個ずつ) とした．

4.3 実験手順

アノテーション済みのウェブページ 300 万件から，正例のデータを 500 件，負例のデータを 500 件取り出し，初期の学習データとした．次に，正例のデータを 5,000 件，負例のデータを 5,000 件を取り出し，評価データとした．残りのデータ約 299 万件のうち，100 万件を取り出し，これらのデータのアノテーションされたラベルを取り除き，ラベル無しデータとした．これらのデータを使い，計 17 回の繰り返し学習を行った．1 回の学習ごとに，識別器によって正例と推定されたデータと負例と推定されたデータをそれぞれ 500 件取り出し，アノテーションを行ったあとに学習データに追加し，繰り返し学習を行う．これにより，学習データは最終的に 18,000 件となる．繰り返し学習を行う際に，評価データで適合率と再現率を求め，それらの調和平均である F 値を求める．また，繰り返し学習を 1 回実施するごとに，学習データで n 交差検定を行ったときに，F 値が最大となる SVM のコストの値を最適値として調整した．

提案手法として，以下の 2 つの手法を用いて精度比較を行う．提案手法 1 は，3 章で述べた提案手法である．一方，重み係数による精度向上への効果を検証するため，提案手法 1 から重み係数 ω を取り除いた手法 (以降，提案手法 2) と比較する．

提案手法 1　3 章で述べた提案手法
提案手法 2　提案手法に対し，重み係数 ω を使用せずにクラスタリングを行う手法

4.4　実験結果

以下の図 2 に評価結果を示す．提案手法 2 は従来手法と比べて大きな差がなく，学習データが 1000 から 20,000 の間では，F 値の差として提案手法 1 の方が平均で 0.0015 高かった．これは，課題 1 のようなケースが少なかったため，精度向上の効果が小さかったと考えられる．一方，提案手法 1 と従来手法を比べたとき，従来手法で追加されたアノテーション対象データが 15,000 件のときの F 値 (0.7984) は，提案手法 1 では 1000 件で得られることが分かった．さらに，学習データの件数が同一のときでも提案手法の精度が常に高く，学習データが 11,000 件のときに差が最大となり，0.04 であった．

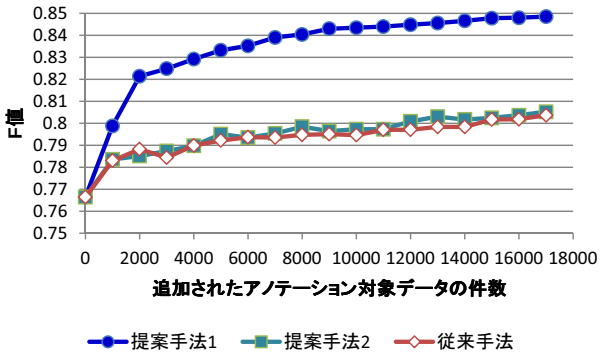

図 2: 提案手法と従来手法との精度比較

5　おわりに

本稿では，文書分類用コーパスを構築する際の人手によるラベル付与作業を軽減するため，クラスタリングに基づく能動学習手法を提案した．類似事例がアノテーション対象データとして選ばれてしまう課題に対し，本稿では SVM のソフトマージンと識別面の間に存在する事例をクラスタリングし，類似しない多種の事例が選ばれることを可能とした．また，クラスタリングを行う際に重み係数を導入することで，重心点に近い事例がアノテーション対象データとして選択されやすくした．評価実験では，提案手法と従来手法を比べたとき，従来手法で追加されたアノテーション対象データが 15,000 件のときの F 値 (0.7984) は，提案手法では 1000 件で得られることを確認した．今後，さらなる大規模のデータに適用し，提案手法の有効性を引き続き検証する．

謝辞

本研究は，(独) 情報通信研究機構の委託研究「インターネット上の違法・有害情報検出技術の研究開発」の一部として実施した．

参考文献

[1] “blogWatcher”, http://blogwatcher.pi.titech.ac.jp/
[2] 小林，松村，木戸，石塚．“知識検索サイトにおける不適切な投稿の分類．” 第 21 回人工知能学会全国大会, 2007.
[3] D.D. Lewis. “Naive (bayes) at Forty: The Independence Assumption in Information Retrieval,” pp.4-15, Springer Verlag, 1998.
[4] C. Cortes and V. Vapnik. “Support-Vector Networks”, Machine Learning, 20, 1995.
[5] G. Schohn, D. Cohn. “Less is More: Active Learning with Support Vector Machines.” Proceedings of the 17th International Conference on Machine Learning (ICML2000), pp. 839-846. 2000.
[6] S. Tong, D. Koller. “Support Vector Machine Active Learning with Applications to Text Classification.” Journal of Machine Learning Research, pp. 999-1006. 2000.
[7] D. D. Lewis and J. Catlett. “Heterogeneous Uncertainty Sampling for Supervised Learning.” In Proceedings of the 11th International Conference of Machine Learning (ICML’94), pp. 148-156, 1994.
[8] Z. Xu, K. Yu, V. Tresp, X. Xu, and J. Wang. “Representitive Sampling for Text Classification using Support Vector Machines.” In Proceedings of the 25th European Conference on IR Research (ECIR’03) pp. 393-407. 2003.
[9] G. H. Ball and D. J. Hall. “A Clustering Technique for Summarizing Multivariate Data.” Behavioural Science, Vol. 12, pp. 153-155.
[10] J. Platt. “Probablistic Outputs for SVMs and Comparisons to Regularized Likelihood Methods.” Advances in Large Margin Classifiers, MIT Press, 1999.
[11] “カテゴリ一覧 — 製品・サービス ネットスター株式会社” http://www.netstar-inc.com/product/category.html
[12] R. Fan, K. Chang, C. Hsieh, X. Wang, C. Lin. “LIBLINEAR: A Library for Large Linear Classification.” Journal of Machine Learning Research, Vol. 9 (Aug. 2008), pp. 1871-1874.
[13] “MeCab: Yet Another Part-of-Speech and Morphological Analyzer”, http://mecab.sourceforge.net
[14] 柳原，松本，小野，滝嶋．“トピック判定における n-gram の組み合わせ手法の検討” 第 7 回情報科学技術フォーラム (FIT2008), Vol. D, pp. 59-61, 2008.
[15] G. Karypis. “CLUTO: A Software Package for Clustering high-dimensional data sets.” Technical Report 02-017, University of Minnesota, Dept. of Computer Science. 2003.