SIRによるテキストの分類
—書き手別の分類を中心に—

金 明哲
札幌学院大学社会情報学部
jin@earth.sgu.ac.jp

1. はじめに

近年，データマイニングとテキストマイニングという言葉をよく耳にするようになった．データマイニングは数値を主とした定型(構造化された)のデータから,テキストマイニングは自然言語で書かれた非定型(構造化されていない)のテキスト(文書・章)から必要となる宝物・情報の掘り出しに関する技法及び行為の総称である．テキストマイニングは自然言語処理とデータマイニングの融合であると考えられる．

テキストマイニングは，テキストからの情報の抽出,テキストの分類,テキストの要約,テキストの検索などテキストの情報処理に関連するタスクである．テキスト分類(text categorization, text classification) は，事前に分類されたカテゴリにテキストを割り当てる問題である．テキスト分類研究が情報処理の分野で広く注目されるようになったのは，1990年代に入ってからである．

情報処理分野で言うテキストの分類は，内容による分類が一般的であるが，著者の同定問題では，文章を書き手別に分類を行う．本研究では，テキストを書き手別に分類するアプローチでテキスト分類について議論を行うことにする．著者の同定問題は百年前から研究が続けてきた．1950年代からはテキストを書き手別に分類するアプローチで文章の書き手を同定する研究が行われている．テキストを内容別に分類するか，書き手別に分類するかは，用いるテキストの特徴ベクトルが異なるだけであり，分類に用いる分類方法は基本的には同じである．

分類方法については，線形回帰モデル，事例ベースモデル，確率モデル,決定木およびルール抽出モデル,ニューラルネットワークモデル，帰納的学習モデル，サポートベクトルマシン方法など多くの分類方法が提案されている．

分類の精度は用いたコレクション及び特徴ベクトルと関係しているので，報告された分類精度に関する評価は絶対的なものではない．しかし，多くの研究では k 近隣法(k nearest neighbor)，SVM法，ニューラルネット法がよいパフォーマンスを示していると報告されている(Sebastiani 2002)．

これらの研究では，学習に用いたテキストの数は約1万前後で，テストに用いたテキストの数は3000〜6000に上る．しかし，書き手不明である文章の著者の同定問題では，学習及びテストに用いる文章は数十しかないのが現状である．このような少サンプルによるテキストの分類に関する比較研究は見られない．

本研究では，SIR(Sibsion's Information Radius, Simbson 1969)によるテキストの分類を試み，かつ広く知られているいくつの分類方法と比較を通じてその有効性を明らかにする．

2. データと分類手法

2.1 用いたデータ

本稿では，青空文庫の菊池寛，森鴎外，夏目漱石，島崎藤村，4人の合計80文章を用いた．特徴ベクトルは助詞の n-gram を用いた(金 1997, 2002a, 2002b, 2003)．本研究では，テキスト i における助詞の n-gram の j パ

ターンの相対頻度をp_{ij}とし，テキストiにおける助詞の n-gram のベクトルは

$$P_i(p_{i1}, p_{i2}, \cdots, p_{ij}, \cdots, p_{im})$$

と記する．ただし

$$\sum_{j=1}^{m} p_{ij} = 1$$

である．

2.2 分類の方法

分類は大きく教師データありの分類と，教師データなしの分類に分けられる．本研究では，教師データありの分類として距離モデル，ニューラルネットワーク，サポートベクトルマシンなどの結果と比較しながら，SIR に基づいたテキスト分類の有効性を実証する．本研究では，以下の方法について比較研究を行う．

- ✧ 距離・類似度による方法
  - ➢ ED(Euclidean distance)
  - ➢ SIR(Sibsion's Information Radius)
  - ➢ COS(Cosine Measure)
- ✧ k-NN(k-Nearest Neighbor)
- ✧ H-NNet(Hidden-layer Neural Network)
- ✧ LVQ(Learning Vector Quantization)
- ✧ SVM(Support Vector Machines)

2.2.1 距離・類似度による方法

判別・識別分析で，最も素朴な方法は距離や類似度による方法である．距離による判別分析では，判別すべきのベクトルと各グループgの中心ベクトルとの距離をD_gとした場合，

$$\min(D_1, D_2, \cdots, D_m)$$

となるグループに属すると判断る．

ここでは，ベクトル間の類似の測度としては SIR(Sibsion's Information Radius, Simbson 1969)と最も広く知られているユークリッド距離，COS の類似度を用いる．SIR は下記のように定義されている．

$$d_{ik} = \frac{1}{2}\sum_{j=1}^{m}\left(p_{ij}\log\frac{2p_{ij}}{p_{ij}+p_{kj}} + p_{kj}\log\frac{2p_{kj}}{p_{ij}+p_{kj}}\right)$$

$$p_{ij} = 0 \Rightarrow p_{ij}\log\frac{2p_{ij}}{p_{ij}+p_{kj}} = 0,$$

$$p_{kj} = 0 \Rightarrow p_{kj}\log\frac{2p_{kj}}{p_{ij}+p_{kj}} = 0$$

2.2.2 K-最近隣法

k-最近隣(k Nearest Neighbor)法は伝統的なパターン分類アルゴリズムである(Cover and Hart 1967)．k-最近隣法は，まず教師データから最も近いものをk個見つけ，その情報に基づいて，判別・識別すべきものとの距離を計算し，最も近いグループに割り当てる．距離の測度としては一般的にはユークリッド距離が使用されている．

2.2.3 ANN 方法

人間の脳で行う情報処理の仕組みをコンピュータで実現するため，人間の脳のニューロンを人口的にモデル化したものをニューラルネットワーク(ANN: Artificial Neural Network)という．ANN はパターン認識や予測などに広く用いられている．ANN にはさまざまなタイプがあるが本研究では，最も広く知られている隠れ層を持つニューラルネットワーク(H-Nnet: Hidden-layer Neural Network) と競合学習型の学習ベクトル量子化(LVQ : Learning Vector Quantization, Kohonen, T. 1996)を用いる．

2.2.4 SVM 方法

最近，サポートベクトルマシン(SVM : Support Vector Machin)が広く知られつつある．SVM はグループを分ける無限に存在する超平面のなか，最もグループわけが良い平面をマージン最大の基準で求める方法である．SVM が提案された当時は基本的には 2 クラス分類器であったが，現在は多クラスの分類器

に，また線形から非線形に拡張されている．

SVM をテキスト分類に適応した例としては Joachims(1998)，Dumais ら(1998)，Li and Yamanishi(1999)，Yang and Liu(1999)などがある．これらの研究では，*k*-NN 方法を含む比較に用いたその他の方法と比べ、そのパフォマンスが高く評価されている．

3. 結果の比較

3.1 学習とテスト

教師データありの分類の問題では，教師用とテスト用のデータが必要となる．大サンプルの場合は教師データとテストデータを分けて用いるが，小サンプルの場合は，N 分割交差検定・確認(N-fold cross-validation)の方法が多く用いられている．N 分割交差検定・確認法は，データセットをランダムに均等に N 等分に分割し，そのなかの N-1 等分を教師データとし，テストは学習に用いていない 1 等分を用いて行う．このようなテストをすべての N 等分について行う．

N 分割交差検定・確認法では，学習テストは N 回行い，結果はその平均値を用いる．このような方法は，サンプル数と分割数 N が小さいときにはランダムに分割を行う際の偏りの影響が大きい．そこで本研究では，教師データをデータセットからランダムに抽出し，残りをテストデータとし，学習・テストの作業を 1000 回繰り返し，その平均値を用いて評価することにした．本研究では，書き手別の 20 文章からそれぞれ 15 をランダムに抽出し教師データとし，それぞれの残り 5(合計 20)をテスト用のデータとした．書き手の判別は，判別すべき文章が複数の書き手の中，誰に帰属するかに関して判別する．そのパフォーマンスは

$$\text{判 別 率} = \frac{\text{正しく判別された数}}{\text{テストに用いた総数}}$$

$$\text{誤判別率} = 1 - \text{判別率}$$

を用いて評価した．

3.2 実験結果

助詞に関しては unigram と bigram のベクトルを用いた．また用いたデータは各変数の平均値を基準とし，降順にソートして用いた．

図 1 に unigram について上位 50 変数(ベクトルの長さ)を用いた場合とすべてを用いた場合に分けた実験結果を示す．横軸は変数の数，縦軸は誤判別率に 100 を乗じた値である．

図 1 では，上位 50 変数を用いた場合と，全部の変数を用いた場合との間には明らかな差が見られない．各手法を比較すると最も誤判別率が低いのが SIR で，その次が H-NNet 法，SVM 法，LVQ 法の順である．

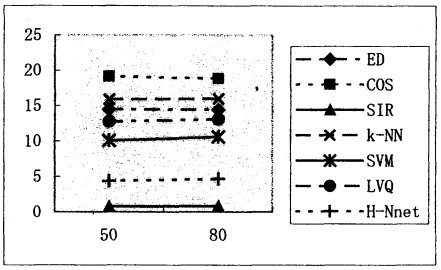

図 1　unigram を用いた誤判別率

用いた bigram のベクトルの長さは 400 にのぼる．特徴ベクトルの長さが分類に与える影響を考察するため，長さを 50，100，150，200，250，300，350，400 のように上位 50 からはじめ 50 ずつ増やしながら分類を試みた．その結果を図 2 に示す．

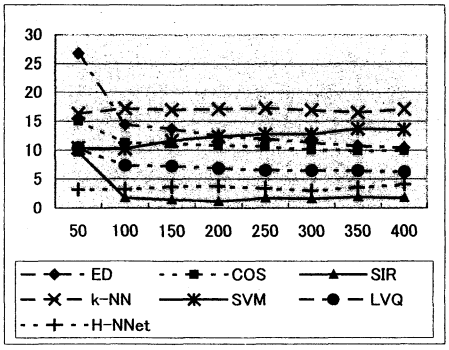

図 2　bigram を用いた誤判別率

特徴ベクトルの上位 50 位を用いた場合は H-NNET 方法による誤判別率が最も低く，その次が SIR 方法である．上位 100 位を用いた場合は，ED, COS, SIR, LVQ 法による誤判別率が急激に減り，SIR 法の誤判別率がもっとも低く，その次が H-NNET 法である．上位 100 から 50 ずつ増やしても誤判別率が低い SIR, H-NNet, LVQ の順序は変わらなかった．SVM 法は特徴ベクトルの変数の数の増加に伴い若干ではあるが誤判別率が増加した．

特徴ベクトルの変数の数の変化に関係なく安定した結果を見せたのは H-NNET 法と *k*-NN 法である．かつ H-NNET は始終よいパホマンスを見せた．SIR は上位 50 変数を用いた場合以外は，誤判別率が最も低い．上位 50 変数を用いた場合と上位 100 変数を用いた場合の誤判別率が大きく変わったのは ED, COS, SIR, LVQ である．これは上位 50 変数にはノイズが多いことが主な原因であると考えられる．

注目されている SVM 法は *k*-NN 法よりはパフォーマンスがよいが，SIR, H-Nnet, LVQ 法よりよいパホマンスは得られなかった．

4. おわりに

本研究では，書き手の同定問題におけるテキスト分類を実例とし，SIR に基づいたテキスト分類方法を提案し，広く使用されているいくつの分類器と比較分析し，その有効性を確認した．SIR に基づいた分類方法は，シンプルで計算量が少ないのにもかかわらずニューラルネットワーク法を含む広く使用されている分類器よりよい結果が得られた．本研究では比率データや確率分布を前提としている．一般のデータの分類への適応の実証や拡張が今後の一つの課題である．

参考文献

1. Cover, T. M. and Hart, P. E. (1967). Nearest Neighbor Pattern Classification. IEEE Transaction on Information theory, IT-B(1), 21-27.
2. Dumails, S. T. et al. (1998). Inductive learning algorithms and representations for text categorization. In Proceedings of CIKM-98, 7th ACM International Conference on Information and Knowledge Management (Bethesda, US), 148-155.
3. Joachims, T. (1998). Text categorization with support vector machines. In Proceedings of ICML-99, 16th International conference on Machine Learning (Bled, SL), 200-209.
4. Li, H. and Yamanishi, K. (1999). Text classification using ESC-based stochastic decision lists. In Proceedings of CIKM-99, 8th ACM International Conference on Information and Knowledge Management (Kansas city, US), 122-130.
5. kohonen, T. (1996). Self-Organizing Maps. Springer-Verlag.
6. Sebastiani , F. (2002). Machine Learning in Automated Text Categorisation. *ACM Computing Surveys*. Vol. 34, No. 1, 1-47. http://faure.iei. pi. cnr. it/ ~fabrizio/
7. Sibson, R. (1969). Information Radius. Z. Wahr. Verw. Geb. 14, 149-160.
8. Vapnic, V. (1995). The Nature of Statistical Learning theory. Springer, New York.
9. Yang , Y. and Liu , X. (1999). A re-examination of Text Categorization Methods, Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR) 42-49.
10. 金　明哲(1997). 助詞の分布に基づいた日記の書き手の認識，計量国語学，20(8)， 357-367.
11. 金　明哲(2002a). 助詞の分布における書き手の特徴に関する計量分析，社会情報，Vol. 11， No. 2， 15-23.
12. 金　明哲(2002b). 助詞の n-gram モデルに基づいた書き手の識別，計量国語学，23 巻 5 号，225-240.
13. 金　明哲(2003). 自己組織化マップと助詞分布を用いた書き手の同定及びその特徴分析，計量国語学，23 巻 8 号(予定).