社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2009－MPS－73 (14)
2009／3／5

相補的な素性選択基準の関係を考慮した文書分類のための素性選択方式

末永高志† 松永務† 関根純†

文書の分類は一般に文書に含まれる素性（単語）の出現情報を文書の特徴として行われる．従って，素性の選択は分類精度に影響を与えるものとして検討が重ねられ，これまで χ^2 統計量や相互情報量といった選択基準が提案されている．ここに，分類精度においては，カテゴリへの帰属を表す網羅性（再現率）とカテゴリにあてはまらない信頼性（適合率）の２つの観点がある．これまで提案された基準はこれら２つの観点のいずれかに応じたもので，特に分類精度の寄与が高くても，低頻度の素性は考慮されない限界があった．本稿では，相補的な分類器の統合を図る集団学習の考え方に基づき，複数の基準を統合する素性選択方式を提案する．提案方式による素性が従来の１つの基準によるものに比べて分類精度に優れること，提案方式により選択された低頻度の素性が分類に効果的に寄与していることを，新聞記事データを用いた実験から明らかにした．

Feature Selection Methods in Text Classification by Combining Complementary Selection Measures

Takashi Suenaga † ,Tsutomu Matsunaga † and Jun Sekine†

In text classification, classifiers are often based on features of terms. Some feature selection measures are proposed, such as document frequency, χ^2-statistics, mutual infomation. The category relevance has two point of views: 1)recall evaluated by classifying positive evidences and 2)precision, rejecting negative ones. These measures are correspond to just a point of view for the requirements. Especially, using just a measure has limitation to select terms with rare but high contributive ones for classification. In this papar, we propose feature selection methods in text classification by combining complementary selection measures, based on the ensemble learning in the machine learning field. Experiments show classification using the smaller term set selected by the proposed method gives higher performance than exitsing measures.

1. はじめに

文書分類に用いる単語（素性）を適切に選択することは，分類プロセスにおける処理時間の軽減や，分類器の過剰適合を回避することにつながり，分類器の構成のためにはカテゴリの分類に寄与する素性の選択が重要である [2]．

文書分類問題の特徴的な課題に，関連するカテゴリラベルを付与することに加え，カテゴリに関連しない文書を排除することがあげられる．具体的には，関連するカテゴリを漏れなく分類するための網羅性の観点と，関連しないカテゴリに誤って分類することを防ぐための，積極的にカテゴリから排除する信頼性の観点である [1]．文書分類に用いる素性に要求される，網羅性と信頼性は背反する要件である．また，低頻度に出現する素性についても無視できない効果があることが指摘されている [4]．選択された素性が活用されるためには，ある程度の頻度で文書に出現することが期待されるが，網羅性や信頼性と出現頻度の取り扱い方法は明らかでない．

分類に寄与する素性の要求に対して，いくつかの選択基準が導入されている [1)~3)]．相互情報量 [3] は網羅性に特化した基準であり，信頼性は考慮していない．χ^2 統計量 [3] は網羅性と信頼性の両方を考慮した基準であるが，低頻度ながら分類に寄与する素性と，高頻度にも関わらず分類の寄与が低い素性の取り扱い方法が明確でない．χ^2 統計量に対して，素性の出現頻度の効果を省略した，簡素化 χ^2 統計量が提案がされている [1]．これは，信頼性よりも網羅性を重視した基準である．

これまで検討された基準はいずれもある観点による

† 株式会社 NTT データ 技術開発本部
R&D Headquarters, NTT data corporation

もので相補的な関係にあるとみられる．上述の文書分類の要件を満たすためには，それらを統合的に扱う必要があると考えられる．

機械学習の分野において，複数の分類器を統合することで，分類の精度向上を目的とした集団学習と呼ばれるアプローチが知られている．その中でも，AdaBoost に代表される，誤分類した学習データの重み付けを変更し再学習を繰り返すことで，多様な分類器群を構築する方式の効果が広く認められている．

本稿では，相補的な関係をもつとみられる素性選択の各々の基準に対して，集団学習の考え方を導入した，文書分類のための素性選択方式を提案する．提案する方式により選択された素性が，網羅性，信頼性の両方の観点に寄与することと，低頻度でもあっても分類に効果的に寄与することを，実データによる分類実験結果により示す．

2. 文書分類のための素性選択基準

文書分類で期待される素性の要求に対して，これまでに検討されてきた素性選択基準について，定義と分類に与える特徴を概説する．

以下，文書分類は二分類問題を対象とし，正例のカテゴリを c，負例のカテゴリを $\bar{c}$，ある素性を t，それ以外の全文書中に出現する素性を $\bar{t}$ とする．また，$P(t)$ は素性 t の出現確率，$P(t,c)$ は素性 t とカテゴリ c の同時確率，$P(c|t)$ は素性 t のもとでのカテゴリ c の周辺確率とする．

χ^2 統計量 [2),3)]

χ^2 統計量 $\chi^2(t,c)$ を素性選択に用いる場合，次式のように展開できる．

$$(\chi^2(t,c)) = \frac{|P(t,c)P(\bar{t},\bar{c}) - P(t,\bar{c})P(\bar{t},c)|^2}{P(t)P(\bar{t})}$$
$$= P(t)P(\bar{t})|(P(c|t) - P(\bar{c}|t)\frac{P(c|\bar{t})}{P(\bar{c}|\bar{t})})P(\bar{c}|\bar{t})|^2$$

これは，素性に対する網羅性と信頼性の双方を考慮した基準と考えられる．

簡素化 χ^2 統計量 [1)]

簡素化 χ^2 統計量 $s\chi^2(t,c)$ は，次式で定義される．

$$s\chi^2(t,c) = P(t,c)P(\bar{t},\bar{c}) - P(t,\bar{c})P(\bar{t},c)$$
$$= P(t)P(\bar{t})(P(c|t) - P(\bar{c}|t)\frac{P(c|\bar{t})}{P(\bar{c}|\bar{t})})P(\bar{c}|\bar{t})$$

これは，網羅性に効果のある素性が選ばれる傾向があると考えられる．

相互情報量 [3)]

相互情報量 $MI(t,c)$ は以下の形式で定義される．

$$MI(t,c) = \log\frac{P(t,c)}{P(t)P(c)}$$

これは，網羅性に寄与する素性が選択される傾向にあると考えられる．

負例の相互情報量

信頼性の観点を積極的に活用する素性選択基準として，相互情報量の考え方に基づいた基準が考えられる．これは，負例のカテゴリ $\bar{c}$ に属する文書と素性 t の独立性を評価する，以下のような基準 $MInega(t,c)$ が相当する．

$$MInega(t,c) = \log\frac{P(t,\bar{c})}{P(t)P(\bar{c})}$$

負例に属する文書に対する相互情報量を考慮しているため，本稿では，上記基準を負例の相互情報量と呼ぶことにする．上述したとおり，これは，信頼性に寄与する素性が選択される傾向にあると考えられる．

文書分類のための素性選択を行うにあたっては，網羅性と信頼性の両面を考慮することが重要である．それに加え，これらの特性が同条件であるならば，文書全体で高頻度の素性の方が分類に寄与する可能性が高くなる．一方で，文書中に出現する素性の頻度分布はいわゆる Zipf の法則に従うことが広く知られており，低頻度の素性の各々が文書分類の結果に与える貢献度は低くても，低頻度の素性全体としてみれば無視できない影響があることが指摘されている [4)]．

χ^2 統計量は網羅性と信頼性の両面を考慮した基準であるが，これらと素性の出現頻度とは異なる観点であり，一つの基準で扱うことは困難と言える．従って，文書分類を目的とした素性選択にあたっては，網羅性や信頼性と出現頻度を同時に考えるだけでなく，出現頻度が少なくとも，網羅性と信頼性のいずれかの向上に高く寄与する素性の観点から選択することが重要と言える．

3. 集団学習に基づく素性選択の統合方式

分類問題の大規模化，多様化に対応するために，複数の分類器を統合する方式が広く知られている．これらの方式では，個々の分類器の性能向上を目指すのではなく，複数の分類器の出力を統合することによって，個々の誤りを軽減し高性能な分類を実現することが期待される．分類器の出力を統合する場合，大きく分けて出力されたカテゴリの度合いを表すパラメータを平均する方式と，出力の多数決を取る方式の二つが広く知られている．

これらを素性選択基準の統合に展開するにあたり，

統合の基となる基準の出力とは，素性ごとに算出された基準値である．この基準値は連続的な値をとるため，パラメータを平均する方式の適用が考えられる．しかしながら，各々の基準値は異なるスケールのため，出力された基準値に適用しても，平均を導出することにはならない．

そこで，各々の素性を基準値によって順位付けすることを考える．これにより同一のスケールで基準間の比較が可能となる．パラメータを平均する方式は，素性ごとに割りあてられた順位の平均値を算出することに対応する．この方式は，より多くの基準で上位に順位付けられた素性が選択されることが考えられる．しかしながら，基準が相補的な関係性をもつのであれば，ある基準では下位であっても，他の基準で上位に順位付けされた素性が有効であることも考えられる．この場面において，順位の平均を用いる方式は機能しない可能性がある．そこで，各々の基準を均等に扱い，順位付けられた素性を上位から順番に選択する方式が考えられる．その他に，ある基準では下位であっても，複数の基準で上位に順位付けされた場合は，順位の分散が大きな値となる．従って，順位の分散を用いる方式が考えられる．

4. 文書分類実験による評価

本章では，文書分類のベンチマークテストに使われる Reuters-21578 の文書データを用いてナイーブ・ベイズ法による分類実験を行い，素性選択結果の評価を行う．

ここで，あるカテゴリに属す文書データを正例とした場合，残りのカテゴリに属す文書データを負例とした．なお，分類実験の評価には，網羅性の基準である再現率と信頼性の基準である適合率の調和平均である F 値を用いる．

4.1 実験に用いるデータ

実験に用いる文書データは，Reuters-21578 から 8 つのカテゴリ acq，crude，earn，grain，interest，money-fx，ship，trade に含まれる記事のみを抽出し，そのうち複数のカテゴリに属す文書を除外した．また，素性となる単語に関しては，文書に含まれる停止語と 1 つの記事のみに出現する素性を除去した．

素性の対象は文書に出現する単語とした．なお，評価対象とするカテゴリ acq は文書数 527，素性の文書頻度 42,210 である．文書データは訓練用と検証用にデータ数を均等に分けた．

4.2 素性選択基準間の相補性の確認

各々の素性選択基準間の相補性を，基準の出力結果の相関係数を基に確認する．具体的には，各々のカテゴリごとに素性の基準値を用いて順位付けを行い，出力されたされた順位に対してスピアマンの相関係数を算出した．分析の対象とする素性選択基準は，2 章で述べた 4 つの基準である．

表 1 に正例をカテゴリ acq とした場合の，各基準による素性選択結果の相関係数をそれぞれ示す．表 1 では，χ^2 統計量は負例の相互情報量との相関係数が 0.413 で正の相関があるものの，残りの基準とはそれぞれ -0.009，-0.007 と相関係数が 0 に近く関連性が低いと言える．また，簡素化 χ^2 統計量と相互情報量は相関係数が 0.988 と相関が高い関係にある一方で，これらと負例の相互情報量との相関係数は，それぞれ -0.631，-0.663 であり負の相関をもつ，すなわち逆の傾向で素性が選択されると言える．

以上から，各々の素性選択基準では相関の意味で関連性が低い，もしくは負の相関を持つ傾向にあり，相補的な関係にあることが実データの結果から確認された．このことから，各々の素性選択基準を統合することで，より効果的な素性選択の実現が期待される．

4.3 各素性選択基準の評価

前節で用いた素性選択基準を用いて素性の順位付けを行い，上位から選択した素性を基にカテゴリ acq の分類実験を行った．図 1 に，横軸を用いた素性の数，縦軸を F 値とした分類の結果を示す．図中の CS は χ^2 統計量，SCS は簡素化 χ^2 統計量，MI は相互情報量，MInega は負例の相互情報量の結果にそれぞれ対応する．

分類結果の最良値は χ^2 統計量を用いた場合で，素性を 10000 種類用いた場合 F 値が 0.84 である．簡素化 χ^2 統計量，相互情報量，負例の相互情報量の優劣はつけがたいと言える．また，簡素化 χ^2 統計量，相互情報量は選択する素性の数を増やすに従い一度精度が落ちてから，再度向上する傾向が示された．このことから，分類に最適な素性の数の存在が想定される

以上から，分類性能としては χ^2 統計量が最もよいことから，次節では，この基準と提案方式を比較し，基準を統合する効果を検証する．

4.4 素性選択の統合方式の評価

本節では，前節で全般的に分類性能が優位であった χ^2 統計量と，3 章で提案した 3 つの方式との比較を行う．図 1 と同様に，カテゴリ acq の分類の結果を図 2 に示す．図中の CS は前節と同様 χ^2 統計量，order は各基準から順番に選択する方式，mean は各基準の順位の平均を用いる方式，variance は各基準の順位の分散を用いる方式の結果にそれぞれ対応する．

表 1　カテゴリ acq の文書における素性選択の基準間の相関係数.
Table 1　Ranking correlations of terms selected from documents of the category "acq".

	χ^2	簡素化 χ^2	相互情報量	負例の相互情報量
χ^2	1.000	-0.009	-0.007	0.413
簡素化 χ^2	-0.009	1.000	0.988	-0.631
相互情報量	-0.007	0.988	1.000	-0.663
負例の相互情報量	0.413	-0.631	-0.663	1.000

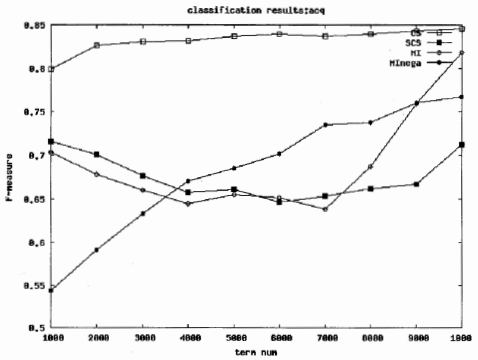

図 1　素性選択基準による素性の分類性能（F 値）：カテゴリ acq.
Fig. 1　F-measures of text classiffiarion for the category "acq" using each term seletion criterion.

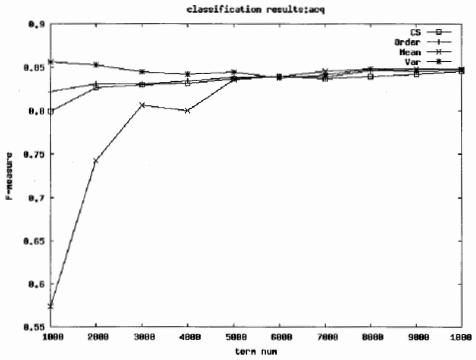

図 2　統合方式と χ^2 統計量の比較（F 値）：カテゴリ acq.
Fig. 2　F-measures of text classiffiarion for the category "acq" using χ^2-stastics and each proposed method.

図 1 と比較すると，素性を追加するに従い一度精度が落ちてから，再度向上する傾向は見られない．統合することで，分類に悪影響を与える素性は排除されたと考えられる．

図 2 から，順位の分散を用いる方式では，用いる素性を追加するにつれ精度が低下しているが，その他では逆の傾向にある．分類性能の最良値は，順位の分散を用いる方式であり，素性を 1000 種類用いた場合に F 値 0.86 が認められた．順位の分散を用いる方式で，少数の素性の種類で最も精度のよい結果となったことから，分類に効果的な素性が効率的に選択されたと考えられる．

以上から，文書分類における網羅性と信頼性の背反する観点に対して，相補的な関係にある素性選択基準を統合する効果が示された．

次に，カテゴリ acq の分類結果の最良値となった順位の分散を用いる方式と χ^2 統計量で選択された素性との違いを比較する．"marine" と "recieved" は順位の分散を用いる方式と χ^2 統計量で選択された素性の例である．素性の文書頻度はそれぞれ 33 と 121 で，χ^2 統計量は高頻度の素性が選択される傾向にある．正例の数はそれぞれ 0 と 19 で，提案する方式では出現頻度が低いが正例では出現しない，すなわち信頼性の寄与が高いと期待される素性が選択されている．

5. ま と め

本稿では，文書分類のための素性選択基準において，相補的な関係性をもつことを明らかにし，これらの統合により，網羅性と信頼性の背反する要件に対応する素性選択方式を提案した．さらに，実データを用いて分類実験を行い，提案した方式で選択された素性を基に，網羅性と信頼性の両方の観点で効果があることを示した．

参 考 文 献

1) Galavotti, L., Sebastiani, F. and Simi, M.: Experiments on the use of feature selection and negative evidence in automated text categorization, *ECDL-00* (2000).
2) Sebastiani, F.: Machine Learning in Automated Text Categorization, *ACM Computing Surveys*, Vol. 34, pp. 1–47 (2002).
3) Yang, Y. and Pedersen, J. O.: A comparative study on feature selection in text categorization, *ICML-97* (1997).
4) 相澤彰子: 低頻度語の利用によるテキスト分類性能の改善と評価, 情報処理学会論文誌, Vol. 44, pp. 1720–1730 (2003).