E-057

# 学習指導要領に基づいた設問の自動分類タスクにおけるモデルの選択に関する研究
# Basic study about comparison of feature model for classification based on course of study

名嘉 真之介† 當間 愛晃‡ 赤嶺有平‡ 山田孝治‡ 遠藤聡志‡
Shinnosuke Naka Naruaki Toma Yuhei Akamine Koji Yamada Satoshi Endo

## 1. はじめに

学校では生徒の学力向上を目的として定期試験が行われている。しかし、その試験結果から生徒の得手不得手の傾向分析が十分に行われていないことや年度毎に行われる学級担任の引き継ぎの際に前年度の結果をうまく引き継げていない現状が、平成 23 年度の沖縄県教育委員会の調べ[1]で分かっている。

これらの問題に対して、年度を跨いだ試験結果のデータを蓄積することで得手不得手のカルテや教員の引き継ぎ資料を作成し、問題解決を図りたい。しかしそのためには、基となる試験結果からこれまで以上のデータ収集が必要不可欠である。

本研究では、試験結果から新たな学習カルテや引き継ぎ資料を作成するための準備段階として、機械学習を使用し学習指導要領[2][3]に基づいた社会科の設問の自動分類を行う。自動分類について十分な精度を得ることができれば、それを基に生徒が正解した設問と間違えた設問を分類して得手不得手の傾向を示す学習カルテを生成することができ、教員の年度毎の引き継ぎ資料としても活用することが期待できる。

## 2. 提案手法

設問に出現する名詞を素性とし、TFIDF により特徴量を算出して特徴ベクトルを生成する。

そして生成した特徴ベクトルに教師データとしてラベルを用意し、機械学習によりマルチラベル分類を行う分類器を構築する。なお用意したラベルは学習指導要領の見出しの項目を採用し、地理 8 ラベル、歴史 20 ラベル、公民 8 ラベルの全 36 ラベルを定義した。

次に今回提案する手法を各段階に分けて説明する。

### 2.1 テストデータの用意

今回扱うテストデータは、H18 年度から H22 年度の沖縄県立高校入試[4]の社会科の設問とする。これらのテストデータにどのラベルが適切なのか調べ、手作業で正事例を用意する。

### 2.2 特徴ベクトルの生成

設問に出現する名詞がその文章を特徴づけているとして、テストデータに MeCab[5]を用いて形態素解析を行い、名詞を抽出する。そして抽出した名詞群から TFIDF(式 1)を算出し、それを基に特徴ベクトルを生成する。

$$TFIDF_{ij} = tf_{ij} \times log(\frac{N}{df_i}) \quad \text{...(式 1)}$$

$tf_{ij}$ = 文書j中の単語iの出現頻度
$df_i$ = 単語iを含む文書の数
$N$ = 全ての文書の数

### 2.3 機械学習

オープンソースデータマイニングツールである WEKA[6]を用いて、生成した特徴ベクトルの機械学習を行う。用いる分類器は基本的に二値分類(T or F)を使用し、今回は J48、NaiveBayes、RandomTree、SimpleLogistic、SMO の 5 種類の分類器を使用する。これらを設問の自動分類に応用する為、対象となるラベルを設問に付けるか否かを分類する分類器はラベルごとに構築する。

そして構築した分類器を使用して、生成した特徴ベクトルにマルチラベル分類を行い、設問の自動分類を行う。

### 2.4 評価方法

精度評価のための基準としては再現率と適合率の調和平均である F 値を使用する。F 値が大きければ大きいほど正しく分類されているとして、分類結果を評価する。

### 2.5 全体概要

以下に今回提案する手法の全体概要を示す(図 1)。

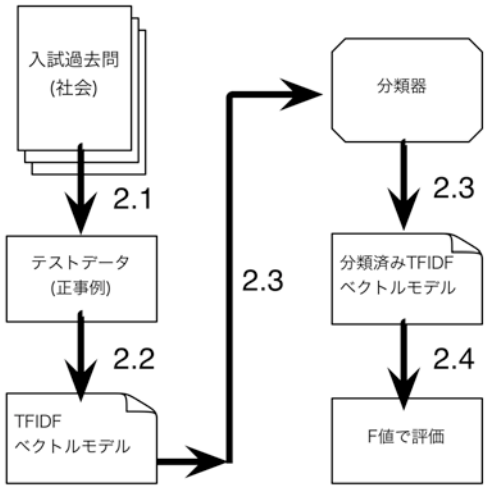


図1 提案手法の全体概要

† 琉球大学大学院理工学研究科情報工学専攻
‡ 琉球大学工学部情報工学科

## 3. 実験

提案手法の妥当性を評価する為、提案手法に基づいた二つの実験を行った。なお、今回は機械学習するにあたって正事例が 1 つもないタグや数個しかないタグが多く見られたため、比較的正事例数の多かった公民タグから選出して実験を行っている。選出したタグは公民分野から K2-1、K2-2、K3-1、K3-2 の 4 タグである。以下に公民タグの一覧と今回使用するタグを示す(表 1)。

表 1　公民タグの一覧と今回使用するタグ

| ラベル名 | ラベルの内容 | 正事例数 | 選出されたタグ |
|---|---|---|---|
| K1-1 | 私たちが生きる現代社会と文化 | 3個 | × |
| K1-2 | 現代社会をとらえる見方や考え方 | 0個 | × |
| K2-1 | 市場の働きと経済 | 10個 | ○ |
| K2-2 | 国民の生活と政府の役割 | 9個 | ○ |
| K3-1 | 人間の尊重と日本国憲法の基本原則 | 10個 | ○ |
| K3-2 | 民主政治と政治参加 | 21個 | ○ |
| K4-1 | 世界平和と人類の福祉の増大 | 5個 | × |
| K4-2 | よりよい社会をめざして | 3個 | × |

### 3.1 実験 1(ノイズワード数の比較)

提案手法で述べた特徴ベクトルの生成について、本実験では正規化ベクトルと非正規化ベクトルの二種類の TFIDF ベクトルを生成し、どちらがより良い名詞の抽出が行えているかの比較を行った。以下にその実験の詳細な手順を示す。

#### 3.1.1 実験方法

1. テストデータから正規化ベクトルと非正規化ベクトルをそれぞれ生成する。
2. 両ベクトルの正事例を使用し、各ラベルの名詞の特徴量を調べ、その特徴量の合計を算出する。
3. 上位 30 単語を表にしてノイズワード(注 1)の数を比較する。そしてノイズワードの少ないモデルが優れているとして評価する。

#### 3.1.2 実験結果

ここでは上位 30 単語中にノイズワードがいくつ出現したのかの数を比較した表を示す(表 2)。

結果は次のようになった。

表 2　ノイズワード数の比較

| | TFIDF(正規化無し) | TFIDF(正規化有り) |
|---|---|---|
| K2-1 | 6 個 | 13 個 |
| K2-2 | 7 個 | 12 個 |
| K3-1 | 7 個 | 17 個 |
| K3-2 | 6 個 | 14 個 |

ノイズワードの数を比較した結果、非正規化ベクトルと比べて正規化 TFIDF ベクトルの方が多くのノイズワードが見られた。

次にその両ベクトルで特にノイズ数に差が大きかった K3-1 について上位 30 単語を示す(表 3)。

表 3　K3-1 タグのノイズワード数の比較

| 順位 | K3-1(正規化無し) | K3-1(正規化有り) |
|---|---|---|
| 1 位 | 開示 | 権利 |
| 2 位 | 権利 | 生活 |
| 3 位 | 花子*** | 開示 |
| 4 位 | 生活 | 自由 |
| 5 位 | 日本国憲法 | 下*** |
| 6 位 | 過半数 | 問*** |
| 7 位 | 請求 | 記号*** |
| 8 位 | さん*** | 語句*** |
| 9 位 | 自由 | 日本*** |
| 10 位 | 賛成 | 下線*** |
| 11 位 | 住民投票 | 部*** |
| 12 位 | 情報 | 次*** |
| 13 位 | 相談 | 日本国憲法 |
| 14 位 | 下*** | 答え*** |
| 15 位 | 文書*** | 工*** |
| 16 位 | 新しい人権 | さん*** |
| 17 位 | 議員 | 花子*** |
| 18 位 | 議院 | 国民 |
| 19 位 | 国民投票 | 関連*** |
| 20 位 | 承認 | 布教 |
| 21 位 | 行政 | 関係*** |
| 22 位 | 防衛 | 過半数 |
| 23 位 | 自衛隊 | 右*** |
| 24 位 | 語句*** | 文書*** |
| 25 位 | 国民 | クーリング・オフ |
| 26 位 | 布教 | リストラ |
| 27 位 | 関係*** | 図*** |
| 28 位 | 問*** | 製造物責任法 |
| 29 位 | 規定 | 独占禁止法 |
| 30 位 | 限度 | 限度 |

※ノイズワードには「***」を付けて加えている

表 3 を見ると、K3-1 の正規化ベクトルには設問特有の名詞である「図」や「記号」「語句」などのノイズワードが多く出現していることがわかる。

よって今回の実験により、テストデータから生成した TFIDF ベクトルでは、非正規化ベクトルの方が正規化ベクトルよりテストデータの特徴を表せると考えられる。

### 3.2 実験 2(ラベルの F 値の比較)

実験 1 で比較した両ベクトルを今度は提案手法に沿ってそれぞれ機械学習させた。F 値を使用し、どちらがより正しく自動分類を行えているかどうか比較を行った。次にその詳細を示す。

#### 3.2.1 実験方法

1. 負事例数(注 2)を調整した両ベクトルをそれぞれ 10 パターンずつ用意する。
2. 両モデルを各分類器で 10 分割交差検定を行い、全てのパターンについて機械学習を行う。
3. 分類精度をすべてのパターンの F 値の平均で評価し、どちらがより正しく分類できているか比較を行う。

#### 3.2.2 実験結果

ここでは両ベクトルについて機械学習した結果をグラフで比較した。結果は次のようになった。

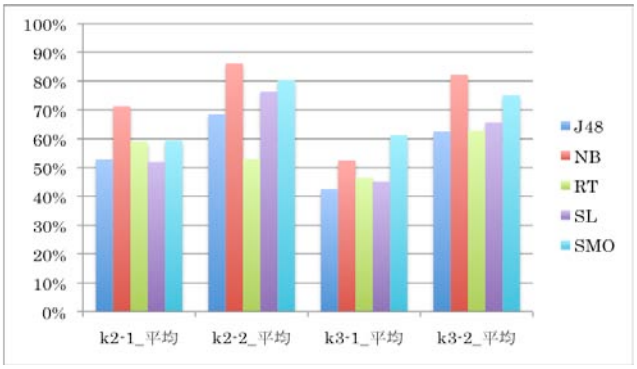


図 2　非正規化ベクトルの平均 F 値

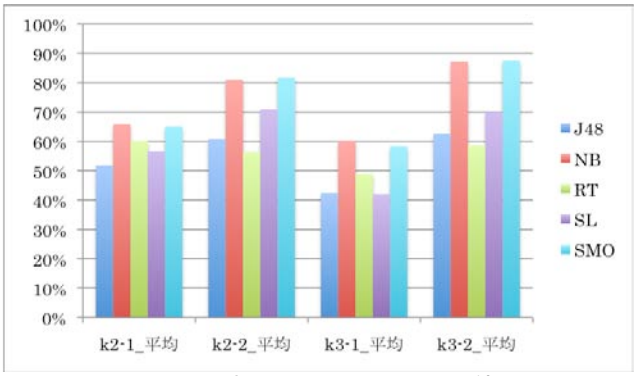


図 3　正規化ベクトルの平均 F 値

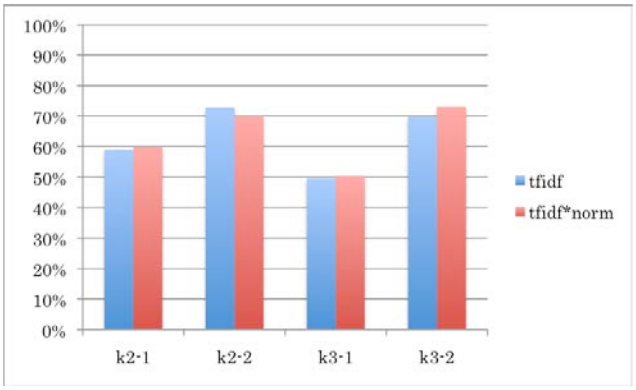


図 4　各分類器の平均 F 値の比較

両ベクトルの F 値について比較した結果、全体的な結果(図 4)としてやや正規化ベクトルの方が F 値は高いという結果になったが、それほど大きさ差は見られなかった。

しかし部分的に見る(図 2,図 3)と、k3-2 の NaiveBayes や SMO など F 値には大きな差が見られる。

また F 値の全体的な傾向として、使用する分類器によって結果にバラつきがある。比較的良い結果が出た分類器としては NaiveBayes や SMO が挙げられ、正規化ベクトルの k3-2 タグだと 8 割強と良い結果となっている。そして比較的悪い結果が出た分類器としては J48 や RandomTree、SimpleLogistic などがあげられ、正規化ベクトルの k3-1 タグだと 4 割程度の結果となっている。

## 4. 考察

実験 1 と実験 2 の結果を踏まえて、設問の自動分類について分類精度の向上には今後どのようなアプローチをすべきか考察していく。

### 4.1 モデルの質の検討方法について

実験 1 の結果として、正規化ベクトルの方がノイズワードは多く検出されたため、機械学習させた結果では非正規化ベクトルの方が良い結果が出ると思われた。

しかし、実験 2 の機械学習の結果を見てみるとさほど大きな差は見られず、やや正規化ベクトルの方が良くなるという結果が出た。このことから次のようなことが考えられる。

1. 二種類の TFIDF ベクトルを生成し、それぞれの特徴量の合計を使用して特徴ベクトルの質の比較を行ったが、特徴量の合計ではベクトルの質をうまく表現できていなかったのではないか。
2. 実験 1 でノイズワードを定義したが、この定義した名詞群が設問の特徴を表す上でノイズワードではなかったのではないか。

よって今後はこれらを踏まえて、モデルの質を検討するために統計的な別のデータ解析の方法を考える。また定義したノイズワードが本当に機械学習に影響を与えているか検討していく必要があると考えられる。

### 4.2 特徴ベクトル内のパターン抽出について

実験 2 では二種類の TFIDF ベクトルをそれぞれ 10 パターンずつ機械学習させ、その結果の F 値の平均をとり、グラフ化して両モデルを比較している。この方法は分類精度を調べるには良いが、個々の特徴ベクトルに存在する特徴を読み取ることができない。

そこで F 値の最高値や最低値を出した特徴ベクトルを調べることによって、個々の特徴ベクトル内に存在する良い F 値が出るパターンや悪い F 値が出るパターンなどを抽出できないだろうか。

パターンとしては特徴ベクトルを構成する負事例を調べ、共通の設問や独自の設問にどんなものがあるか、出現する単語の特徴量にはどのような特徴があるかなどを使用することが考えられる。よって今後は特徴ベクトルを構成する負事例のパターンについていろいろな角度から調べる必要がある。

### 4.3 TFIDF ベクトルについて

今回の実験では、TFIDF をベースとした 2 種類の特徴ベクトルの生成を行い、機械学習させた。その結果、ラベルによっては F 値が 8 割強と良い結果を出ているラベルもあれば、F 値が 4 割程度のラベルも見られる。よって F 値に

バラつきがあるため、設問を自動分類するための特徴ベクトルとしてはまだ不完全であると考えられる。

また名詞を特徴とした TFIDF ベクトルは意味のある単語が出現しやすくベクトルの中身のデータ解析がしやすいが、名詞以外の品詞を完全に省いてしまうために完全な設問の特徴とはいえないのではないか。

よって、全ての品詞を対象とした特徴ベクトルの生成や n-gram のような隣り合う文字グループから特徴を抽出し生成した特徴ベクトルの生成方法を検討する必要があると思われる。

## 5. 検証

今回は 4.2 で考察した特徴ベクトル内のパターン抽出について、今回生成した特徴ベクトルの F 値がどの程度バラつきがあるか調べることによって特徴あるパターンが抽出できるか検証した。

次に各ラベルの 10 パターンの特徴ベクトルの F 値の最高値と最低値の差を表にした(表 4、表 5)。

表 4　非正規化ベクトルの F 値の最大変位

| 分類器 | K2-1 | K2-2 | K3-1 | K3-2 |
|---|---|---|---|---|
| J48 | 0.45** | 0.22 | 0.33 | 0.20* |
| NaiveBayes | 0.25 | 0.16* | 0.30** | 0.15 |
| RandomTree | 0.22 | 0.28 | 0.31** | 0.19* |
| SimpleLogistic | 0.19 | 0.29 | 0.54** | 0.09* |
| SMO | 0.58** | 0.29 | 0.29 | 0.17* |
| 分類器平均 | 0.27** | 0.13 | 0.22 | 0.12* |

表 5　正規化ベクトルの F 値の最大変位

| 分類器 | K2-1 | K2-2 | K3-1 | K3-2 |
|---|---|---|---|---|
| J48 | 0.30 | 0.39 | 0.41** | 0.29* |
| NaiveBayes | 0.44** | 0.44** | 0.32 | 0.08* |
| RandomTree | 0.46** | 0.28* | 0.41 | 0.29 |
| SimpleLogistic | 0.28 | 0.23 | 0.29** | 0.21* |
| SMO | 0.59** | 0.31 | 0.35 | 0.10* |
| 分類器平均 | 0.27** | 0.17 | 0.16 | 0.10* |

※**：最大変位　*：最初変位

結果として、分類器平均でみると 2 種類のベクトルとも K2-1 の変位が 4 つのラベルの中で一番大きいことがわかった。よって、K2-1 の特徴ベクトルの中で最大の F 値を持つ特徴ベクトルと最低の F 値を持つ特徴ベクトルの違いを発見できれば良い F 値が出るパターンを抽出できることを期待できる。

また分類器別に見ていくと、SimpleLogistic や SMO で最大変位が 0.5 以上のラベルが見受けられる。これらについても対応する特徴ベクトルを調べることで良い F 値が出るパターンを抽出への貢献を期待できる。

### 5.1 安定した学習精度と過学習について

今回の検証で K3-2 の変位が比較的小さいことがわかった。恐らくその要因としては、K3-2 は選出した 4 つのラベルの中で一番多くの教師データを持つことが挙げられる。よって、今回提案した手法で学習の精度を安定させるには、K3-2 が持つ教師データ数があれば、学習を安定させることができるのではないかと考えられる。よって今後の新たな検討内容として、全てのラベルの教師データ数を K3-2 程度にして学習させることで F 値の変位が小さくなるか検討する必要がある。

また K3-2 の教師データ数以上の教師データを用意し学習を行うことで、過学習がどこで発生し学習精度が落ちるか調べ、適切な教師データ数を検討する必要がある。

## 6. おわりに

本研究では学習指導要領に基づいた設問の自動分類をするための基礎研究として、入試問題の社会科における設問の自動分類方法を提案した。

実験 1 では設問に使用される名詞群の TFIDF を算出し、それを基に正規化ベクトルと非正規化ベクトルを生成した。その後、両ベクトルの名詞の特徴量の合計を使用し、上位 30 単語に出現するノイズワード数の比較を行い、両ベクトルの質を調べた。

実験 2 では両ベクトルを提案手法に沿って機械学習させ、その精度を F 値で評価し、グラフで比較を行った。

考察では二つの実験結果から、ベクトルの質の検討方法について、特徴ベクトル内のパターン抽出の方法について、TFIDF ベクトルについて考え、今後の分類精度の向上に向けての考察をした。

検証では考察で考えたパターン抽出について、特徴ベクトルの F 値のバラつきを最大値と最小値の変位で表し、両ベクトルでの最大変位を調べてパターン抽出ができそうな特徴ベクトルの検討を行った。またそこで新たに学習精度を安定方法や過学習について検討を行った。

最後に今後の課題として、統計学的な別のアプローチで実験結果のデータ解析を行うことや特徴ベクトル内のパターン抽出すること、n-gram のような TFIDF とは異なる方法で生成されるモデルの作成を行うことなど考察や検証の部分で挙げたことの実験・検証を行う。そして、設問の自動分類精度の向上を目指していく。

### 参考文献

[1] “学校教育における指導の努力点”, 沖縄県教育委員会, (2011)

[2] “中学校学習指導要領” 文部科学省(2008, 3)

[3] “中学校学習指導要領解説 社会編” 文部科学省(2008, 9)

[4] ”平成 23 年度受験 沖縄県 県立高校入試問題”, 富士教育出版社

[5] “MeCab” http://mecab.sourceforge.net/

[6] “WEKA 3:  Data Mining Software in JAVA” http://www.cs.waikato.ac.nz/ml/weka/

(注 1)

定義として、ノイズワードとは問題文に出現する特有の問題内容とは関係無い名詞とする。

例：図、問、記号、語句、下線、次、答えなど

(注 2)

今回使用する教師データはラベル毎に正事例となる設問以外を全て負事例とするため、正事例に比べて負事例の数が多くなる。そのままの状態で機械学習をさせるとデータの不均衡が発生し、正しく学習が行われない可能性がある。よって正事例の数に合わせて全負事例の中からランダムで同じ数だけの負事例を抽出して TFIDF ベクトルを生成することでデータの不均衡を回避している。