D-007

# 機械学習のための区間属性の提案
# Intervals as Attributes for Machine Learning

廣川　佐千男†1　　杉原　亨‡
Sachio Hirokawa　Toru Sugihara

## 1. はじめに

データの分析をする場合、一つの対象データがたった一つの数値として表現されている事はない。例えばアンケートデータの場合、回答の中には、単純な項目選択や、5 段階のリッカート尺度だけでなく、回答者の名前、年齢、性別などのプロフィールデータが含まれることもある。アンケートによっては、回答者の身長や体重などの実数値もある。また、自由記述の文章による回答が含まれることもある。数値として表現されるデータでも、順序尺度か間隔尺度かで、取り扱いに注意しなければならない[5]。データ分析の典型として二群の比較があるが、対象を表す数値がどのような母集団に属するか、分布がどのようになっているかに応じて、妥当な分析方法を選択しなければならない[7]。二群の違いを検定して有意に違うことが分かったとしても、その違いの 要因を明らかにする必要がある。ところが、因子の妥当性として使われる p 値について、問題が指摘されている[2]。統計的検定は、対象とする二群の違いが統計的有意性を調べる。二群を 比較するという意味では、機械学習による２クラス分類も利用すること ができる。検定での因子分析に対しては、属性の重要性を評価する 属性選択の研究がある[3,4]。

本発表では、文書中の単語の出現情報や単純選択肢のような名義尺度や 5 段階のような順序尺度や体重や計測値などの実数値という、多種類のデータ により記載される分析対象を、統一的にベクトル化する方式を提案する。 具体的には、一つの数値として表現される属性をその数値を含む複数の 区間としてとらえ、それら区間を仮想的な単語とすることで、全ての属性 を名義尺度で表現する。この提案手法を大学評価や教育改善のために多くの組織で 実施されている学生調査[1]のアンケートに適用した。具体的 には、[6]で行った地方短期大学学生を対象としたアンケート で、資格取得した学生とそうでない学生の比較を間隔属性を使うベクトル化 である提案手法により行い、標準的なベクトルよりも高い識別性能が 得られることを示した。

## 2. 地方短期大学学生への意識アンケート

本稿で対象とするのは地方短期大学の卒業生 120 名についてのアンケート 結果である。このアンケートはいくつかの目的があるが、 本稿では、在学中の資格を取得した学生（107 名）と、そうでない学生（13 名） の違いを明らかにする。[6]ではマンホイットニー検定により分析を 行っている。本稿では、資格を取得した学生とそうでない学生 の識別問題として、機械学習を適用し、ベクトル化による識別性能の比較と、 識別の主要因を分析する。

アンケートは Q1 から Q15 までの 15 種類の質問項目からなる。 いくつかの項目はさらに複数の小項目から構成される。 各小項目の質問は、単純な選択肢（名義尺度）を選ぶものと、5 段階での順序尺度を選ぶものの 2 種類ある。

表 1　アンケート質問項目（順序尺度）

| 質問番号 | 項目数 | 質問内容 |
|---|---|---|
| Q7 | 5 項目 | 期待と入学後印象 |
| Q9 | 5 項目 | 授業に対する考え方 |
| Q11 | 6 項目 | 役立ったか |
| Q13 | 1 項目 | 先生に在学中相談 |
| Q14 | 1 項目 | 先生に卒業後相談 |
| Q15 | 6 項目 | 学びの必要性 |

Q10 が本稿で分析の目的変数とする項目で、資格や免許の取得の有無を尋ねる項目である。 Q1, Q2, Q3, Q4, Q5, Q6, Q8, Q11, Q12 はそれぞれ 性別、専攻、卒業見込み、入学方法、入学理由、 他大学進学検討の有無、成績、卒業後の進路について複数の選択肢（名義尺度） を選択する形式の質問である。 残りの、Q7, Q9, Q11, Q13, Q14, Q15 はそれぞれ複数の小 項目からなる質問で「とてもそう思う(5)、まあそう思う(4)、 どちらとも言えない(3)、あまり思わない(2)、まったく思わない(1)」のように 順序尺度の質問である(表 1)。

## 3. 区間属性による数値の名義尺度化

本稿で分析対象としてアンケートの回答は、単純な選択肢（名義尺度）の回答は 0,1 の論理値で表現し、5 段階のリッカート尺度の回答は実数値としてベクトル化すれば一人の学生の回答は 61 次元実数値ベクトルとして表現できる。属性名と値を組とする単純名義尺度化によれば 282 次元の論理値ベクトルとなる。本稿ではさらに、ある属性がある値となっていることを、値そのものではなく、その値を含む可能な全区間を列挙することで表現する。例えば、大項目で 7 番目の質問の中の 2 番目小項目質問に対する 5 段階選択肢について 4 を選択したとき、単純名義尺度化だと Q7-2:4 という単語として表現される。区間属性法では、Q7-2:[3,5]のように、[3,5]のように 4 を含む区間を使って表す。

## 4. 資格取得学生と資格非取得学生の機械的判別

3 種類のベクトル化（標準的数値化、単純名義尺度、区間属性）について 資格取得した学生 107 名を正例として機械学習の SVM を適用した。 ただし、標準的数値でのベクトル化の場合には SVM-light で下の範囲でパラメータのグリッドサーチを行った。 ただし、$C = 10^n$ (n=－3~+3) (線形カーネル、多項式カーネルの場合のパラメータ)、 $\gamma = 10^n$ (n=－3~+3)（ガウシアンカーネル RBF の場合のパラメー

†九州大学
‡関東学院大学

タ）。 また、単純名義尺度、区間属性によるベクトル化については 線形カーネルの SVM を適用し属性選択[4]を適用した。

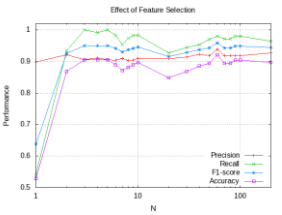


図 1　属性選択の効果

表 2 では、標準的数値化については最適パラメータについての性能、 単純名義尺度と区間尺度については、最適属性数での性能を示した。 最後の 2 行は参考のため、属性選択を適用せず全属性を使った 場合の性能を示した。 recall 以外の、precision,,F1-score,acuracy の評価指標については、 本稿で提案する区間尺度で属性選択を適用した N=60 の場合が、最も高い 識別性能となっている。recall についても、提案手法は 98%という 高い性能となっている。属性選択の個数 N を変化させたときの識別 性能をプロットしたのが図 1 である。識別性能が最適なのは N=60 だが、 N=3 でもほとんど変わらない高い性能を示している。

表 2　識別性能比較

| モデル | parameter | precision | recall | F1-score | Accuracy |
|---|---|---|---|---|---|
| 線形 | C=1 | 0.8929 | 1.0000* | 0.9426 | 0.8929 |
| 多項式 | C=1000 | 0.9061 | 0.8906 | 0.8973 | 0.8198 |
| RBF | C=100 γ=0.01 | 0.8852 | 1.0000* | 0.9386 | 0.8852 |
| 単純名義尺度 | N=200 | 0.9161 | 0.9643 | 0.9391 | 0.8897 |
| 区間属性 | N=60 | 0.9377* | 0.9805 | 0.9577* | 0.9214* |

## 5. 資格取得学生の特徴

ウィルコクソンの検定での p 値を求めたものが表 5 である。 0.05 未満で有意なものは 6 個しかない。 一方、図 1 でも示したように、区間属性法では N=3 までの 6 個の属性で最適属性選択の N=60 の場合とほぼ同じ識別 性能になっている。具体的には、正の特徴属性は、 Q9-2:[3,5]、Q9-2:[3,4]、Q12:1 の 3 つと、 負の特徴属性は Q9-2:[1,2]、Q13:4、Q13:[4,5]の 3 の 合計 6 個で F1-socre, Accuracy がそれぞれ 0.9502、0.9055 となっている。そして Q9-2 の質問は「先生に良い授業を行うよう、 要求する権利をもっている」であった。この質問に強く肯定 した学生が資格を取得した学生であり、そう思わない学生が 資格を取得しなかった学生といえる。つまり、消費者として の意識の高さが資格取得の主要要因であることが確認できた。 一方、表 3 の p 値が 0.05 未満の 6 個の属性については、SVM でのスコアを並べた表 4 の上位 3 位までには現れていない。

表 3　重要属性の p 値

| p 値 | 番号 | 質問内容 |
|---|---|---|
| 5.477e-05 | 8-4 | 成績　不可失格割合 |
| 0.01375 | 8-3 | 成績　可割合 |
| 0.01602 | 8-1 | 成績　優割合 |
| 0.02726 | 9-1 | 授業への考え方　学費を払っているので多くの授業を受けたい |
| 0.04065 | 9-4 | 授業への考え方　良い授業を要求する権利を持っている |
| 0.0464 | 8-2 | 成績　良割合 |

表 4　重要属性の SVM スコア

| 順位 | スコア(正) | 属性 | スコア(負) | 属性 |
|---|---|---|---|---|
| 1 | 0.0973 | Q9-2:[3,5] | -0.115 | Q13:4 |
| 2 | 0.0941 | Q12:1 | -0.0905 | Q9-2:[1,2] |
| 3 | 0.0928 | Q9-2:[3,4] | -0.0702 | Q13:[4,5] |
| 4 | 0.0756 | Q9-4:[4,5] | -0.0701 | Q8-4:[20,30] |
| 5 | 0.0753 | Q13:[1,3] | -0.0688 | Q9-4:[1,3] |
| 6 | 0.0732 | Q11-3:4 | -0.0684 | Q12:6 |
| 7 | 0.0725 | Q9-2:3 | -0.0683 | Q8-4:20 |
| 8 | 0.0683 | Q11-2:[2,4] | -0.0646 | Q13:[3,4] |
| 9 | 0.0675 | Q13:[2,3] | -0.0625 | Q7-3:[3,4] |
| 10 | 0.1629 | Q9-4:4 | -0.1704 | Q7-5:5 |

## 6. まとめと今後の課題

対象をベクトルとして表現し、正例と負例に識別することが機械学習の典型的な利用方法である。対象について出来るだけ多くの情報を使いたい。しかし、自由記述のテキストやアンケートの選択肢や、測定値のように多様なデータがあり、共通に扱うのは困難であった。本発表では、実数値のベクトル化として、その値を含む複数の区間を利用する区間属性法という手法を提案する。ユーザーアンケートに適用して、従来より高い性能で正例ユーザーの識別が出来る事を示した。今後、観測数値や文章などの混合データへの効果を検証する予定である。

### 参考文献

[1] A.W. Astin (1984), Student Involvement: A Development Theory for Higher Education, Journal of College Student Development,Vol.40, No.5,pp.518-529
[2] M. Baker (2016), Statisticians issue warning on P values, Nature, Vol.531, pp.151, https://www.nature.com/news/polopoly_fs/1.19503!/menu/main/topColumns/topLeftColumn/pdf/nature.2016.19503.pdf
[3] J.Li, K. Cheng, S. Wang, F. Morstatter, R.P. Trevino, J. Tang, H. Liu, (2017), Feature selection: A data perspective, ACM Computing Surveys, Vol.50, No.6, [94]. https://doi.org/10.1145/3136625
[4] T. Sakai, S. Hirokawa (2012), Feature Words that Classify Problem Sentence in Scientific Article, Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services, pp.360-367
[5] S. S. Stevens (1946), On the Theory of Scales of Measurement, Science Vol.103, Issue 2684, pp. 677-680
[6] T. Sugihara(2016),A Study on the Survey for Consumer Awareness and Behaviors of Local Junior College Graduates in Japan,Proceedings of 5th IIAI International Congress on Advanced Applied Informatics, pp.1200-1204
[7] 高木英行 (2014), 使える！統計検定・機械学習 --- I ---2 群間の有意差検定,システム／制御／情報, vol.58, no.8, pp.346-352