

JohnsonSU 分布を用いた異字種パターン検出に関する一検討

鈴木雅人 † 北越大輔 † 松本章代 ‡
 † 東京工業高等専門学校情報工学科 ‡ 東北学院大学教養学部

1 はじめに

大学・高等等の入試において、採点誤りによる社会問題が多発していることは周知のとおりである。この問題を改善するために、著者らは入試採点を支援するシステムの構築に関する研究 [1] を行っており、その中で、採点誤りの自動検出法に関する研究を行っている。この自動検出法は、設問ごとに収集した全答案画像の特徴量から特徴の母集団分布を推定し、仮説検定によって大きく逸脱している答案画像を誤答と判断して、採点誤りを検出するものである [2]。しかし、母集団分布の推定で用いる部分的正規分布は、正規分布に多少の歪みを加えた分布であり、表現できる歪みには限界がある。そこで本稿では、分布の表現自由度が高い JohnsonSU 分布を用いて分布のあてはめを行い、採点誤りの検出精度改善を試みる。

2 採点誤り検出アルゴリズム

設問に対する採点が正しいかどうかを判断するために、その設問の答案に記載された内容が正しいかどうかを判定する。提案する採点誤り検出アルゴリズム [2] の概要を図 1 に示す。一般に、手書きの答案画像を直接認識して採点誤りを検出することは難しいため、提案アルゴリズムでは、各設問に対する全答案画像を収

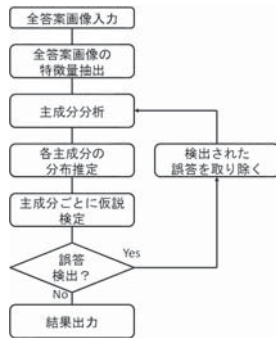


図 1: 採点誤り検出アルゴリズム

A Study on the Detection Algorithm of a Different Type of Letters using JohnsonSU distribution
 †Masato SUZUKI †Daisuke KITAKOSHI ‡Akiyo MATSUMOTO
 †Department of Computer Science, Tokyo National College of Technology
 ‡Faculty of Liberal Arts, Tohoku Gakuin University

集し、それらの画像の特徴を反映した特徴量を抽出する。次に主成分分析を行った後、各主成分軸上の特徴分布を推定し、大きく逸脱しているデータを仮説検定を用いて逐次的に検出して、誤答の可能性のある答案を検出する。以下に主な工程の詳細を述べる。

2.1 特徴量

本アルゴリズムでは、解答画像から抽出する特徴量として、方向線素特徴量 (Directional Element Feature) を用いる。この特徴量は、文字画像を局所領域に細分し、線分の傾きを 4 方向に量子化して 196 個の数値を数え上げたもので、もとは手書き文字認識に用いられたものである。そのため、入試の答案用紙に記入されるような、字形にばらつきのある文字を高精度に認識することは困難であるが、多くのパターンの中から明らかに違う構造を持つ文字画像を検出することは可能であると考えられる。

2.2 分布のあてはめ

同じ解答欄の答案画像から得られた方向線素特徴量を用いて、その画像群に関する共分散行列から、196 個の主成分を抽出する。画像群の特徴量を各主成分軸に投影し、JohnsonSU 分布のあてはめにより、母集団分布の推定を行う。JohnsonSU 分布は 4 つのパラメータ $\gamma, \delta, \lambda, \xi$ をもつ分布で、確率密度関数 $p(x)$ は次式で与えられる。

$$p(x) = \frac{\delta}{\lambda\sqrt{2\pi}\sqrt{z^2+1}} \exp\left\{-\frac{1}{2}(\gamma + \delta \sinh^{-1} z)^2\right\}$$

$$\text{ただし、 } z = \frac{x - \xi}{\lambda}$$

$p(x)$ の期待値を m 、分散を σ^2 、歪度を β_3 、尖度を β_4 とすると、それらは次のような式で与えられる。

$$m = \xi - \lambda\sqrt{\omega} \sinh \Omega$$

$$\sigma^2 = \frac{\lambda^2}{2}(\omega - 1)(\omega \cosh 2\Omega + 1)$$

$$\beta_3 = \omega(\omega - 1) \frac{\{\omega(\omega + 2) \sinh 3\Omega + 3 \sinh \Omega\}^2}{2(\omega \cosh 2\Omega + 1)}$$

$$\beta_4 = \frac{\omega^2(\omega^4 + 2\omega^3 + 3\omega^2 - 3) \cosh 4\Omega}{2(\omega \cosh 2\Omega + 1)^2} + \frac{4\omega^2(\omega + 2) \cosh 2\Omega + 3(2\omega + 1)}{2(\omega \cosh 2\Omega + 1)^2}$$

ここに、 $\omega = \exp(\delta^{-2})$, $\Omega = \gamma/\delta$ である。学習データから、これら4つの統計量の不偏推定量を求めると、上式を用いた非線形連立方程式を解くことによって、 $\gamma, \delta, \lambda, \xi$ を求めることができるので、母集団分布を JohnsonSU 分布で推定することが可能である。

2.3 誤答検出アルゴリズム

全ての主成分に対し、推定した母集団分布を用いて、それぞれの標本がこの母集団から抽出されたものかどうかを、仮説検定により判断する。一般に、誤答は一部の主成分において、他の標本から非常に大きく離れた値を持つため、そのような主成分においては、歪度の絶対値が大きくなると考えられる。また、主成分軸を表す固有ベクトルは、対応する固有値の小さい方に、より計算誤差が蓄積されるため、誤答の判別には、値の大きい固有値に対応する固有ベクトルで表される主成分軸が有効であると考えられる。以上のことから、本アルゴリズムでは、値の大きい固有値を30個取り出し、それらに対応する固有ベクトルの中から、歪度の絶対値が閾値(本稿では0.55)より大きい主成分軸を抽出し、有意水準5%の仮説検定を行って誤答を検出する。すなわち、各主成分軸の半数以上で棄却される答案画像のうち、もっとも棄却回数の多いものを誤答と推定し検出する。また、もしそのような答案画像がなければ、答案画像群の中に誤答は含まれないと判断し、検出アルゴリズムを終了する。

3 評価実験

提案アルゴリズムによって、誤答検出にどのような効果があるか検証するため、性能評価実験を行った。実験に使用したデータは、著者らが所属する高専において、入試と同様に学生に答案用紙に解答を記入してもらい、それらを画像として保存したものである。12個の設問に対してそれぞれ400件のデータを収集し、のべ4800件の解答画像を収集した。今回は、数学の入試を想定し、正答は記号や数字からなる数式としたが、中にはより複雑な誤答を記載している場合もあり、誤答は全体の20%である。

本実験では、検出率および一致率を下記のように定義し、2つの尺度で性能評価を行った。その結果を図2に示す。

$$(\text{検出率}) = \frac{(\text{検出誤答数})}{(\text{誤答数})}$$

$$(\text{一致率}) = \frac{(\text{検出誤答中の真の誤答数})}{(\text{検出誤答数})}$$

部分的正規分布を用いた手法 [2] と比べると、検出率はわずかに上昇しているが、必ずしも本提案手法によ

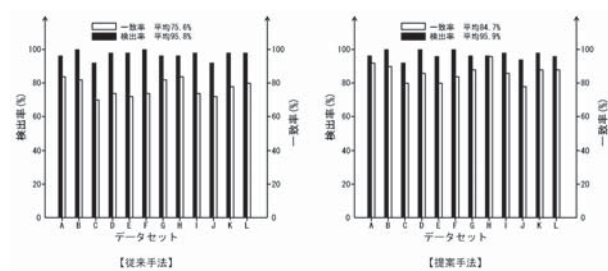


図 2: 性能評価実験結果

る効果が出ているとは言い難い。しかし、一致率は大きく改善されていることがわかる。これは、本来は正答であるものが誤答として検出されている件数を絞ることができることを意味しており、本アルゴリズムによって検出された誤答の確認作業時間を、飛躍的に改善することができる。

一方、従来の部分的正規分布による計算時間は、1つの設問に対して3~5分であるのに対し、提案手法では30~40分程度の時間を要する。部分的正規分布は正規分布を継ぎ合わせた分布であるため、仮説検定における α 点は、標準正規分布への変換によって容易に計算することができるのに対し、JohnsonSU分布では、個々の分布に対する α 点を求めるのに数値積分が必要となるためである。

4 まとめ

本稿では、入試における採点誤りを自動検出するために、JohnsonSU分布および仮説検定を用いた手法を提案した。従来の部分的正規分布を用いた手法と比べると、検出率はほぼ同等であるが一致率が改善され、検出された誤答の確認作業時間を大きく改善できることがわかった。しかし、提案手法による処理には膨大な時間がかかってしまう。今後は、実際の入試採点業務を想定して、処理時間を大幅に改善する方法を検討する必要がある。尚、本研究の一部は科学研究費補助金(基盤研究(C)課題番号22500170)の助成によるものである。

参考文献

- [1] 向山和宏, 鈴木雅人, 北越大輔, “入試採点支援における採点責任者業務支援システムの開発”, 2012年信学全大, D-9-33, Mar, 2012.
- [2] 鈴木雅人, 北越大輔, “分的正規分布に基づくパターン類別法による入試採点誤り検出の検討”, 信学技報 E2012-17, pp.7-12, Aug. 2012.